

**SOME PROBLEMS IN THE THEORY AND
APPLICATION OF THE METHODS OF NUMERICAL
TAXONOMY**

David Wishart

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



1970

Full metadata for this item is available in
St Andrews Research Repository
at:

<http://research-repository.st-andrews.ac.uk/>

Please use this identifier to cite or link to this item:

<http://hdl.handle.net/10023/10121>

This item is protected by original copyright

SOME PROBLEMS IN THE THEORY AND
APPLICATION OF THE METHODS OF
NUMERICAL TAXONOMY

by

DAVID WISHART, B.Sc.

ABSTRACT

Several of the methods of numerical taxonomy are compared and shown to be variants of a tripartite grouping procedure associated with a generalised intercluster similarity function involving ten computational parameters. Clustering by the techniques of hierarchic fusion, monothetic division and iterative relocation is obtained using different arithmetic combinations of the function parameters to both compute similarities and effect changes in cluster membership. The combinatorial solution for Ward's method is found, and the centroid sorting combinatorial solution is extended for size difference, shape difference, dispersion and dot product coefficients.

It is suggested that clusters are characterised more by the choice of similarity criterion than by the choice of method, and it is demonstrated that some common criteria such as distance and the error sum of squares are inclined to force spherical 'minimum-variance' classes. These are contrasted by 'natural' classes, which correspond to closed density surfaces defined for a multivariate sample space by the underlying probability density function.

ProQuest Number: 10166976

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10166976

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

A method for mode-seeking is developed from this probabilistic model through various theoretical and experimental phases, and it is shown to perform slightly better than iterative relocation with the minimum-variance criteria using several Gaussian test populations.

A fast algorithm is proposed for the solution of the Jardine-Sibson method for generating overlapping classes, and it is observed that this technique finds natural classes and is closely related to the probabilistic model.

Some aspects of computational procedures are discussed, and in particular, it is proposed that a generalised system involving a statistical language, conversational mode package and program suite could be developed from a basic subroutine system. Paging and simulation techniques for the organisation of direct-access data files are suggested, and a comprehensive package of computer programs for cluster analysis is described.

SOME PROBLEMS IN THE THEORY AND
APPLICATION OF THE METHODS OF
NUMERICAL TAXONOMY

by

DAVID WISHART, B.Sc.

Dissertation submitted for the degree
of Doctor of Philosophy of the
University of St. Andrews.

(May, 1970)



Th 5691

PREFACE

In October 1961, I matriculated at the University of St. Andrews and read for a degree in Mathematics at St. Salvator's College, graduating in June 1965 with the Ordinary Degree of Bachelor of Science. Following my appointment in January 1966 as computer programmer at the Computing Laboratory, I matriculated for the degree of Master of Science in Computational Science. In March 1969, I was admitted as a candidate for the degree of Doctor of Philosophy under Resolution of the University Court, 1967, No. 1, and credited with the time spent towards the M.Sc. degree. Throughout the period 1966-70 I was supervised by Professor A. J. Cole, the head of the Department of Computational Science.

DECLARATION

I declare that the following thesis
is a record of research work carried out by me,
that the thesis is my own composition, and that
it has not been presented in application
for a higher degree previously.

David Wishart

CERTIFICATE

I certify that David Wishart has satisfied the conditions of the Ordinance and Regulations and is thus qualified to submit the accompanying thesis in application for the degree of Doctor of Philosophy.

A. J. Cole

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor Professor A. J. Cole for his sustained guidance and criticism throughout this work. My thanks are due to everyone who has contributed towards my experience with the methods of numerical taxonomy through their comments on and evaluation of results. In this respect, I am particularly grateful to Dr. R. M. M. Crawford of the Department of Botany, University of St. Andrews, for our fruitful collaboration in the studies in plant ecology. I also acknowledge the contributions of Professor D. C. D. Pocock of Erindale College, University of Toronto, and Dr. A. H. Dawson of the Department of Geography, University of St. Andrews, in geographic and urban regionalisation studies, and Professor W. I. Card and Dr. T. Taylor of the Department of Medicine in relation to Mathematics and Computing of the University of Glasgow, for their useful discussions on the classification of diseases. I am similarly indebted to Mr. S. V. Leach for our joint work on Platonic prose rhythm and chronology. I wish also to express my thanks to Dr. T. Calinski of the Poznan College of Agriculture, Poland, for showing me the manuscript of his paper with J. Harabasz before publication, to Dr. D. F. Merriam of the University of Kansas for his help with the publication of my computer programs, and to all the programmers who have been concerned with the implementation of these programs on other computers. Finally, my thanks are especially due to my wife Doreen for her invaluable help with the development of my programs and preparation of documentation, to Mrs. M. Cunningham for typing this thesis so very well, and to Mr. T. McQueen for printing the end result.

CONTENTS

| | | |
|--|---------|-----|
| <u>INTRODUCTION</u> | | 1 |
| <u>CHAPTER 1: DATA OPERATIONS AND MEASURES OF SIMILARITY</u> | | 4 |
| 1.1 BASIC DATA CONSIDERATIONS | | 4 |
| 1.2 TRANSFORMATIONS AND WEIGHTING SCHEMES | | 14 |
| 1.3 PRINCIPAL COMPONENTS ANALYSIS | | 25 |
| 1.4 SIMILARITY COEFFICIENTS | | 39 |
| <u>CHAPTER 2: HIERARCHIC FUSION</u> | | 51 |
| 2.1 GENERAL PROCESS | | 51 |
| 2.2 METHODS | | 54 |
| 2.3 DISCUSSION | | 66 |
| <u>CHAPTER 3: DIVISIVE METHODS</u> | | 72 |
| 3.1 MONOTHEMIC DIVISION | | 72 |
| 3.2 POLYTHEMIC DIVISION | | 87 |
| <u>CHAPTER 4: ITERATIVE RELOCATION</u> | | 96 |
| 4.1 GENERAL PROCESS | | 96 |
| 4.2 METHODS | | 99 |
| 4.3 DISCUSSION | | 109 |
| <u>CHAPTER 5: GENERALISED TRIPARTITE PROCEDURE</u> | | 113 |
| 5.1 INTERCLUSTER SIMILARITY FUNCTION | | 113 |
| 5.2 METHODS | | 120 |
| 5.3 DISCUSSION | | 124 |

CONTENTS

| | | | | | |
|--|----|----|----|----|-----|
| <u>CHAPTER 6: THE PROBABILISTIC MODEL</u> | .. | .. | .. | .. | 129 |
| 6.1 MINIMUM-VARIANCE TECHNIQUES | . | .. | .. | .. | 129 |
| 6.2 NATURAL-CLASS METHODS | .. | .. | .. | .. | 138 |
| 6.3 SINGLE-LEVEL MODE ANALYSIS | .. | .. | .. | .. | 143 |
| 6.4 HIERARCHICAL METHOD | .. | .. | .. | .. | 145 |
| 6.5 DENSITY FUNCTION AND LARGE POPULATIONS | . | .. | | | 149 |
| 6.6 AVERAGE DISTANCE AS DENSITY ESTIMATE | .. | .. | | | 152 |
| 6.7 IMPROVED ALGORITHM | . | .. | .. | .. | 153 |
| | | | | | |
| <u>CHAPTER 7: EXPERIMENTAL TESTS</u> | .. | .. | .. | .. | 166 |
| 7.1 ITERATIVE RELOCATION TESTS | .. | .. | .. | .. | 168 |
| 7.2 HIERARCHICAL MODE ANALYSIS | .. | .. | .. | .. | 180 |
| 7.3 CONCLUSIONS | .. | .. | .. | .. | 185 |
| | | | | | |
| <u>CHAPTER 8: COMBINATORIAL COEFFICIENTS</u> | .. | .. | .. | .. | 188 |
| 8.1 LANCE-WILLIAMS' 4-PARAMETER MODEL | . | .. | .. | .. | 188 |
| 8.2 WARD'S METHOD | . | .. | .. | .. | 191 |
| 8.3 CENTROID SORTING | .. | .. | .. | .. | 195 |
| 8.4 COMBINATORIAL ALGORITHM | .. | .. | .. | .. | 198 |
| | | | | | |
| <u>CHAPTER 9: K-PARTITION</u> | . | .. | .. | .. | 205 |
| 9.1 JARDINE-SIBSON ALGORITHM | .. | .. | .. | .. | 205 |
| 9.2 COLE-WISHART ALGORITHM | . | .. | .. | .. | 207 |
| 9.3 CLUSTER RECOGNITION ALGORITHM | .. | .. | .. | .. | 217 |
| 9.4 DISCUSSION | .. | .. | .. | .. | 226 |
| | | | | | |
| <u>CHAPTER 10: COMPUTER TECHNIQUES</u> | .. | .. | .. | .. | 230 |
| 10.1 TOWARDS GENERALISED STATISTICAL SYSTEMS | .. | .. | | | 230 |
| 10.2 DATA STORAGE AND RETRIEVAL TECHNIQUES | .. | .. | | | 238 |
| 10.3 FEATURES OF CLUSTAN | .. | .. | .. | .. | 248 |

INTRODUCTION

Numerical taxonomy could be described as the branch of multivariate statistics which is concerned with the simplification and description of observational data. Two general problems arise in connection with data simplification. Firstly, an observer who wishes to describe statistically a 'concept' or 'frame of reference' which he has chosen for study may encounter difficulty in the selection of relevant variables. Secondly, methods of analysis must be found which fit both the types of data and the way in which they are to be treated. Selection of variables which characterise a complex concept, such as 'areal class structure of a town' or 'hominoids', depends largely on the observer's personal idea of which measurable attributes contribute usefully towards the variability within his sampling frame. At this stage, considerations of methodology should not arise, for the observer must be completely free to make any measurements which he thinks are important. The statistician must therefore design methods which take into account all aspects of sampling, making allowances for such undesirable effects as strong multiple correlations, the inclusion of poor variables, weak orthogonal components, and so on. It is very easy to sidestep the consideration of data and formulate rules about what the methods are to do in a given sampling scheme, implying that if the data do not suit these rules then it is the data which must change and not the methods.

One of the least studied aspects of the subject is structure in the multivariate sample space, and there is sometimes a very dangerous tendency to dismiss this topic altogether - I refer to writers who begin tidily with "a suitable similarity matrix" and proceed to define rigorous procedures based on intuitive ideas of what should be done with similarity coefficients. Another hazard is the 'logic' deduced from descriptive arguments which use the M-dimensional euclidean space as a model, 'thinking' of it in terms of 3-dimensional reality. Who is to say that M-space behaves like 1, 2 or 3-space? There are some excellent examples of the unpredictability of M dimensions (Day, 1970). On the other hand, we cannot afford to be so confident as to reject altogether the euclidean model as an indicator; yet there are those who adopt a more topological approach to the subject, and appear to regard the euclidean model as a rather trivial specialisation (e.g. Jardine, et al, 1967).

One source of information that we cannot possibly ignore is the wealth of empirical evidence which appears throughout the literature in the justification of individual procedures. It was surely empirical studies which revealed the chaining effect of single linkage and fragmentation in association analysis. Also, the fact that no satisfactory definition of the concept 'cluster' exists makes the collective examination of empirical results and intuitive procedures one of the areas of

investigation which are most likely to lead to a formal theory. Consequently, an important part of the work reported here has been the development of a system of computer programs for cluster analysis which is available to workers in all disciplines for the collective empirical study of existing methods. The system has been carefully designed to make it easy for new procedures to be added - indeed, despite the rather complex use of magnetic disk and tape (Chapter 10), the actual clustering programs are totally machine independent; thus a new procedure can now be introduced at about 50 installations throughout the world without any changes in data set assignments or the job control language specifications for each machine and operating system.

This thesis is concerned more with the treatment of data and the properties of methods. An attempt has been made to survey the range of methods, making generalisations where possible in order to deduce their common properties. In a sense, the work constitutes a classification of classification methods using as data the kernel of each technique considered. Every care has been taken to avoid fixing standards, so that the conclusions may be free from conjecture based on pre-defined requirements. However, a few concessions have to be made in order to permit the discussion of multivariate structure, and it is hoped that the topologist, in particular, will allow the use of the euclidean model for illustrations.

CHAPTER 1: DATA OPERATIONS AND MEASURES OF SIMILARITY

1.1 BASIC DATA CONSIDERATIONS

The general objective of cluster analysis is to find a grouping of N individuals into k classes which is 'meaningful', and constitutes a useful classification of the population. This very vague statement is about as near as we ever get to generalising the methods of numerical classification. In order to be more specific we must devise numerical models to represent populations, specify structural limits for the k classes, interpret the notion 'meaningful' in relation to actual problems or abstract generalisations, and explain how the results can be used. We can, however, state the general classification result as follows:

1. There shall be k groups of individuals such that each group contains at least one individual.
2. Each individual may be assigned to no group (if it is 'unclassifiable'), one group only (if disjoint clusters are required), or more than one group (when overlapping clusters are permitted).

For the most part we shall be concerned with techniques that derive disjoint clusters (although Chapter 9 concerns a method dedicated to finding overlapping clusters), and our first problem is that of defining a numerical model to represent a wide range of observational material.

Continuous data

'Continuous' or 'quantitative' data are measurements of quantities which range on a 'continuous' scale. We can easily distinguish between quantitative and qualitative (see below) observations, but it is sometimes less easy to say when a continuous variable can no longer be treated as such, and should be regarded as an ordered multistate character (see below). For example, we may regard population in countries as continuous (the range is 'continuous' on a scale from a few thousand to 600 million); similarly, population in cities and towns, boroughs, or wards may be treated as continuous; however, is it reasonable to treat population in houses or rooms of houses as a continuous variable, when the range is only about 1 to 10? Thus we have encountered the first demand for subjective decision, namely, the choice between the treatment of semi-quantitative data as either continuous or multistate variables.

A typical small raw continuous data matrix is shown in Table 1.1.1. Six variables (the number of service establishments per 1000 population for six categories) are measured for nine individuals (census divisions of the USA). We denote by X_{ij} the value of variable j for the i th individual (the j th element of row i in Table 1.1.1), and define the mean and variance for variable j by

mean
$$U_j = \frac{1}{N} \sum_{j=1}^M X_{ij} \quad (1.1.1)$$

| Census Division | Personal | Business | Auto. Repair | Misc. Repair | Amusement | Hotels etc. |
|--------------------------|----------|----------|--------------|--------------|-----------|-------------|
| 1 New England | 2.56 | 0.57 | 0.53 | 0.69 | 0.43 | 0.46 |
| 2 Middle Atlantic | 2.70 | 0.72 | 0.54 | 0.72 | 0.41 | 0.25 |
| 3 E.N. Central | 2.10 | 0.50 | 0.52 | 0.68 | 0.46 | 0.30 |
| 4 W.N. Central | 2.11 | 0.47 | 0.71 | 0.84 | 0.56 | 0.53 |
| 5 S. Atlantic | 1.74 | 0.38 | 0.49 | 0.53 | 0.42 | 0.42 |
| 6 E.S. Central | 1.38 | 0.25 | 0.38 | 0.41 | 0.33 | 0.22 |
| 7 W.S. Central | 2.04 | 0.45 | 0.68 | 0.80 | 0.45 | 0.40 |
| 8 Mountain | 1.92 | 0.57 | 0.70 | 0.78 | 0.55 | 1.24 |
| 9 Pacific | 2.37 | 0.87 | 0.82 | 0.87 | 0.51 | 0.63 |
| Mean U_j | 2.10 | 0.53 | 0.60 | 0.70 | 0.46 | 0.49 |
| Variance S_j^2 | 0.144 | 0.029 | 0.017 | 0.020 | 0.005 | 0.084 |
| Standard Deviation S_j | 0.38 | 0.17 | 0.13 | 0.14 | 0.07 | 0.29 |
| Minimum | 1.38 | 0.25 | 0.38 | 0.41 | 0.33 | 0.22 |
| Maximum | 2.70 | 0.87 | 0.82 | 0.87 | 0.56 | 1.24 |
| $P_j = S_j/\text{Range}$ | 0.288 | 0.274 | 0.295 | 0.304 | 0.304 | 0.284 |

Table 1.1.1. A typical continuous data matrix comprising six variables measured for nine individuals (census divisions of the U.S.A.)

variance
$$S_j^2 = \frac{1}{N} \sum_{i=1}^M (X_{ij} - U_j)^2 \quad (1.1.2)$$

where N is the number of individuals, and M the number of variables. The standard deviation is therefore given by S_j .

Table 1.1.1 also shows the mean, variance, standard deviation, minimum and maximum for each of the six variables; we observe that the category 'personal' has the largest variance (0.144) and range (1.32), while 'amusements' has the smallest variance (0.005) and range (0.23).

Numerical classification techniques are invariably concerned with the comparison of individuals or groups of individuals (clusters) in terms of the set of M variables. Perhaps the most common measure of the similarity between two individuals is the 'squared euclidean distance' coefficient: each individual is represented by a point in M-dimensional space whose coordinates are the associated M variable scores, and we may compute the squared distance d_{ik}^2 between two points i and k using the formula

$$d_{ik}^2 = \frac{1}{M} \sum_{j=1}^M (X_{ij} - X_{kj})^2 \quad (1.1.3)$$

Thus, for example, using the data of Table 1.1.1 d_{12}^2 is

$$\begin{aligned} d_{12}^2 &= \frac{1}{6} \left\{ (2.56-2.70)^2 + (0.57-0.72)^2 + (0.53-0.54)^2 + \right. \\ &\quad \left. (0.69-0.72)^2 + (0.43-0.41)^2 + (0.46-0.25)^2 \right\} \\ &= 0.0142 \end{aligned}$$

$(X_{ij} - X_{kj})^2$ is the component of d_{ik}^2 which is attributed to variable j , and the mean of this component for the N^2 possible d_{ik}^2 coefficients is

$$\begin{aligned} E(d^2)_j &= \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N (X_{ij} - X_{kj})^2 \\ &= \frac{1}{N^2} \sum_i \sum_k [(X_{ij} - U_j) - (X_{kj} - U_j)]^2 \end{aligned}$$

which reduces, after expansion, to

$$\begin{aligned} \frac{1}{N^2} \sum_i N S_j^2 + N(X_{ij} - U_j)^2 \\ = 2S_j^2 \end{aligned}$$

It follows that the expected contribution to d^2 of each variable is proportional to the variance, and hence the distance coefficient is biased towards variables with high variance. From Table 1.1.1, we see that the mean of the distance component for the category 'personal' is 0.288, while for 'amusements' the mean is 0.01. Variable 1 is therefore weighted by the factor 288, and consequently 'amusements' has almost no influence on the resulting distance coefficients. For this reason it is customary to 'standardise' continuous data so that the contributions of each variable are of equal importance. This is achieved by replacing each X_{ij} with

$$X_{ij} = X_{ij}/S_j \quad (1.1.4)$$

or, more usually,

$$X_{ij} = \frac{X_{ij} - U_j}{S_j} \quad (1.1.5)$$

Both standardisations transform the variables to unit variance, so that the expected contribution to d^2 of each variable is 2. (1.1.5) is the more usual formula since the resulting scores have zero mean; strong deviations from the mean are then easily observed as deviations from zero in the X_{ij} 's. Table 1.1.2 shows the standard scores obtained using formula (1.1.5) with the data of Table 1.1.1. Since the expected component of d^2 for each vector of standard scores is 2, the expected value of d^2 for M independent continuous variables will also be 2; however, this result is of little value since independence can seldom be assumed.

| Census Division (Sample) | Variables | | | | | |
|--------------------------------|-----------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1.19 | 0.23 | -0.51 | -0.09 | -0.40 | -0.12 |
| 2 | 1.56 | 1.10 | -0.44 | 0.13 | -0.69 | -0.84 |
| 3 | -0.01 | -0.18 | -0.59 | -0.16 | 0.03 | -0.67 |
| 4 | 0.02 | -0.36 | 0.87 | 0.98 | 1.49 | 0.12 |
| 5 | -0.94 | -0.88 | -0.82 | -1.22 | -0.55 | -0.26 |
| 6 | -1.88 | -1.63 | -1.67 | -2.07 | -1.86 | -0.94 |
| 7 | -0.16 | -0.47 | 0.64 | 0.69 | -0.11 | -0.32 |
| 8 | -0.47 | 0.23 | 0.79 | 0.55 | 1.34 | 2.56 |
| 9 | 0.70 | 1.97 | 1.72 | 1.19 | 0.76 | 0.47 |

Table 1.1.2. Standard scores derived from the data of Table 1.1.1 using equation (1.1.5)

Binary data

Since it is not always possible to obtain continuous or semi-

continuous data, we must adopt the alternative 'qualitative' or 'binary' mode. This enables the recording and manipulation of 'qualities', of which the most fundamental are binary attributes that can exist in one of two states: present or absent. In fact, we shall see later (Sect. 1.2) that all other qualitative data can be reasonably transformed into 2-state attributes, so that the binary mode becomes the single alternative to continuous data.

The binary data for N individuals which possess or lack M attributes is usually recorded using an $N \times M$ binary matrix such as Table 1.1.3. By convention, 'presence' of an attribute is

| | | Binary attributes | | | | | | | | | |
|-------|--|-------------------|---|---|---|---|---|---|---|---|----|
| CASES | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| 3 | | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |

Table 1.1.3. A typical binary data matrix for 10 attributes and 4 individuals

recorded by 1, while 0 denotes absence: thus, in Table 1.1.3, attribute 1 is present for individuals 1, 3 and 4, but absent for individual 2. If we denote by α_{ij} the contents of the j th element of row i , then individual i may be thought of as a point in M -space

having coordinates

$$(\alpha_{i1}, \dots, \alpha_{ij}, \dots, \alpha_{iM})$$

While such data are clearly discrete (the points lie only at the vertices of an M-dimensional hypercube), there appears to be no reason why the intuitive rules of continuous data cannot be equally well applied to binary data using the above spatial representation. In fact, we may evaluate the mean

$$U_j = \frac{1}{N} \sum_{i=1}^M \alpha_{ij} = p_j$$

where p_j is the probability associated with attribute j , and variance (from 1.1.2)

$$\begin{aligned} S_j^2 &= \frac{f_j(1-p_j)^2 + (N-f_j)p_j^2}{N} \\ &= p_j(1-p_j) \end{aligned} \tag{1.1.6}$$

and then standardise using equation (1.1.5) so that α_{ij} is replaced by

$$X_{ij} = \frac{\alpha_{ij} - p_j}{\sqrt{p_j(1-p_j)}} \tag{1.1.7}$$

Standardisation of binary data is not usual, but it has been suggested by Lance and Williams (1966c) and Williams et al (1966), and the latter writers state the following case:

"It might plausibly be suggested that the joint presence of two rare attributes (or joint absence of two common ones) is more meaningful than the joint presence of two

common attributes (or joint absence of two rare ones).

To weight such joint occurrences appropriately, the

attributes are standardised before the analysis begins."

Such arguments can be seen to be dangerous when we extrapolate to an extreme, such as attributes having only 1% occurrence. From (1.1.6), the attribute variance is

$$s_j^2 = .01(.99) = .0099$$

so that those individuals i which possess such an attribute have standardised coordinate (1.1.7) of

$$X_{ij} = \frac{1 - .01}{\sqrt{.0099}} = 9.95$$

while all other individuals k have coordinate

$$X_{kj} = \frac{-.01}{\sqrt{.0099}} = -0.1$$

Hence the component of distance is zero in the comparison of a possessing individual i with any other possessing individual i , or a non-possessing individual k with any other k , while in the comparison of any possessing individual i with a non-possessing individual k the component of distance is

$$(9.95 - (-0.1))^2 = (10.05)^2 = 101$$

Although it is still true that the overall mean of this component is 2, it is clear that the value 101 is sufficiently large to completely separate all those individuals which possess the rare

attribute j , to the extent that ordinary data are worthless. Since binary data are already 'normalised', in the sense that the range of values for each attribute is 1, it is probably safer in general to use unstandardised 1/0 data, thereby avoiding the weighting of rare attributes (see also Sect. 1.2).

Multistate characters

The idea of a binary attribute existing in one of 2 states can be extended to that of a multistate attribute which can exist in one of R states. We shall distinguish between two types: ordered multistate and unordered multistate characters, which differ according to whether two particular states can be said to be more closely related than two others. An example of an unordered multistate character is 'colour of hair' - Table 1.2.2. In this instance, we cannot convincingly say that two colours are more closely related than any other two. The converse is an ordered multistate character, for which there is a very definite relationship between the states that must be taken into account. Table 1.2.3 contains the example of 'age' coded as an ordered multistate character. We shall assume that a certain sample population of ladies, although unwilling to state their actual ages, were prepared to say whether they were in their 'teens, twenties, and so on. In this case, we must certainly take account of the strong relationship 'teens-twenties in comparison with the weak relationship 'teens-forties.

1.2 TRANSFORMATIONS AND WEIGHTING SCHEMES

In the previous section four types of observational measurements were introduced, together with such fundamental operations as standardisation and the computation of distances. We must now consider ways of transforming data of one type into another type so that unbiased measures of similarity (such as distance) may be evaluated using standard formulae. The four possible data types previously introduced are:

Continuous

Binary

Ordered Multistate

Unordered Multistate

but we shall restrict data for analysis to only the continuous and binary computation modes. The following transformations, which have been discussed by Wishart (1969d), enable any of the four data types to be expressed in either of these two computation modes, and all possible combinations of these transformations are summarized in Table 1.2.5.

T_{BC}: Binary to Continuous

It has already been suggested that binary data may be treated as 1/0 M-dimensional coordinate vectors which may or may not be standardised. Hence, in order to compare binary and continuous data in the continuous mode we simply code the binary attributes 1.0 for 'presence' or 0.0 for 'absence' and treat the resulting

scores as quasi-continuous. This transformation is illustrated in Table 1.2.1.

| | | | | | | | | |
|---|-----------|------------|----------|---|-------------------------------|--|-----|-----|
| Binary | | Continuous | | | | | | |
| <table border="1" style="border-collapse: collapse;"> <tr><td>'present'</td><td style="text-align: center;">1</td></tr> <tr><td>'absent'</td><td style="text-align: center;">0</td></tr> </table> | 'present' | 1 | 'absent' | 0 | T_{BC} \longrightarrow | <table border="1" style="border-collapse: collapse;"> <tr><td style="text-align: center;">1.0</td></tr> <tr><td style="text-align: center;">0.0</td></tr> </table> | 1.0 | 0.0 |
| 'present' | 1 | | | | | | | |
| 'absent' | 0 | | | | | | | |
| 1.0 | | | | | | | | |
| 0.0 | | | | | | | | |

Table 1.2.1. Transformation from binary data into the continuous computation mode

T_{UB} : Unordered multistate to Binary

For an unordered multistate character having R states we create R binary attributes, such that each state is coded as the 'presence' of one attribute. T_{UB} is illustrated with the 4-state character 'colour of hair' in Table 1.2.2.

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|-------|--------|-----|---|-------|---|-------|---|-------------------------------|---|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unordered multistate | | Binary | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1" style="border-collapse: collapse;"> <tr><td>White</td><td style="text-align: center;">1</td></tr> <tr><td>Red</td><td style="text-align: center;">2</td></tr> <tr><td>Brown</td><td style="text-align: center;">3</td></tr> <tr><td>Black</td><td style="text-align: center;">4</td></tr> </table> | White | 1 | Red | 2 | Brown | 3 | Black | 4 | T_{UB} \longrightarrow | <table border="1" style="border-collapse: collapse;"> <tr> <td></td> <td style="text-align: center;">1</td> <td style="text-align: center;">2</td> <td style="text-align: center;">3</td> <td style="text-align: center;">4</td> </tr> <tr> <td style="text-align: center;">1</td> <td style="text-align: center;">1</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> </tr> <tr> <td style="text-align: center;">2</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> </tr> <tr> <td style="text-align: center;">3</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">0</td> </tr> <tr> <td style="text-align: center;">4</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> </tr> </table> | | 1 | 2 | 3 | 4 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 0 | 1 |
| White | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Red | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Brown | 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Black | 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 1 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 0 | 1 | 0 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 0 | 0 | 1 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 0 | 0 | 0 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Table 1.2.2. Transformation from an unordered multistate character into the binary computation mode

T_{OB} : Ordered multistate to Binary

Given R ordered states, (R-1) binary attributes are scored

such that the j th of these attributes is coded 'present' only for those individuals having a character state code greater than j . This has the effect of introducing a more positive match between adjacent states than between distant state codes. Table 1.2.3 illustrates T_{OB} for a 4-state character, and it is seen that, for example, the component of unstandardised distance ranges from 0 (total match) through 1 and 2 to 3 (in the comparison of states 1 and 4).

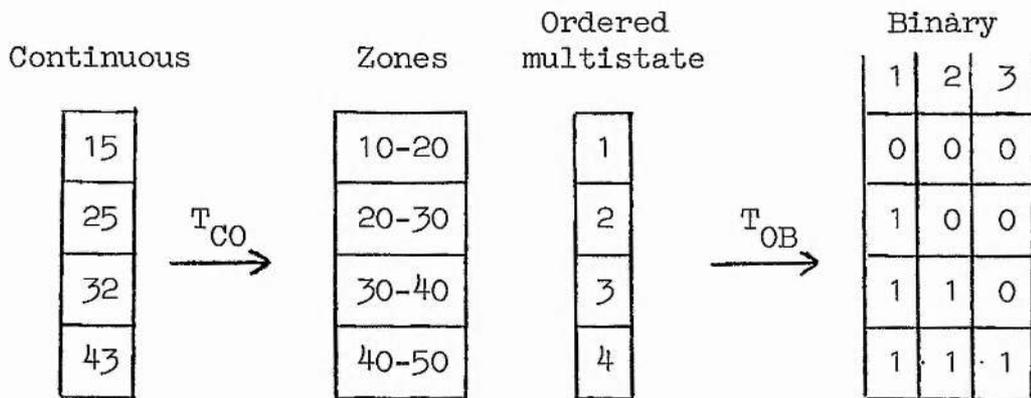


Table 1.2.3. Transformations T_{CO} and T_{OB} which convert a continuous variable to ordered multistate, and ordered multistate to binary, respectively.

T_{CO} : Continuous to Ordered multistate

To transform continuous data into the binary computation mode the first step requires the computation of an ordered multistate character such that each state j corresponds to one of R ranges of values for the continuous variable. Thus we can code the continuous variable 'age' as 1, 2, 3, or 4 according to whether

a person's age was in the range 10-20, 20-30, 30-40, or 40-50. Table 1.2.3 shows transformation T_{CO} for this variable, and it is seen that T_{CO} can then be combined with T_{OB} to complete the transformation T_{CB} from continuous data to binary.

T_{OC} : Ordered multistate to Continuous

This transformation, illustrated in Table 1.2.4, is weakest when generalised. Depending on the relationships between the R states of an ordered multistate character, numeric values are chosen to replace the state codes accordingly. In the example, we replace codes 1, 2 and 3 by the same numeric values. The selection of these substitute codes is crucial, even if the resulting vector of scores is standardised, because the differences of the codes from the overall mean are reflected proportionately by the standard scores. Every attempt should be made to associate the inserted codes with the means of the ordered intervals which they represent. For example, a percentage variable which has been zoned 0-30, 30-50, 50-70, and 70-100 could have substitute codes 15, 40, 60 and 85 if the distribution is known to be rectangular, or 22, 43, 57 and 78 if it is a normally distributed variable.

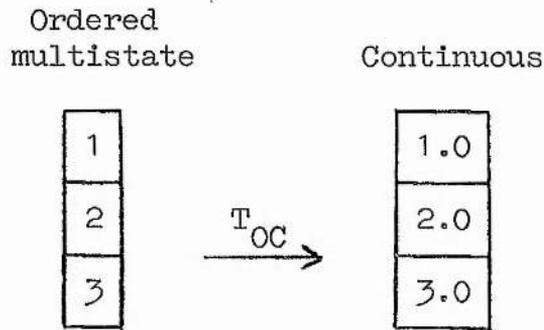


Table 1.2.4. Transformation from an ordered multistate character into the continuous computation mode

Table 1.3.5 shows how these five transformations can be combined to convert data of all four types into the two basic computation modes.

| | | TO | |
|----------------------|-------------------|-------------------|---|
| FROM | Binary | Continuous | |
| Binary | - | T_{BC} | |
| Continuous | $T_{CO} + T_{OB}$ | | - |
| Ordered multistate | T_{OB} | T_{OC} | |
| Unordered multistate | T_{UB} | $T_{UB} + T_{BC}$ | |

Table 1.2.5. Summary of the transformations from the four possible data types into the two computation modes Binary and Continuous

Standardisation

In Sect. 1.1 it was argued that continuous data should be standardised so that each variable has a mean distance component of 2. It was further argued that when binary data are standardised then the distance component of rare binary attributes can bias coefficient values to such an extent that other data are worthless. This complaint can also be applied to continuous data which is "ill-conditioned", in the sense that some values do not conform to a nicely rounded density function. For example, if we were to count the numbers of theatres in a survey of British towns then almost all the values will be near 1, and not exceeding about 5. By comparison, the count for London would be of the order 50, and the standardised value would probably also be about 50. The squared distance component of a coefficient which compares any other town with London would therefore be of the order 2500, which does not compare at all favourably with the mean 2. We can assume that if a classification problem exists, then the density function for some variables cannot be expected to exhibit nice unimodal distributions. Standardisation must therefore be used with care, and can dangerously influence the results if applied to ill-conditioned data.

Normalisation

In the comparison of binary, multistate and continuous data, Parks (1969a, 1969b) uses a normalisation technique which ensures

that $\max(d_{ik}^2) = 1$. Binary data are treated as variable values 0.0 and 1.0 (transformation T_{BC}); continuous data are recorded as

$$X_{ij} = \frac{X_{ij} - \min(X_{.j})}{\max(X_{.j}) - \min(X_{.j})} \quad (1.2.1)$$

where $\min(X_{.j})$ and $\max(X_{.j})$ denote the bounds of variable j .

Multistate characters are transformed using T_{OC} , where the substitute code for the j th of R states is

$$\frac{(j-1)}{(R-1)}$$

Hence the multistate character codes 1-5 become 0.0, 0.25, 0.5, 0.75 and 1.0 (Parks does not appear to distinguish ordered from unordered multistate characters).

The great disadvantage of this technique is that the resulting scores are determined by single bound values. In the previous example of theatres in British towns, the range of values for this variable throughout towns other than London would be either 0 to 5/5 if London were excluded, or 0 to 5/50 if London were included. Thus the inclusion of a 'rare' extreme individual is seen to radically influence the coding of the rest of the population (it should be noted that mistakes in the recording or coding of data might have the same effect).

The second important disadvantage of Parks' normalisation is the inadvertent weighting he imposes by choosing the range rather

than the standard deviation as normalising factor with continuous data. If we use Parks' normalisation (1.2.1) instead of standardisation (1.1.5) with the data of Table 1.1.1, then we obtain the standard deviations P_j and variances P_j^2 which are shown at the bottom of Table 1.1.1. Since the component of distance for any variable j has a mean of $2S_j^2$, we see immediately that Parks weights variables 4 and 5 more heavily than variable 2 by a factor of 1.24. Also, if we were to add binary data to continuous using Parks' normalisation, then any binary attribute having 50% occurrence has standard deviation (from 1.1.6) $S_j = 0.5$ and variance $S_j^2 = 0.25$. Thus such an attribute is weighted roughly $3\frac{1}{2}$ times heavier than variable 2, and $2\frac{1}{2}$ times heavier than variable 4 of Table 1.1.1. It would seem, therefore, that Parks' normalisation is undesirable when used with mixed-mode data.

Weighting Considerations

Table 1.2.5 contains a summary of the proposed transformations from all four variable types into either binary or continuous data. These transformations do not account for any further weighting which is designed to standardise or normalise the observations, and methods must therefore be devised for eliminating or reducing bias. In the case of continuous data, all transformations yield one variable and therefore the origin of the data is not strictly important: we may consider normalisation or standardisation

of each continuous variable regardless of its derivation, provided that account is made for ill-conditioned variables.

With the binary case, the creation of more than one binary variable from a multistate character (which encompasses all three other types) immediately introduces a bias. For any multistate character, let there be R states such that for N individuals the j th state occurs f_j times. We define the probability of the j th state as $p_j = f_j/N$, and now deduce the mean component of binary distance for multistate characters.

Unordered multistate (T_{UB}).

Table 1.2.2 shows the four possible binary patterns obtained from the transformation (T_{UB}) of an unordered 4-state character to binary. The number of times that two individuals will match scores is given by

$$t(0) = f_1^2 + f_2^2 + \dots + f_R^2$$

and the total number of comparisons between individuals is N^2 . Hence, the frequency of a mismatch, for which the component of binary distance is 2, is given by

$$t(2) = N^2 - \sum_{j=1}^R f_j^2$$

The mean component of distance on an unordered multistate character is therefore

$$\begin{aligned}
 E(d^2) &= 2(N^2 - \sum_{j=1}^R f_j^2)/N^2 \\
 &= 2(1 - \sum_{j=1}^R p_j^2) \\
 &= 2 \sum_{j=1}^R p_j(1-p_j)
 \end{aligned}$$

Ordered multistate (T_{OB}).

In Table 1.2.3 a continuous variable is transformed by T_{CO} into an ordered multistate character, which is then transformed by T_{OB} into 3 binary attributes. We observe that attribute j is 'present' for all state codes greater than j . In this example, we can write the expected frequencies of each component of binary distance (from 1 to 3) as follows:

$$\begin{aligned}
 t(1) &= f_1 f_2 + f_2 f_3 + f_3 f_4 \\
 t(2) &= f_1 f_3 + f_2 f_4 \\
 t(3) &= f_1 f_4
 \end{aligned}$$

The total sum of binary distance components t can now be written as:

$$\begin{aligned}
 t &= f_1 f_2 + f_2 f_3 + f_3 f_4 \\
 &\quad + 2(f_1 f_3 + f_2 f_4) \\
 &\quad + 3(f_1 f_4)
 \end{aligned}$$

which, in general, is obtained with the formula

$$\sum_{j=1}^{R-1} f_j \sum_{j'=j+1}^R (j' - j) f_{j'}$$

The mean distance component for a multistate character is therefore

$$\sum_{j=1}^{R-1} p_j \sum_{j'=j+1}^R (j' - j)p_{j'}$$

As shown in Sect. 1.1, the mean distance component of a continuous variable is $2S_j^2$, which reduces to $2p_j(1-p_j)$ in the case of a single unstandardised binary attribute. It would seem, therefore, that an adequate standardisation is obtained if the scores for each type of variable are divided by the appropriate mean distance component; in the case of multistate characters, each binary attribute produced by T_{UB} or T_{OB} should be weighted by the overall mean distance component. However, this scheme holds only for the manipulation of distances. We should now extend the principle of evaluating mean coefficient contributions to any similarity measure. For example, with the dot product coefficient (Sect. 1.4)

$$S_{ik} = \frac{1}{M} \sum_{j=1}^M X_{ij} X_{kj}$$

the mean component of variable j is seen to be

$$\frac{1}{N^2 M} \sum_i \sum_k X_{ij} X_{kj} = \frac{1}{M} U_j^2$$

In this case, therefore, it would appear that each variable j is unbiased only when divided by U_j^2 .

The choice of appropriate weighting schemes which yield unbiased variable vectors clearly requires considerable investigation; this discussion is confined solely to pointing out some of the

problems which exist. In general, we shall assume that an appropriate weighting scheme has been chosen, and the desired transformations have been completed. The data for classification will therefore take the form of an $N \times M$ matrix of either binary presence/absence (1/0) scores $[\alpha_{ij}]$ or continuous variable values $[x_{ij}]$.

1.3 PRINCIPAL COMPONENTS ANALYSIS

Berry (1961) suggests that a transformation to principal component scores will eliminate the redundancies incurred when several variables display a single pattern of concomitant variation. Each pattern of correlated variables is replaced by a single component which represents the pattern, and the point distribution can be described approximately in terms of a smaller number of uncorrelated component variables. It is certainly true that, as Berry claims, the transformation will in some instances save considerable computation, especially when a large number of initial variables is used. The analysis is also of interest in its own right, because inevitably any classification obtained from the data will be a function of the initial variables, and the isolation of the major factors present as a result of the choice of variables gives an indication of the terms of reference to which the classification applies (the classifications obtained from quadrat sampling of a town on socio-economic variables may be

completely different from those derived, for example, from health variables). It cannot be stressed too strongly that the results obtained from any classification technique are dependent on the original choice of variables, and therefore the derivation of 'meaningful' principal components can be extremely helpful where clarification of the frame of reference is required.

In principal components analysis, the original coordinate axes are rotated to a new set of orthogonal axes so that the major axis (factor) is the line of best fit through the point swarm (that is, it accounts for the maximum amount of variance), and successive factors are similar lines of best fit subject to the constraint that they must be, in each case, orthogonal to each of their predecessors. The result can be demonstrated by a set of points which lie in a plane through a three-dimensional space. The first axis will lie along the line of best fit, the second is orthogonal to the first, and the third which must be orthogonal to the plane that contains the points is therefore redundant. If factor scores (coordinates) are obtained for the first two principal axes, then the distance between any two points under this system will be identical to the distance measured in the original three-variable system. In the general case, the first few components will usually account for a large proportion of the overall variance in the point distribution, and when scores,

computed on these factors alone, are used for measuring distances, good approximations to the true distances in M-space are achieved. This reduction from M variables to a few (f, say) factors corresponds to a projection of the point swarm from M-space into f-space with the minimum possible distortion of the point orientation. Factor loadings, or eigenvectors, obtained¹ from the product-moment correlation matrix for the six variables of Table 1.1.1 are shown in Table 1.3.1, and the corresponding transformation to factor scores is given in Table 1.3.2.

| | Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|------------|----------|-------|-------|-------|-------|------|-------|
| Component: | 1 | 0.31 | 0.40 | 0.47 | 0.48 | 0.44 | 0.31 |
| | 2 | -0.63 | -0.41 | 0.14 | -0.06 | 0.33 | 0.55 |
| | 3 | -0.16 | -0.43 | 0.31 | 0.39 | 0.24 | -0.69 |
| | 4 | 0.54 | -0.50 | -0.51 | 0.09 | 0.42 | 0.14 |
| | 5 | -0.21 | 0.43 | -0.25 | -0.39 | 0.68 | -0.31 |
| | 6 | 0.37 | -0.24 | 0.58 | -0.67 | 0.09 | -0.02 |

Table 1.3.1. Factor loadings (eigenvectors) obtained by principal components analysis of the USA census data of Table 1.1.1

¹ for a fuller account of the method of evaluating principal components analysis see either Cooley and Lohnes (1964), or Morrison (1967).

| Census Division (Sample) | Components | | | | | |
|--------------------------------|------------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | -0.04 | -1.11 | 0.50 | 0.60 | -0.23 | -0.12 |
| 2 | 0.21 | -2.19 | 0.40 | 0.12 | -0.01 | 0.06 |
| 3 | -0.62 | -0.35 | -0.30 | 0.29 | 0.36 | 0.18 |
| 4 | 1.43 | 0.76 | -1.08 | 0.47 | 0.21 | -0.07 |
| 5 | -1.94 | 0.58 | 0.16 | -0.03 | 0.21 | -0.16 |
| 6 | -4.13 | 0.59 | 0.11 | -0.46 | -0.04 | 0.03 |
| 7 | 0.25 | 0.13 | -0.90 | -0.21 | -0.58 | 0.04 |
| 8 | 1.97 | 2.14 | 1.01 | 0.21 | -0.11 | 0.08 |
| 9 | 2.87 | -0.55 | 0.10 | -0.98 | 0.18 | -0.05 |

Table 1.3.2. Factor scores for the six components of the USA census data, obtained by principal components analysis

The contributions of components 5 and 6 to the distance coefficient for any two samples may be seen to be small by comparison with the four major components. There is little difference between the distance measures obtained using (a), all six standard scores and (b), the first four component scores. When distances are obtained using all six factor scores (as adopted by Berry), the results are identical to those derived from the standard scores, and in this instance, the principal components analysis is ineffective.

There seems to be no clear rule for determining the number of factors that should be chosen to define the components' sub-

space. H. F. Kaiser (1959) suggests a rule for principal components analysis where significant components are those which account for an eigenvalue not less than unity, but whether this rule should be adopted for classification methods is doubtful, for, on the basis of the unity rule, only the first two factors obtained from the present census data would be adopted (Table 1.3.3). While it is true that combined they account for 88.5 per cent of the overall

Variation Explained

| Component | Total | Percentage | Cumulative Percentage |
|-----------|-------|------------|-----------------------|
| 1 | 3.94 | 65.7 | 65.7 |
| 2 | 1.37 | 22.8 | 88.5 |
| 3 | 0.39 | 6.5 | 95.0 |
| 4 | 0.22 | 3.6 | 98.6 |
| 5 | 0.07 | 1.2 | 99.8 |
| 6 | 0.01 | 0.2 | 100.0 |

Table 1.3.3. Analysis of the variance within the USA census data of Table 1.1.1 by principal components analysis

variance, the sizeable contributions to the distance measures of the scores for factors 3 and 4 in Table (1.3.2) suggest that, for classification purposes, the unity rule would involve an oversimplification and create excessive distortion. However, the dichotomy that exists concerning the choice of relevant components is not present in this instance, for clearly too many components

cannot be selected. The suggestion by D. F. Morrison (1967) that components should be chosen which together explain some arbitrary percentage of the total variance seems to be more pertinent to classification methods, and a level of 90 or 95 per cent of the variance would be reasonable.

More recently, Berry (1965) has proposed an additional standardisation of the component scores prior to classification. This effectively destroys all relationship between distance measures obtained using the original standard scores and the new component scores. The components are standardised in such a way that they have equal importance, a state which is clearly not substantiated by the many applications of principal components analysis. Problems now arise concerning the number of components which should be used, and the incorporation of components which do not have meaningful interpretations. But what is most important is that the technique has the effect of creating a 'synthetic' frame of reference in which inter-sample similarities no longer correspond to the observed relationships. Classification methods which derive similarity measures from eigenvectors normalised in this way produce their results from a frame of reference which differs completely from that of the original data; it is therefore suggested here that Berry's normalising procedure should be avoided.

The adoption by D. M. Ray and Berry (1965) of an additional rotation from the principal components solution to a normal Varimax frame of reference has certain advantages concerning the interpretation of factors. This may be adopted when meaningful principal components cannot be derived and a factor analysis appraisal of the data structure is desired. It is not clear, however, whether Ray and Berry use scores computed from Varimax factors for their similarity measures. If this is the case, and the only axes rotated are those corresponding to the f eigenvectors which would otherwise be used to compute distances, then the distances using the f Varimax factor scores will be the same as those derived from the f major component scores. On the other hand, if more than f axes are rotated to a Varimax solution, then more dimensions will usually be required to compute accurate distances since the Varimax rotation does not result in an optimal variance solution as obtained by principal components. The effect of Varimax is to share out the large variance explained by the major components among the lesser components in order to obtain factors which lend themselves to easier interpretation. Distances calculated from the major Varimax factors are still good approximations to those obtained using standard scores, but are less accurate than those derived using principal components loadings.

It is therefore recommended here that, when a reduction in

the number of dimensions used to compute distance coefficients is desired, then factor scores obtained from those eigenvectors associated with the major principal components, which together account for an arbitrary proportion of the overall variance, should be used. A Varimax solution may be obtained as an auxiliary investigation but should not be used in conjunction with classification procedures.

Scatter Diagrams

One of the most useful functions of principal components analysis is that of a diagrammatic tool in the interpretation of cluster structures and relationships. Firstly, we may plot a scatter diagram using any two components as orthogonal axes and their scores as point coordinates (it is customary to plot component I against component II - the principal plane). Supposing that the population of N individuals has been assigned to k disjoint clusters (coded from 1 to k), then instead of plotting stars or crosses on the scatter diagram we may plot cluster codes. Using the principal plane we obtain the best possible 2-dimensional representation of cluster distributions in M -space, and it is often the case that a very large proportion of the overall M -space variance is displayed (see, for example, Appendix Ie). It is now proposed here that each cluster may be represented by a circle in the principal plane whose centre is located at the mean and radius proportional to the square root of the

joint variance of the cluster distribution. An important requirement is that the ratio of the lengths of the axes should be equal to the ratio of the latent roots. If this is not the case, then the displayed interpoint distances will not correspond to the actual distances in space. Thus, for any two components x and y having variances S_x^2 and S_y^2 we require that $u_x/u_y = S_x/S_y$, where u_x and u_y are the lengths of the axes. The circle for cluster t will therefore be centred on the mean (\bar{x}_t, \bar{y}_t) where

$$\bar{x}_t = \frac{1}{k_t} \sum_{i \in t} x_i$$

and have radius $\sqrt{S_{xt}^2 + S_{yt}^2}$ where

$$S_{xt}^2 = \frac{1}{k_t} \sum_{i \in t} (x_i - \bar{x}_t)^2$$

is the variance of component x for the subset of individuals belonging to cluster t . Examples can be found in Figure 1.3.2 and Appendices Ia and Ie.

Molecular Models

The above technique may be extended for the representation of cluster distributions in the principal 3-space. Each cluster is designated by a sphere having centres $(\bar{x}_t, \bar{y}_t, \bar{z}_t)$ and radius $\sqrt{S_{xt}^2 + S_{yt}^2 + S_{zt}^2}$ where x , y and z are the first three principal components. Although this structure cannot easily be drawn, we may treat it as a molecule and construct a 3-dimensional model

which can then be photographed from different angles (e.g. Boyce, 1969). Alternatively, we may use a computer program to plot different views of the molecular-type cluster structure: such a program has been written at St. Andrews by P. G. Adamson in Assembler code for the IBM 1620 (Cole and Adamson, 1969), and is currently being translated into Fortran IV for the IBM 360 series.

Rotating Principal 3-Space

The following method enables a 3-dimensional distribution to be orthogonally projected on to any 2-dimensional plane defined by its normal, and can therefore be used to plot cross-sectional scatter diagrams of principal 3-space (or any other space).

A plane is defined by a normal VC, where $V(v_x, v_y, v_z)$ is the viewpoint and $C(c_x, c_y, c_z)$ the centre of vision. Let (x_i, y_i, z_i) be the coordinates of the i th point in the principal 3-space, then we can transform the origin to V and rotate the axes so that VC coincides with the new Z axis. If we now compute the coordinates (X_i, Y_i, Z_i) of the i th point with reference to the new axes, then (X_i, Y_i) are the point's coordinates in the plane which is orthogonal to VC. It is easy to show (Cole, 1966) that

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = \begin{bmatrix} -(c_y - v_y)/\lambda & (c_x - v_x)/\lambda & 0 \\ -(c_x - v_x)(c_z - v_z)/\lambda & -(c_y - v_y)(c_z - v_z)/\lambda & \lambda \\ c_x - v_x & c_y - v_y & c_z - v_z \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix}$$

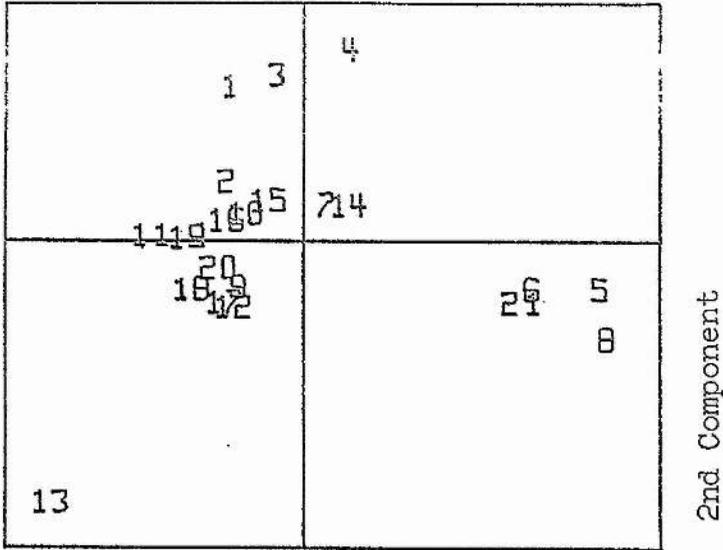
where $\lambda^2 = (c_x - v_x)^2 + (c_y - v_y)^2$. Hence we may write

$$X_i = - (c_y - v_y)x_i/\lambda + (c_x - v_x)y_i/\lambda$$

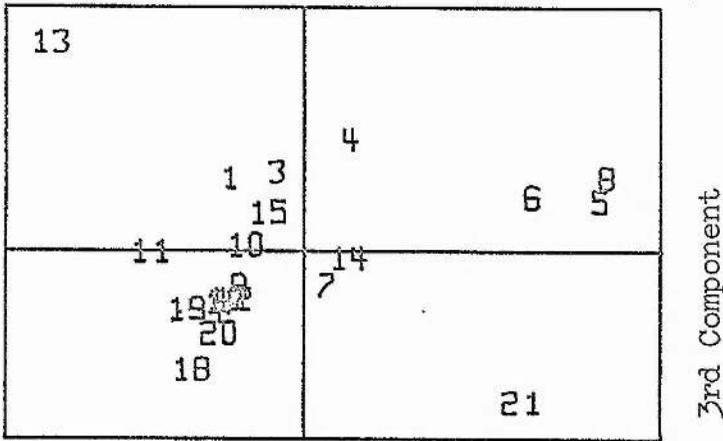
$$Y_i = - (c_x - v_x)(c_z - v_z)x_i/\lambda - (c_y - v_y)(c_z - v_z)y_i/\lambda + z_i\lambda$$

These coordinates hold when $\lambda > 0$; that is, provided that $c_x \neq v_x$ or $c_y \neq v_y$. If $c_x = v_x$ and $c_y = v_y$ then VC has been chosen parallel to the Z-axis. It follows that the x-y plane is already orthogonal to VC, and therefore the required coordinates (X_i, Y_i) are (x_i, y_i) .

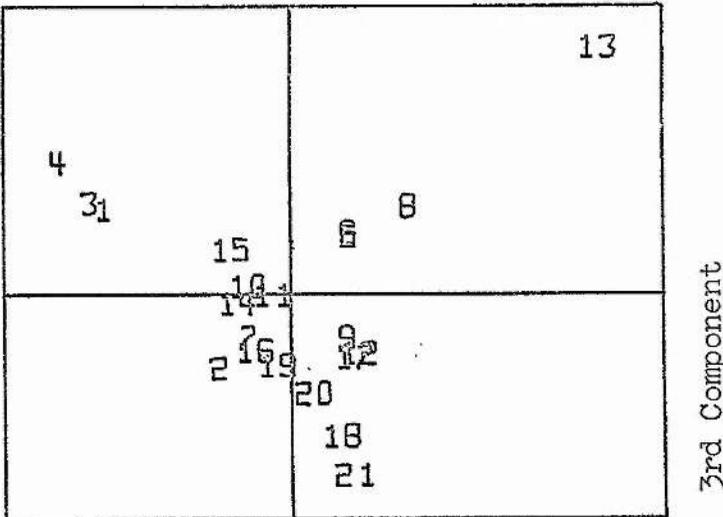
Figure 1.3.1 shows six projections of a 21-point distribution from principal 3-space on to planes orthogonal to VC, where C is the origin of coordinates in each case and V has been assigned to six different points (the coordinates of V are indicated below each diagram). Figure 1.3.1A shows the principal plane (component I versus II) while figures 1.3.1B and 1.3.1C show components I versus III and II versus III, respectively. The other three diagrams are obtained without viewing along an axis. Perhaps the most interesting aspect of these diagrams is that point 21 is seen to be separated from the group (5,6,8), although this is not demonstrated on the principal plane (Figure 1.3.1A). Indeed, it seems that the best principal 3-space groupings are (13) (21) (1,3,4) (5,6,8) and (the rest). Furthermore, figure 1.3.1E indicates these cluster separations to best advantage: the failure of figure 1.3.1A to show the isolation of point 21 can be attri-



(A) $V(0,0,1)$ 1st Component

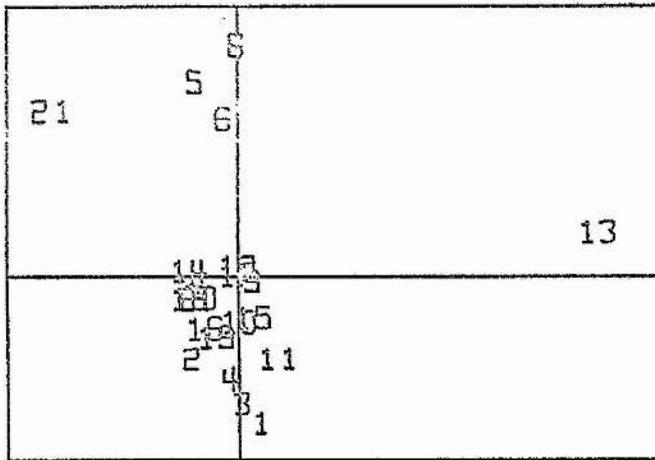


(B) $V(0,1,0)$ 1st Component

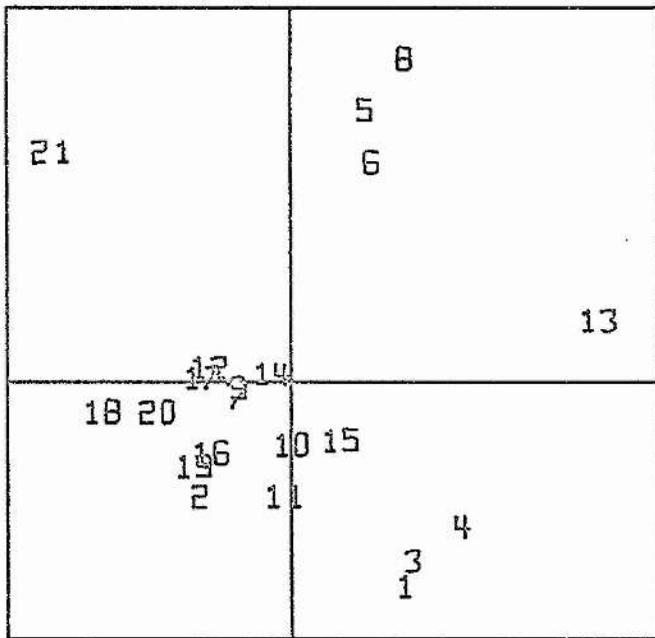


(C) $V(1,0,0)$ 2nd Component

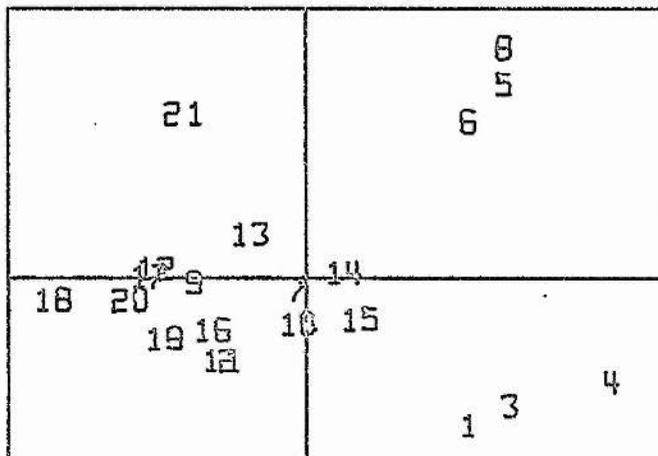
Figure 1.3.1. Six projections of a 21-point distribution in principal 3-space on to planes defined by normals which connect the viewpoint V with the origin $(0,0,0)$.
(continued on page 37)



(D) V(1,1,1)



(E) V(1,1,0)



(F) V(1,1,-1)

Figure 1.3.1. (continued)

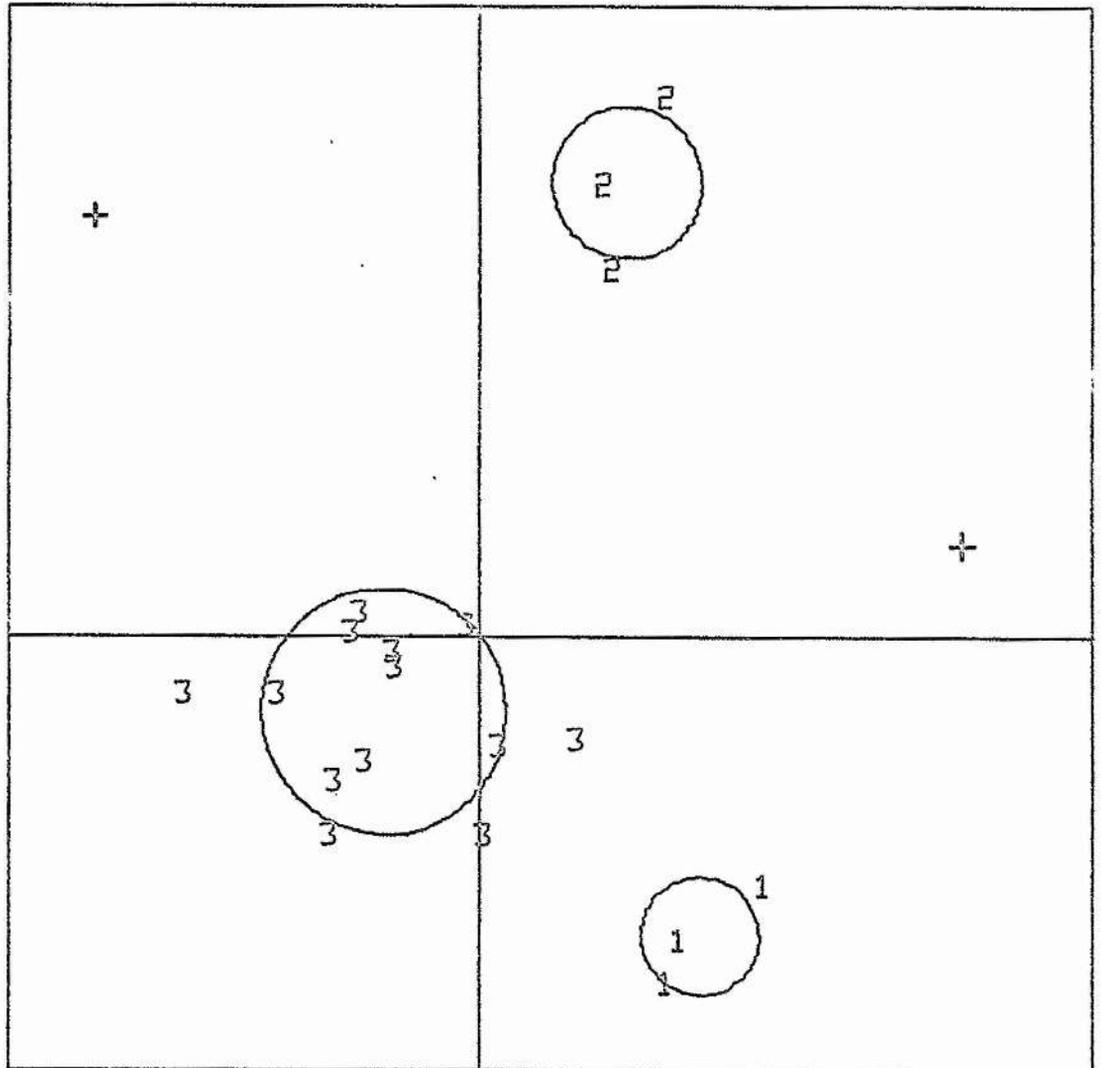


Figure 1.3.2. Scatter diagram showing three clusters extracted from figure 1.3.1(E), with circles plotted proportional to the joint cluster variance.

buted to the overwhelming influence of (the rest) which accidentally causes (21) to be projected on top of (5,6,8). Figure 1.3.2 shows these three clusters plotted on to the plane of figure 1.3.1E, using circles to represent within-group variance (the single objects 13 and 21 have been omitted).

1.4 SIMILARITY COEFFICIENTS

Computational data have previously been described in two modes (binary and continuous), but in order to extend to some group statistics we must also consider the collective data associated with a cluster of individuals. We shall denote the submatrix of $[X_{ij}]$ corresponding to the group of individuals $i \in t$ (a cluster of size k_t) by $[X_{ij}]_t$. Similarly, the corresponding submatrix of the binary matrix $[\alpha_{ij}]$ is denoted by $[\alpha_{ij}]_t$. There is now a considerable temptation to generalise individuals as clusters of size 1, so that all data can be expressed in these two forms; however, this would preclude the simplification of binary similarity coefficients (using the 2 x 2 table - see below), and does not allow group 'disorder' statistics to be treated separately according to whether single individuals or clusters are being compared (since a single individual has zero variance or entropy). We therefore consider the following four data types:

- (1) Continuous individual data $[X_{ij}]$
- (2) Binary individual data $[\alpha_{ij}]$

- (3) Continuous cluster data $[X_{ij}]_t$
- (4) Binary cluster data $[\alpha_{ij}]_t$

Associated with each data type there is a class of statistics. Some of these statistics (for example, distance) appear within more than one class, and some intercluster measures (such as variance) degenerate to constant multiples of distance when used to measure the similarity between two individuals. Table 1.4.1 shows the compatibilities between the data types and statistic classes. The only transformation (Sect. 1.2) that is considered sufficiently general to be adopted is T_{BC} , which assumes a legitimate mapping from the binary attribute space into the metric space of continuous data. It seems that this is a reasonable generalisation since the computation of binary centroids (attribute probability vectors) implies the valid representation of binary data in metric space, and therefore concepts of cluster structure, disorder and entropy may be discussed without regard for the special properties of the raw data.

Some writers choose to distinguish between 'similarity' and 'dissimilarity' coefficients, the distinction being determined by whether the coefficient increases or decreases with increasing similarity. Indeed, it is often the case that a reasonable coefficient is manipulated by its originator in order to reverse the type from similarity to dissimilarity, or vice-

TYPE OF MEASURE

| TYPE OF DATA | INDIVIDUAL | | INTERCLUSTER | |
|----------------------------------|-------------------------------|------------------------|--------------------|-------------------------------------|
| | BINARY (2 x 2) S_{ik} | CONTINUOUS S_{ik} | BINARY S_{pq} | CONTINUOUS S_{pq} |
| BINARY $[\alpha_{ij}]$ | YES | YES (T_{BC}) | YES | YES (T_{BC}) $k_p = k_q = 1$ |
| CONTINUOUS $[x_{ij}]$ | / | YES | / | YES $k_p = k_q = 1$ |
| BINARY CLUSTER $[\alpha_{ij}]_t$ | / | YES (T_{BC}) | YES | YES (T_{BC}) |
| CONTINUOUS CLUSTER $[x_{ij}]_t$ | / | YES | / | YES |

Table 1.4.1. Compatibilities between the four types of computational data and the four classes of similarity measures. Transformation T_{BC} is used to indicate when binary data may be treated successfully as continuous scores; $k_p = k_q = 1$ denotes that cluster statistics may be used with individuals, treating them as groups of size 1.

versa. For example, Gower (1967) uses Sokal's binary matching coefficient $S_{ik} = (A + D)/M$ (see below) in the form $(2(1 - S_{ik}))^{\frac{1}{2}}$ so that it may be treated as a distance statistic. It is simple to show that with almost all methods the square root and addition of a constant make little overall change to the analysis, and serve merely to simplify notation. It is much easier to treat S_{ik} as a particular instance of a 'similarity coefficient', and in fact, the term 'similarity' will be used hereafter to denote all coefficients, including those of type 'dissimilarity'. It is assumed therefore that appropriate computation tests will be reversed for dissimilarities.

We shall now review several similarity coefficients, stating the formulae without further elaboration. For general surveys of similarity measures the reader is referred to Ball (1966), Boyce (1969), Lance and Williams (1966b), Orloci (1968c) and Sokal and Sneath (1963); all references to relevant papers are made against each statistic, together with the 'CODE' of the coefficient within the 'CLUSTAN' suite of computer programs (Chapter 10).

Continuous Individual Data

Let $(X_{i1}, \dots, X_{ij}, \dots, X_{iM})$ be the continuous data scores for the i th individual, then the similarity between two individuals i and k is defined, using summations for $j = 1$ to M , as follows:

Distance:
$$d_{ik}^2 = \frac{1}{M} \sum (X_{ij} - X_{kj})^2$$

CLUSTAN CODE 1

(Ball, 1966; Boyce, 1969; Gower, 1967; Jancey, 1966; Lance and Williams, 1966b, 1966c, 1967a; Macnaughton-Smith, 1965; Orloci, 1966, 1968c; Sneath, 1966; Thorndike, 1953; Williams et al, 1966; see also Sect. 1.1 and Sect. 1.2)

Correlation:
$$\frac{M \sum X_{ij} X_{kj} - \sum X_{ij} \sum X_{kj}}{\sqrt{[M \sum X_{ij}^2 - (\sum X_{ij})^2][M \sum X_{kj}^2 - (\sum X_{kj})^2]}}$$

CLUSTAN CODE 3

(Ball, 1966; Boyce, 1969; Lance and Williams, 1966b, 1966c, 1967a; Orloci, 1966, 1967a, 1968c; Sokal and Michener, 1958; Williams et al, 1966)

Similarity Ratio:
$$\frac{\sum X_{ij} X_{kj}}{\sum X_{ij}^2 - \sum X_{ij} X_{kj} + \sum X_{kj}^2}$$

CLUSTAN CODE 28

(Ball, 1966; Rogers and Tanimoto, 1960)

Dot Product:
$$\frac{1}{M} \sum X_{ij} X_{kj}$$

CLUSTAN CODE 26

(Ball, 1966; Orloci, 1967a, 1968c)

Cosine or Normalised Correlation:

CLUSTAN CODE 27
$$\frac{\sum X_{ij} X_{kj}}{\sqrt{\sum X_{ij}^2 \sum X_{kj}^2}}$$

(Ball, 1966; Boyce, 1969; Orloci, 1967b)

Size Difference: $\frac{1}{M^2} (\sum X_{ij} - \sum X_{kj})^2$

CLUSTAN CODE 29

(Boyce, 1969; Penrose, 1954; Sokal and Sneath, 1963)

Shape Difference:

$$\frac{1}{M} \sum (X_{ij} - X_{kj})^2 - \frac{1}{M^2} (\sum X_{ij} - \sum X_{kj})^2$$

CLUSTAN CODE 30

(Boyce, 1969; Penrose, 1954; Sokal and Sneath, 1963)

Dispersion: $\frac{1}{M} \sum (X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k)$

CLUSTAN CODE 32

(Orloci, 1966, 1967a) where $\bar{X}_i = \frac{1}{M} \sum_j X_{ij}$

Nonmetric or Canberra Metric:

CLUSTAN CODE 36
$$\frac{\sum |X_{ij} - X_{kj}|}{\sum (X_{ij} + X_{kj})}$$

(Lance and Williams, 1966b, 1966c, 1967a, 1967b; Williams et al, 1966)

Binary Individual Data

The data associated with an individual i is represented by the vector $\underline{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ij}, \dots, \alpha_{iM})$ where $\alpha_{ij} = 1$ or 0 according to whether i possesses or lacks attribute j . In order to compare two individuals i and k we construct the following 2 x 2 table:

$$\begin{array}{c}
 \alpha_k \\
 + \\
 - \\
 \end{array}
 \begin{array}{c}
 \alpha_i \\
 + \\
 - \\
 \end{array}
 \begin{array}{|c|c|}
 \hline
 A & B \\
 \hline
 C & D \\
 \hline
 \end{array}$$

where A is the number of attributes possessed by both i and k, B is the number possessed by k but lacked by i, and so on. We may write

$$\begin{aligned}
 A &= \sum \alpha_{ij} \alpha_{kj} \\
 B &= \sum (1 - \alpha_{ij}) \alpha_{kj} \\
 C &= \sum (1 - \alpha_{kj}) \alpha_{ij} \\
 D &= \sum (1 - \alpha_{ij})(1 - \alpha_{kj}) \\
 A+B &= \sum \alpha_{kj} \\
 A+C &= \sum \alpha_{ij} \\
 A+B+C+D &= M
 \end{aligned}$$

It seems that the 2 x 2 binary table contains all the information that we need to know about a pair of individuals i and k, for we may now write the similarity coefficients previously defined for continuous data as simple functions of these cell counts:

Distance:
$$d_{ik}^2 = \frac{1}{M} \sum (\alpha_{ij} - \alpha_{kj})^2 = \frac{B+C}{M}$$

CLUSTAN CODE 2

Correlation:
$$\frac{AD - BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}}$$

CLUSTAN CODE 18

Similarity Ratio: $\frac{A}{A+B+C}$

CLUSTAN CODE 5

Dot Product: $\frac{A}{M}$

CLUSTAN CODE 13

Cosine or Normalised Correlation:

CLUSTAN CODE 16 $\frac{A}{\sqrt{(A+B)(A+C)}}$

Size Difference: $\left(\frac{B-C}{M}\right)^2$

CLUSTAN CODE 20

Shape Difference: $\frac{M(B+C) - (B-C)^2}{M^2}$

CLUSTAN CODE 31

Dispersion: $\frac{AD - BC}{M^2}$

CLUSTAN CODE 33

Nonmetric or Canberra Metric:

CLUSTAN CODE 37 $\frac{B+C}{2A+B+C}$

Sokal and Sneath (1963) review the following additional coefficients defined from the cell counts, which are therefore limited to binary individual data:

Simple Matching Coefficient: $\frac{A+D}{M}$

CLUSTAN CODE 4

CLUSTAN CODE 6 $\frac{2A}{2A+B+C}$

CLUSTAN CODE 7 $\frac{2(A+D)}{2(A+D) + (B+C)}$

CLUSTAN CODE 8 $\frac{A}{A + 2(B+C)}$

CLUSTAN CODE 9 $\frac{A+D}{(A+D) + 2(B+C)}$

CLUSTAN CODE 10 $\frac{A}{B+C}$

CLUSTAN CODE 11 $\frac{A+D}{B+C}$

CLUSTAN CODE 12 $\frac{(A+D) - (B+C)}{M}$

CLUSTAN CODE 14 $\frac{1}{2}(\frac{A}{A+B} + \frac{A}{A+C})$

CLUSTAN CODE 15 $\frac{1}{4}(\frac{A}{A+B} + \frac{A}{A+C} + \frac{D}{B+D} + \frac{D}{C+D})$

CLUSTAN CODE 17 $\frac{AD}{\sqrt{(A+B)(A+C)(B+D)(C+D)}}$

CLUSTAN CODE 19 $\frac{AD - BC}{AD + BC}$

Pattern Difference: $\frac{BC}{M^2}$

CLUSTAN CODE 21 (Sneath, 1968)

Continuous Cluster Data

Let cluster t contain k_t individuals, then we define the centroid of t as $\underline{U}_t = (U_{1t}, \dots, U_{jt}, \dots, U_{Mt})$, where

$$U_{jt} = \frac{1}{k_t} \sum_{i \in t} X_{ij}$$

The total variance S_t^2 for cluster t is given by

$$S_t^2 = \frac{1}{Mk_t} \sum_j \sum_{i \in t} (X_{ij} - U_{jt})^2$$

All of the continuous coefficients stated previously may now be used to measure the similarity between two clusters by representing the clusters by their centroids. That is, we replace each X_{ij} and X_{kj} by U_{jp} and U_{jq} in order to obtain the intercluster similarity S_{pq} .

In addition we have the following special measures of intercluster similarity which take into account within-group structures:

Error Sum of Squares:

CLUSTAN CODE 24 Let $E_t = \sum_{i \in t} \sum_j (X_{ij} - U_{jt})^2$ be the error sum of squares for cluster t , then $S_{pq} = E_{p+q} - E_p - E_q = d_{pq}^2 \frac{k_p k_q}{(k_p + k_q)}$

(Beale, 1969; Bolshev, 1969; Calinski, 1969; Calinski and Harabasz, 1970; Forgey, 1965; Gower, 1967; Jancey, 1966; Orloci, 1967b; Ward, 1963; Wishart, 1969c; see also Sect. 2.2 and Sect. 8.4)

Average Distance:

CLUSTAN CODE 22

$$\frac{1}{M k_p k_q} \sum_{i \in p} \sum_{k \in q} \sum_j (x_{ij} - x_{kj})^2$$

$$= d_{pq}^2 + S_p^2 + S_q^2$$

(Ball and Hall, 1966; Thorndike, 1953; Wishart, 1969e)

Variance:

CLUSTAN CODE 34

$$\frac{1}{M} S_{p+q}^2$$

(Ball and Hall, 1966)

Binary Cluster Data

Suppose that for a cluster t comprising k_t individuals, attribute j is possessed by f_{jt} individuals, then we define the probability associated with the j th attribute as

$$p_{jt} = \frac{1}{k_t} f_{jt} = \frac{1}{k_t} \sum_{i \in t} \alpha_{ij}$$

Thus the centroid of cluster t is given by

$$\underline{p}_t = (p_{1t}, \dots, p_{jt}, \dots, p_{Mt})$$

The variance S_t^2 of cluster t is now easily shown to be

$$S_t^2 = \frac{1}{M} \sum_j p_{jt}(1 - p_{jt})$$

We can apply binary cluster data to the continuous coefficients by using, firstly, the centroid \underline{p}_t as a point representing cluster t (with the coefficients for continuous individual data), and secondly, using \underline{p}_t and S_t^2 with the three continuous intercluster

similarity criteria. The CLUSTAN CODES for the binary versions of these similarity measures are

Error Sum of Squares: CLUSTAN CODE 25

Average Distance: CLUSTAN CODE 23

Variance: CLUSTAN CODE 35

In addition, we define the 'information content' of cluster t as:

$$I_t = Mk_t \log k_t - \sum [f_{jt} \log f_{jt} + (k_t - f_{jt}) \log (k_t - f_{jt})]$$

whence the increase in information arising from the fusion of two clusters p and q is obtained from:

Information Gain:

CLUSTAN CODE 40

$$\Delta I_{pq} = I_{p+q} - I_p - I_q$$

(Hyvarinen, 1962; Lambert and Williams, 1966; Lance and Williams, 1966b, 1966c, 1967a, 1967c, 1968; Macnaughton-Smith, 1965; Orloci, 1968a, 1968c, 1969; Williams et al, 1966; see also Sect. 2.2, Sect. 3.1 and Sect. 4.2)

CHAPTER 2: HIERARCHIC FUSION

2.1 GENERAL PROCESS

One of the earliest forms of hierarchical clustering (Sneath, 1957; Sokal and Sneath, 1963) was a grouping procedure associated with some similarity 'sorting level' or 'threshold' chosen by the user. For example, with Sneath's method (single linkage - Sect. 2.2) an individual joins a group if the highest similarity between the entrant and any member of the group is not less than the chosen threshold S^* ; also, two groups are joined if the highest similarity between the groups (i.e. between two individuals, one from each group) satisfies the same test. It is easily shown that, for an agreed similarity matrix and threshold S^* , methods can be devised for group-forming by this criterion which all produce the same final and unique solution.

By contrast, Sørensen's method (complete linkage - Sect. 2.2) clusters individuals or groups by requiring the least inter-group similarity to equal or exceed the threshold S^* , and cannot be given a procedure which produces a unique result. Sørensen (1948) suggests several criteria for determining the choice of fusion when a dichotomous situation arises (due to equal least similarities occurring for an individual with two or more groups): the first standby decision attaches the entrant to the largest of the groups; if the groups all have the same size, then admission is determined by the highest average similarity. As Sokal and

Sneath (1963) point out, the resultant clustering will vary depending on the order of the initial fusions.

One solution (Sokal and Michener, 1958) to the problem is to select monotonic decreasing thresholds chosen at regular short intervals, and to use the clustering obtained at each level as initial grouping for the next. This has the effect of reducing the number of fusions that occur at any one level, so that they are well-ordered. The first disadvantage of this technique is that if the intervals between successive thresholds are too short, then it may happen that no additional fusions occur from one level to the next, in which case the step is totally unproductive. Secondly, the choice of threshold must be made by the user, who, if he is a casual user, will require guidance in his selection. These considerations lead naturally to the hierarchical or 'agglomerative' algorithm which eliminates the choice of threshold completely, and can be generalised as follows:

- 1) Define a criterion of similarity $S(p,q)$ between two groups p and q , where the groups may contain one or more individuals.
- 2) Start with N groups (each comprising a single individual), and compute the between-group similarities $S(i,j)$ - otherwise called the similarity matrix (Sect. 1.4).
- 3) Fuse those two groups p and q which are most similar. On the first cycle, these will be the two most similar individuals in the entire population. The structure of the new group is then

resolved, and the new similarities $S(p+q,r)$ between $(p+q)$ and all other groups r are calculated.

4) Return to 3 and continue fusing groups successively until $N-1$ cycles have been performed.

We observe that Sneath's and Sørensen's methods are obtained when $S(p,q)$ is defined as the highest and lowest between-cluster similarity, respectively. Step 3 of the algorithm effectively chooses the similarity threshold from the highest $S(p,q)$ available - we must, therefore, now query what happens when there are two or more identically highest $S(p,q)$'s available. It seems that this is a question which cannot be given a generalised answer, since such a solution would (a) have to be tailored to meet the characteristics of each particular method, and (b) would be an intuitive decision anyway (for example, Sørensen's). In practice, we can say that for truly continuous data it is highly unlikely that equal similarities, let alone equal highest similarities, will occur. Secondly, most methods perform some further computation on elements of the similarity matrix (e.g. average linkage and centroid) so that the dichotomy is less likely to occur during later stages of these analyses. Thirdly, if we have a choice between two highest similarities $S(p_1,q_1)$ and $S(p_2,q_2)$ then, from within-group homogeneity considerations, we may compare the 4 and 2-cluster levels: the order of fusions in between is not important. Finally, if the choice is between $S(p_1,q)$ and $S(p_2,q)$ then we must decide which

way q should go. In this case, the new similarity $S(p_1+q, p_2)$ determines the characteristics required of the clusters - Sneath's method would set $S(p_1+q, p_2) = S(p_1, q)$ by definition, while Sørensen's would almost certainly set $S(p_1+q, p_2) < S(p_1, q)$. It would appear that, any arbitrary decision such as the first cluster pair with highest similarity which is encountered is as good a general solution as any other.

2.2 METHODS

Linkage Techniques

(1) Single linkage. The criterion of the similarity between two clusters is defined as the highest similarity between two individuals, one from each cluster. This method, which is generally attributed to Sneath (1957; see also Sokal and Sneath, 1963) has evidently been proposed independently by McQuitty ('Linkage analysis', 1957; 1961; 1967a) and Gengerelli (1963), and is also associated with minimum spanning tree techniques (Florek, et al, 1951; Gower and Ross, 1969). The method is well-known for its 'chaining' effect (Forgey, 1964, 1965; Needham, 1965a; Williams, et al, 1966; Hodson, et al, 1966; Lance and Williams, 1967a; Jardine and Sibson, 1968; Shepherd and Willmott, 1968; see also Chapter 6) which produces long straggling clusters. This is generally considered to be undesirable, especially with large populations for which the method tends to isolate the distribution core as one cluster and single peripheral individuals as the others (see, for example, figure 2.3.1).

The hierarchical algorithm is developed by several writers (Williams, et al, 1966; Johnson, 1967) who select, at each fusion, those two clusters which contain the closest pair of individuals or the 'nearest neighbours'. The hierarchical algorithm is sometimes referred to as 'nearest neighbour', and it is simple to show that it derives all the N possible groupings which can be obtained with Sneath's original algorithm using any threshold.

(ii) Complete linkage. A group of individuals comprises a cluster provided that no two individuals have a similarity which is less than the threshold (Sørensen, 1948; Sokal and Sneath, 1963). This is the exact opposite of single linkage in the sense that the farthest neighbours must satisfy the similarity criterion; when d^2 is used, spherical or 'tight' clusters are obtained. Macnaughton-Smith (1965), McQuitty ('Syndrome analysis', 1966a) and Johnson (1967) evidently suggest the hierarchical algorithm whereby two clusters are fused if the resulting least similarity between pairs of members is greatest. That is, using d^2 the diameter of the resulting cluster must be minimum. The method depends for its fusion decision on the vagaries of pairs of points, and is therefore rather unstable; furthermore, the diameter constraint is probably too severe, and sometimes a type of chaining is observed (Wishart, 1969d; Crawford, et al, 1970; see Appendices Ia and Id).

(iii) Average linkage. Sokal and Michener (1958), with their 'unweighted variable group' method, were evidently the first to take

into account group structure in clustering. Using product-moment correlation coefficients to measure the similarities between individuals, they define the similarity between two clusters as the average of all the similarities between pairs of individuals, one from each cluster. The hierarchical method is proposed by Ray and Berry (1965), Lance and Williams ('group average', 1966a, 1967a), and McQuitty ('similarity analysis', 1966b), and the concept of average linkage as a compromise between the single and complete linkage extremes is discussed by Sokal and Sneath (1963), Hodson, et al (1966), Proctor (1966) and Sneath (1966a). Average linkage is also used to augment single linkage as a de-chaining (Shepherd, 1966; Shepherd and Willmott, 1968) and counter-chaining (Carmichael, et al, 1968) mechanism. On the whole, the method seems to behave well (see Appendix Ic); however, the hierarchical algorithm has been known to chain with very large populations (e.g. Wishart, 1969d; Crawford, et al, 1970).

(iv) Median linkage. As an alternative compromise between the single and complete linkage extremes, Kendrick and Proctor (1964) propose a median linkage method which they say is 'easier than the mean¹ and unaffected by outlying values'; Proctor (1966) further claims that 'in the absence of a computer program it is easier to obtain than the arithmetic average¹'. The similarity

¹ centroid sorting - see next paragraph.

between two clusters is defined as that similarity between two individuals, one from each cluster, which represents the median of all between-cluster links; that is, one-half² of the inter-cluster similarities are less than the median. This method will behave very much like average linkage, but is considerably more difficult to programme (despite the authors' claims). At the fusion of two clusters, the similarities between all other clusters and the new group must be obtained from a search of each submatrix of the similarity matrix which contains the between-group similarities for a cluster pair. Each search must also include an ordering mechanism to isolate the new median, which must then be stored elsewhere (the similarity matrix has to be retained in full). By contrast, average linkage has a very nice 'combinatorial solution' - see Sect. 8.1.

(v) Proportional Link linkage (Sneath, 1966a). As its name implies, proportional link linkage would combine two clusters if a specified proportion ϕ of the between-cluster similarities exceeded a chosen threshold (a similar suggestion is made by Shepherd and Willmott, 1968). The hierarchical procedure would require that the similarity $S_{pq}(\phi)$ between two clusters be defined

²if either of the clusters has an even number of members, then there will be an even number $2r$ of between-cluster similarities. For convenience we shall adopt the $(r+1)$ th highest similarity as the median in this case.

as that between-cluster similarity which is the $(\phi k_p k_q)$ th member of the ordered list of between-cluster coefficients, where k_p, k_q are the cluster sizes, and $\phi k_p k_q$ is a rounded-up¹ integer. Hence $\phi = 0, \frac{1}{2}$ and 1 exactly reproduce single, median and complete linkage. Fusion would be defined for those clusters p and q for which $S_{pq}(\phi)$ is greatest. Although theoretically nice, the method unfortunately suffers the same computational disadvantages of median linkage, and does not seem to have been programmed or used.

Centroid Sorting

One very attractive generalisation of the hierarchical fusion method is 'centroid sorting', for which a group of one or more individuals is represented by a point located at the group's mean or centroid. This concept permits us to compare two groups in terms of any quantitative similarity criterion (see Chapter 1): for example, we may use the distance separating their centroids as measure of similarity, or the cosine of the angle between two lines connecting the origin with the centroids. Obviously the cosine criterion is dependent, while distance independent, on the position of the origin, and therefore centroid sorting permits us to compare very different similarity criteria within the same clustering framework (see Chapter 5).

¹ the term 'rounded-up' is used here in the special sense that for any value $K \leq \phi k_p k_q < (K+1)$, we choose the number $(K+1)$, excepting the case when $\phi \equiv \frac{1}{2}$, where we use $k_p k_q$.

Sokal and Michener (1958) seem to have been the first to adopt centroid sorting, naming it the 'weighted variable-group' method, and in this instance they use the product-moment correlation coefficient as similarity criterion.

In general, with the hierarchical fusion algorithm, we fuse two clusters p and q and then compute the centroid or mean of the new cluster distribution: then the similarities $S_{pq,r}$ between all other clusters r and the new cluster (p+q) are computed using the two cluster centroids as if they were single individuals. The next fusion is then indicated for those two clusters having highest similarity, and the process is repeated N - 1 times.

Ward's Method

Ward (1963) proposed a method for hierarchical fusion which is probably one of the most used procedures of its kind, particularly in the social sciences. The 'disorder' within a cluster is measured by the sum of the squared distances of the points from the cluster mean; hence if X_{ijt} is the value of the jth variable for the ith point of cluster t, which contains k_t points, then

$$E_t = \sum_{i=1}^{k_t} \sum_{j=1}^M (X_{ijt} - U_{tj})^2$$

where U_{tj} is the mean of the jth variable for cluster t. The total 'error sum of squares' E is then defined by Ward as the sum of the E_t values for all T clusters -

$$E = \sum_{t=1}^T E_t$$

With hierarchic fusion, two clusters p and q are chosen for fusion in order to minimise E : that is, they are fused if the increase in E

$$I_{pq} = E_{p+q} - E_p - E_q$$

is minimum. This method is independently proposed by Orloci (1967b), and E is considered by Edwards and Cavalli-Sforza (1965) in their exhaustive polythetic divisive method (Sect. 3.2), and Beale (1969) for iterative relocation (Sect. 4.2). Wishart (1969c) found the combinatorial transformation for I_{pq} , and proposed an efficient computer algorithm for this and other hierarchical methods (see Sect. 8.4). It is fairly straightforward to prove (Sect. 8.2) that

$$I_{pq} = d_{pq}^2 k_p k_q / (k_p + k_q)$$

(equation 8.2.4), where k_p , k_q are cluster sizes, and d_{pq}^2 is the distance between the cluster centroids. Using this form, the method can be included in the group of 'centroid sorting' techniques, where clusters are represented by their centroids and the similarity criterion is I_{pq} , as stated above.

Gower's Median

In introducing his median strategy, Gower (1967) writes:

"In the geometrical interpretation we assume that if S_{ij}

is a similarity between individuals i and j , then the distance between their point representations p_i and p_j is $\left[2(1 - S_{ij})\right]^{\frac{1}{2}}$. Gower (1966) has shown that the latent vectors of the similarity matrix, scaled so that the sum of squares of the elements of the r th vector is equal to the r th latent root, gives directly a set of coordinates with this distance property."

In fact, Gower (1966) considers precisely one definition of S_{ij} , namely Sokal's matching coefficient (Chapter 1)

$$S_{ij} = \frac{A + D}{M} = 1 - \frac{B + C}{M} = 1 - d_{ij}^2$$

where d_{ij}^2 is the binary squared distance measure. Hence for the matching coefficient the distance property holds. Gower (1967) then goes on to define the combinatorial solution (see Sect. 8.1) for the fusion of two clusters p and q by centroid sorting as

$$S_{pq,r} = S_{pr} k_p / (k_p + k_q) + S_{qr} k_q / (k_p + k_q) + (1 - S_{pq}) k_p k_q / (k_p + k_q)^2 \quad (2.2.1)$$

which is identical with the centroid sorting combinatorial formula (Sect. 8.1) on substitution of $1 - d_{pq}^2$ for each S_{pq} . At this point, in connection with an involved discussion of weighting schemes, Gower suggests that we may "wish to give each cluster unit weight, regardless of the number of individuals in it" from which he deduces the alternative combinatorial formula

$$S_{pq,r} = \frac{1}{2}(S_{pr} + S_{qr}) + \frac{1}{4}(1 - S_{pq}) \quad (2.2.2)$$

and when $1 - d_{pq}^2$ is substituted for the matching coefficient S_{pq} , this is the same as

$$d_{pq,r}^2 = \frac{1}{2}(d_{pr}^2 + d_{qr}^2) - \frac{1}{4} d_{pq}^2 \quad (2.2.3)$$

Formula (2.2.3) is correctly interpreted (for distances only) by Lance and Williams (1967a) as a 'median' strategy, in the sense that the new cluster formed by the fusion of p with q is assigned to a point midway between the points representing p and q (the median of the line which connects p with q), regardless of cluster sizes. Formula (2.2.2) will produce an identical fusion hierarchy when S_{pq} is the matching coefficient: however, in deriving (2.2.1) and (2.2.2) Gower (1967) appears to generalise the strategy for all similarity coefficients. The point should be made that Gower's 'median' and 'centroid' sorting strategies, as defined geometrically above, are only obtained with the distance measures d_{pq}^2 or $(B + C)/M$, or the complement $(A + D)/M$; any other similarity coefficient will either not satisfy the geometrical interpretations of (2.2.1) and (2.2.2), or will require additional proof (see, for example, Sect. 8.3).

This dichotomous situation is best resolved by adopting equation (2.2.3) in connection with the distance statistics d_{ij}^2 and $(B + C)/M$, these being the only measures discussed here which satisfy the geometrical interpretation of Gower's median method.

Information Statistic

The 'information' and 'information gain' statistics I and ΔI are generally introduced (Hyvarinen, 1962; Macnaughton-Smith, 1965; Williams, et al, 1966; Lance and Williams, 1966b; Orloci, 1968a, 1968c, 1969a, 1969b) for binary data, thus:

Shannon (1948) defined the quantity 'information' for a finite discrete probability function taking R states by adopting the entropy function H (Tolman, 1938; Brillouin, 1962) as a measure of the 'disorder' of a system. If p_r is the probability associated with the r th state, then entropy is given by

$$H = - \sum_{r=1}^R p_r \log p_r$$

For classification purposes, it is usual to consider the case when there exist only two possible states (presence and absence) of a binary attribute j - however, Hyvarinen (1962) and Orloci (1968a, 1968c, 1969b) continue with the general case of R_j states associated with each multistate character j in order to adapt the binary result to semiquantitative data (such as species density counts within stands). Hence we can write

$$H_j = - \left[p_j \log p_j + (1 - p_j) \log (1 - p_j) \right]$$

for the binary case. Clearly, when $p_j \rightarrow 1$, $H_j \rightarrow 0$ since $\log p_j \rightarrow 0$, and similarly when $p_j \rightarrow 0$, $H_j \rightarrow 0$; in fact, the value of H_j achieves a maximum at $p = \frac{1}{2}$ (Shannon, 1948). If, in a classification process

we derive a group of individuals for which the j th attribute is either completely absent or completely present, then p_j will be 0 or 1 respectively, and we conclude that the group is well-defined for that attribute.

This statistic is further generalised by Shannon for Markoff chain processes to the case when there are M events j each having entropy H_j , so that the total disorder of the system may be measured by the total entropy (or average information)

$$\bar{H} = \sum_{j=1}^M H_j$$

which, for 2-state data, reduces to

$$\bar{H} = - \sum_{j=1}^M \left[p_j \log p_j + (1 - p_j) \log (1 - p_j) \right]$$

By introducing the factor n (group size) we obtain the 'working formula' for total information

$$I = n\bar{H} = Mn \log n - \sum_{j=1}^M \left[f_j \log f_j + (n - f_j) \log (n - f_j) \right]$$

where the f_j 's are attribute frequencies ($f_j = p_j n$).

In the context of the hierarchical fusion process for binary data, three variants of these 2-state statistics are used to measure dissimilarity. MacArthur and MacArthur (1961) have adopted total entropy \bar{H} ($= I/n$) to measure diversity, while Lambert and Williams (1966) use I . In either case, we fuse those two clusters

whose resulting \bar{H} or I is minimum. Alternatively, we can define 'information gain' ΔI at fusion (Macnaughton-Smith, 1965; Williams, et al, 1966; Lance and Williams, 1966b) as

$$\Delta I_{pq} = I_{p+q} - I_p - I_q$$

and combine those two groups whose ΔI is minimum. Since the p_j 's constitute centroid coordinates, all three variants of the information statistic may be included within the 'centroid sorting' category, since each group at fusion is represented by its centroid. It is for this reason that Williams et al (1966) and others describe 'information analysis' as another variant of centroid.

Kullback (1959) has evidently established a relationship between χ^2 and these functions under random sampling, and Macnaughton-Smith (1965) has compared I with $\sum \chi^2$ (see also Association analysis, Sect. 3.1). Using the χ^2 relationship, Lambert and Williams (1966) have set up a null-hypothesis test of confidence whereby fusion is terminated when $2\Delta I \leq \chi^2$ (M d.f.). However, it has been generally admitted that this significance test is conservative, and unreliable when used in the context of monothetic division, particularly when M is large (Lambert and Williams, 1966; see also Sect. 3.1). Although Lance and Williams (1966b) appear to be fairly satisfied with the test when used for hierarchic fusion, it would appear that frequently too many final groups are indicated. For instance, with the 450 quadrat x 37

species Andean survey data (Crawford, et al, 1970 - Appendix Ia), 35 clusters were indicated by the significance test at $p = 0.01$. Clearly, this test needs further improvement, and cannot be confidently used in its present form¹.

In practice, information appears to behave very much like the error sum of squares, and at the present time the only realistic approach to determining a cut-off point on the hierarchy is to look for large relative 'jumps' in the fusion coefficient and then examine the prior grouping for 'meaningful' clusters. Sneath (1969) has noted that the information statistic dislikes small groups when large clusters are around: it is readily seen that a single peripheral individual, if grouped to a large cluster, is unlikely to modify the p_j 's extensively, regardless of its attribute structure. Hence, the statistic can be accused of tending to force clusters of equal size (see also figure 2.3.2).

2.3 DISCUSSION

One of the most attractive features of the generalised hierarchical algorithm is that the entire procedure can be represented by a 'dendrogram' or 'linkage tree'. Every individual is

¹Orloci (1968a, 1968c, 1969b) merely mentions a relationship between 2I and χ^2 established by Kullback (1959), but does not adopt the significance test, and Lance and Williams (1968) admit that the number of degrees of freedom is usually very large, and the test of significance correspondingly weak.

allocated a node (point) on a graph, and each fusion conveyed by connecting the two branches associated with the fused groups. These connections are usually drawn parallel to points on a coefficient scale which correspond to the fusion coefficient values, so that large jumps in the coefficient can be readily observed. An example is shown in figure 2.3.2(B5) where the large jump from the 2 to 1 cluster level could be 'interpreted' as the transition from a 'well-ordered' to 'disordered' classification.

This device is used by several writers (Sneath, 1966a; Williams, et al, 1966; Lance and Williams, 1966b, 1967a) for the visual comparison of hierarchical methods. Figures 2.3.1 and 2.3.2 are used by Williams et al (1966) to compare single linkage with centroid sorting for five different similarity criteria. The most striking aspect of figure 2.3.1 (single linkage) is the consistent 'chaining' effect throughout, while figure 2.3.2 (centroid sorting) shows the phenomenon known as 'coefficient reversals', for which the fusion coefficient values are not always monotonic decreasing. Another important feature of centroid sorting (shown in figure 2.3.2) is that different similarity criteria often produce very different results. We see that the information gain statistic (B5 of 2.3.2) appears to produce (force ?) a nicely nested hierarchy within which variation of cluster size seems to be reduced, while the other criteria show differing ten-

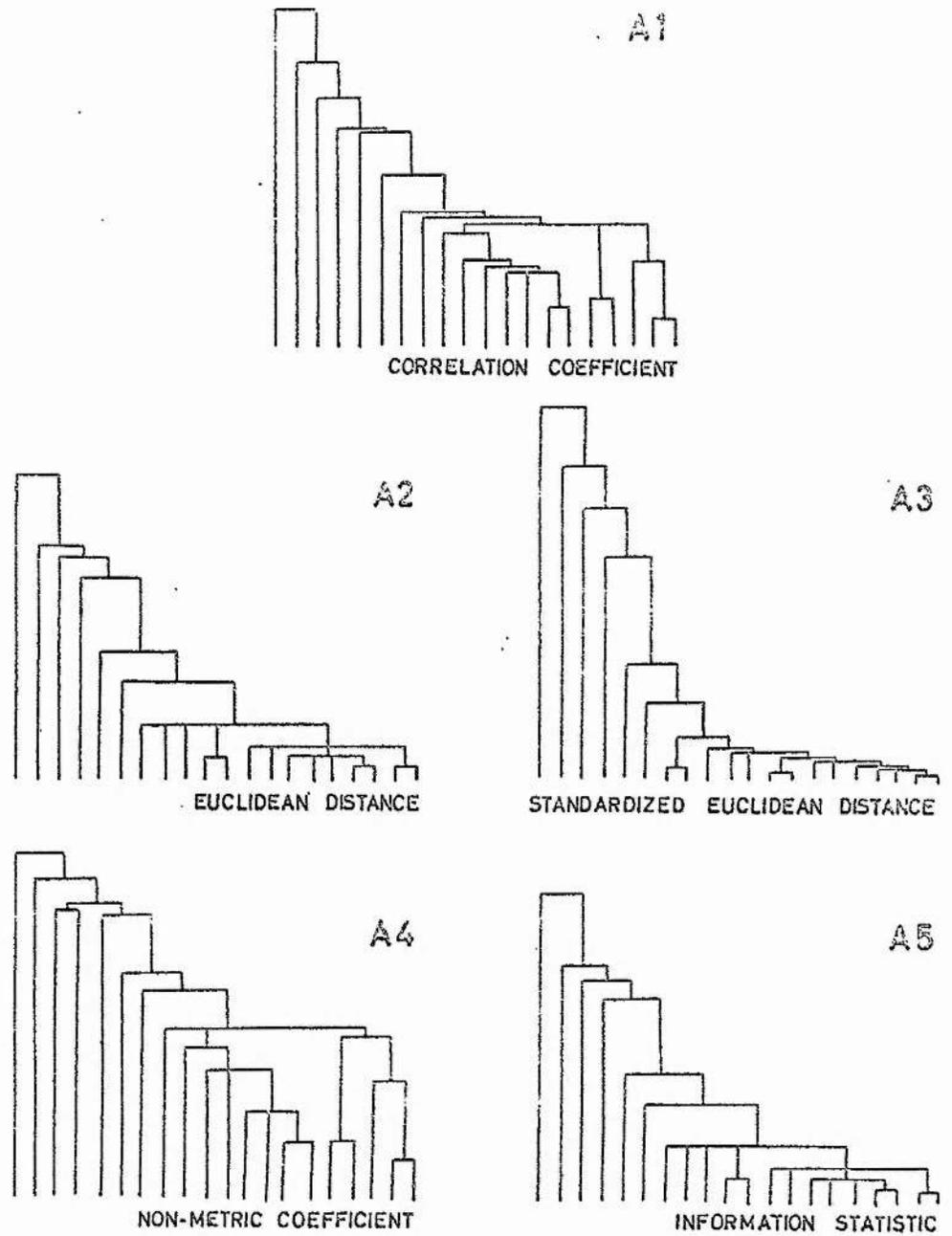


Figure 2.3.1. Dendrograms for single linkage using five different similarity criteria with a 20 quadrat ecological survey (Williams, et al, 1966). Reproduced with kind permission of the authors and the editor of the Journal of Ecology.

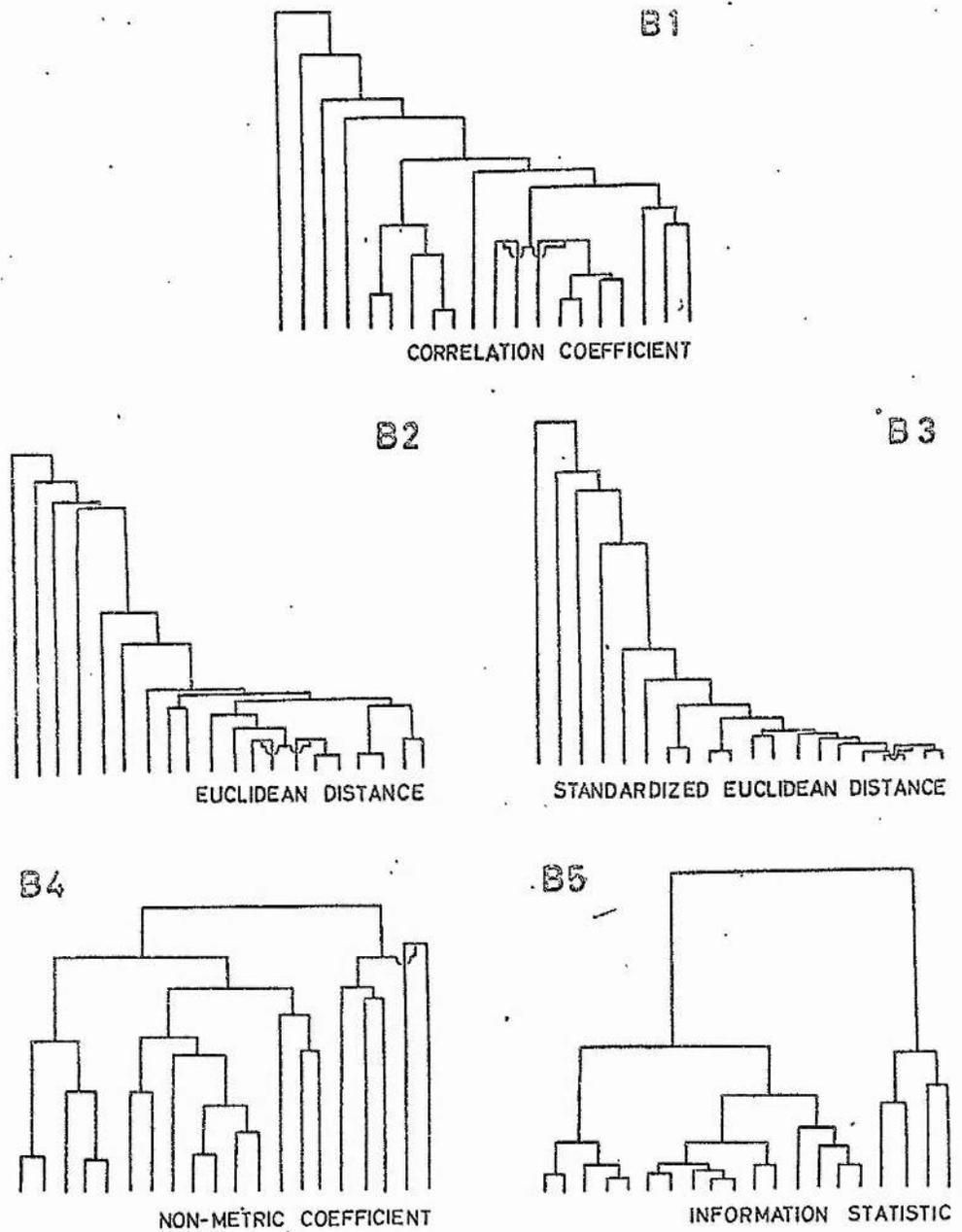


Figure 2.3.2. Dendrograms for centroid sorting using five different similarity criteria (Williams, et al, 1966). Reproduced with kind permission of the authors and the editor of the Journal of Ecology.

dencies to chain (see also Appendices Ia, Ic and Id).

Two interesting aspects of single linkage are shown by figure 2.3.1. Firstly, it is apparent that on several occasions the fusion coefficient is constant for two or more successive fusions. This happens because the similarity matrix contains repeated similarity values resulting from the use of binary data (76 attributes; see also Sect. 2.1), and these single-link coefficients appear as repeated thresholds on the dendrograms. By contrast, hardly any recurring thresholds appear on the centroid sorting dendrograms, because unique centroid coordinates are computed after the early fusions. The second interesting aspect of single linkage is that the euclidean distance dendrogram (unstandardised - A2) is repeated for the information statistic (A5). This shows that the information statistic is proportional to distance when treated solely as a measure of the similarity between two individuals (as is the case with the linkage methods).

At this point, we could attempt to 'explain' the various styles of hierarchy using an involved discussion of weighting schemes and apparent cluster positions (Proctor, 1966; Gower, 1967; Sneath, 1969), or space-dilating/conserving/distorting concepts (Lance and Williams, 1967a). However, such arguments are never completely clarifying or convincing in the absence of a good definition of what is required of the classification

methods, and judgment on the similarity criteria is deferred until later, using the iterative relocation model, rather than hierarchic fusion, with empirical trials (Chapter 7).

CHAPTER 3: DIVISIVE METHODS

3.1 MONOTHEMIC DIVISION

Given the usual binary data matrix for N individuals which either possess or lack a total of M binary attributes, we compute the M x M matrix of chi-square coefficients, defined as

$$\chi_{jk}^2 = \frac{(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

where A,B,C and D are the conventional cell counts for attributes j and k (hence A + B + C + D = N). Next, the columns of this matrix are summed, yielding the vector

$$\left(\sum_{j \neq 1} \chi_{j1}^2, \dots, \sum_{j \neq k} \chi_{jk}^2, \dots, \sum_{j \neq M} \chi_{jM}^2 \right)$$

and we select that attribute k' for which $\sum_{j \neq k'} \chi_{jk'}^2$ is maximum. The population is now divided into two subsets defined by either the presence or absence of attribute k', and these subsets are then examined individually for further subdivision.

Thus far, we have described a neat (if lengthy) method which we can use to obtain 2, 4, 8 or more groups by successive subdivision. However, unless some fundamental law of nature has been overlooked, there are obvious disadvantages to group-forming by powers of 2. An alternative is to obtain 8 groups (say) as above, and then plot a division tree as the converse of a dendrogram, where each division node is placed on a scale determined by its max $\sum \chi^2$ value. The example in figure 3.1.1 shows such a tree,

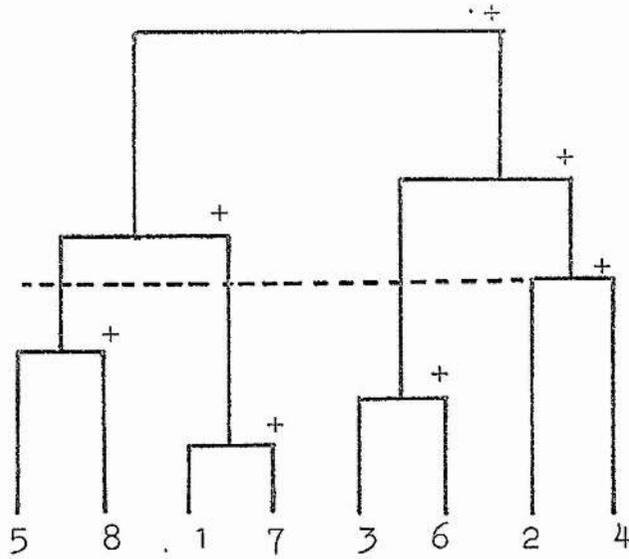


Figure 3.1.1. Nested subdivision tree showing cut-off point which yields 5 clusters.

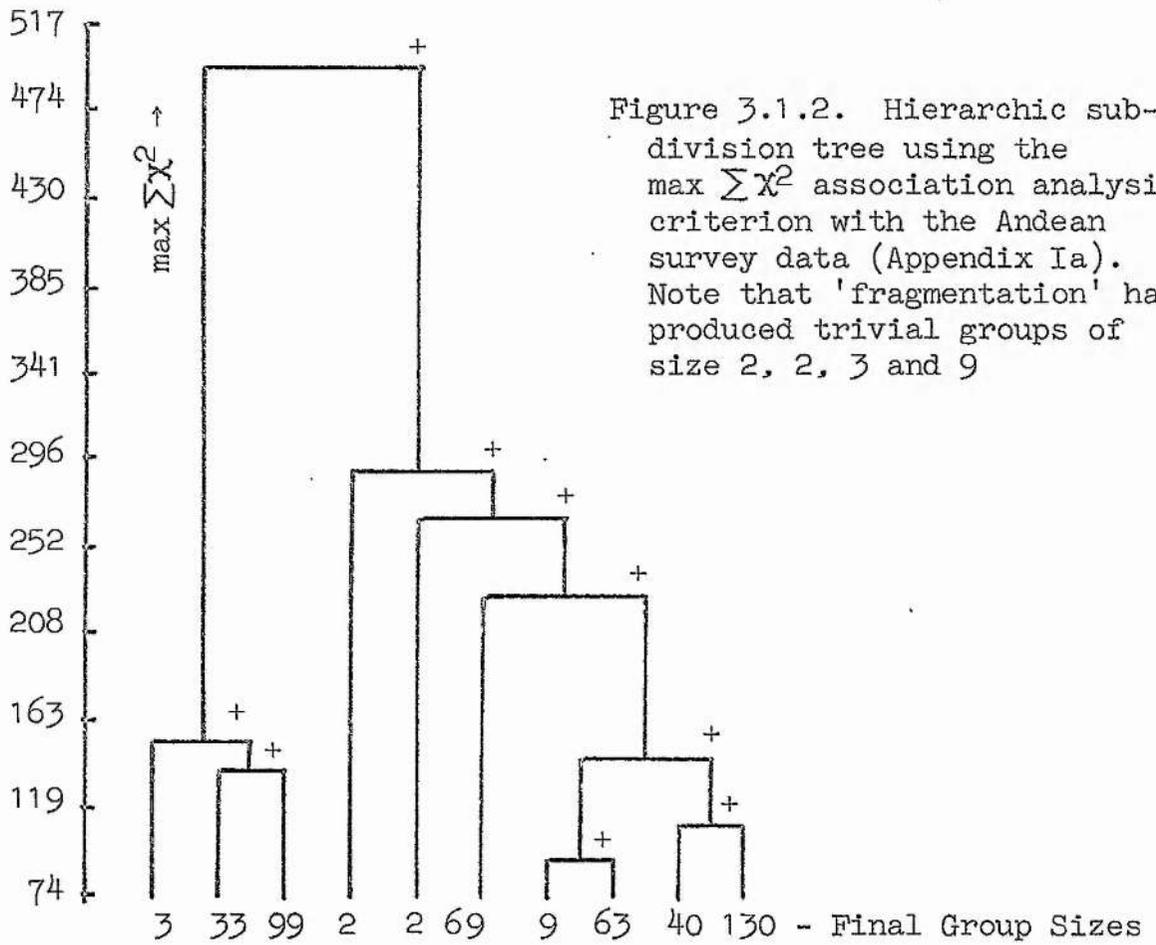


Figure 3.1.2. Hierarchic subdivision tree using the $\max \sum \chi^2$ association analysis criterion with the Andean survey data (Appendix Ia). Note that 'fragmentation' has produced trivial groups of size 2, 2, 3 and 9

where the cut-off point for 5 groups is indicated by a dotted line: we shall call this method 'nested subdivision'.

The technique suffers from the necessity for unprofitable divisions (in the example, the three divisions 5-8, 1-7, and 3-6 are needed to obtain 5 clusters), or alternatively, from the dichotomy that arises when the cut-off has to be drawn across one of the final branches (if we require 6 groups from figure 3.1.1, should the cut-off be placed at the 3-6 division, or through a higher uncalculated point on stem 2 or 4?). The latter situation is resolved by defining 'hierarchic subdivision', being the exact antithesis of 'hierarchic fusion', which works as follows:

1. Divide the total population into 2 groups according to the $\max \sum \chi^2$.
 2. Examine both of the two subgroups thus obtained for their most profitable further subdivisions, and divide that subgroup which has the highest $\max \sum \chi^2$.
 3. Return to 2, and continue to divide the 'best' group at each cycle until a specified number of final groups has been obtained.
- It should be noted that 'hierarchic subdivision' also suffers from unnecessary work: when k final clusters have been obtained, k-2 further subdivisions have been examined. However, this 'look-ahead' information can be supplied to the user at termination, and is, therefore, not totally wasted. Nested subdivision has been used (Crawford et al, 1970) as an alternative to 'stopping

rules' (see below), but there appears to be no instance in the literature of the more general hierarchic subdivision.

Association analysis

The above example of a monothetic divisive process uses the $\max \sum \chi^2$ criterion of Williams and Lambert (1960) which they originally employed for 'normal association analysis'. As previously noted, the method is rather lengthy when the number of attributes (M) is large; the division of a subgroup requiring computation time proportional to m^2 , where m is the number of attributes which are neither totally present or totally absent from the subgroup. It is evident from later applications of association analysis that $\max \sum \chi^2$ tends to 'fragment the analysis by initially splitting off outliers from the population' (Lance and Williams, 1965), a phenomenon which produces a division tree which is analagous to the 'chained' dendrogram of single-linkage. This is quite evident in four empirical studies by Lambert and Williams (1966) whose first divisions found pairs of clusters having the following sizes: (5,15), (6,70), (15,55), and (7,46), where the first figure in each case represents the size of cluster defined by the presence of the division attribute. This phenomenon is also reported by Crawford and Wishart (1966 - see Appendix Ia) who found it necessary, with a 263 quadrat x 142 attribute survey, to impose the arbitrary restriction that the

division attribute should have at least nine quadrat occurrences: even so, the first division yielded clusters of size 32 and 231. If the method is used in its original form with the Andean survey data (Crawford et al, 1970; Appendix Ia) comprising 400 quadrats with 37 attributes, then the hierarchic subdivision tree of figure 3.1.2 is obtained. In fact, Crawford et al (1970) had to use their minimum attribute frequency criterion with these data in order to obtain the more acceptable published result.

Perhaps as a direct consequence of the 'fragmentation' phenomenon, various authors (Williams and Lambert, 1959, 1960, 1961; Macnaughton-Smith, 1965; Lance and Williams, 1965; Lambert and Williams, 1966; Gower, 1967) have proposed variants of association analysis which are designed to 'improve' the technique. Also, the complex subject of 'stopping rules' has been discussed at great length by these writers in an attempt to combine a statistical test of significance with a division strategy that continues subdividing along each arm of the tree until the stopping criterion can be imposed. We have now reached the stage (Lance and Williams, 1965) at which the user may choose from the following division criteria:

$$\max \sum_{j \neq k} \chi_{jk}^2 \quad (3.1.1)$$

$$\max \sum_{j \neq k} \sqrt{\chi_{jk}^2} \quad (3.1.2)$$

$$\max \sum /AD-BC/ \quad (3.1.3)$$

$$\max \sum (AD-BC)^2 \quad (3.1.4)$$

Options (3.1.1) and (3.1.2) are further confused by the occasional use (Williams and Lambert, 1960) of Yates' correction in the computation of the χ^2 coefficients. However, Macnaughton-Smith (1965) recommends neither Yates' correction nor option (3.1.2), and evidently advocates the original $\max \sum \chi^2$. Lance and Williams (1966b) argue the converse from utilitarian considerations, concluding that option (3.1.2) is the 'best general-purpose solution'. One notable aspect of association analysis which should be mentioned is that Professor Williams and his associates do not appear to have discussed the method since 1966 (Lance and Williams, 1966c): does this mean that they now favour the hierarchical fusion (Lance and Williams, 1967a) and iterative relocation (Lance and Williams, 1967c) procedures, or monothetic division using the information statistic (Macnaughton-Smith, 1965; Lance and Williams, 1967b)? In a comparison between information-analysis (hierarchical fusion) and association analysis, Lambert and Williams (1966) conclude:

1. that information-analysis has a 'better theoretical structure'.
2. 'intrinsic and extrinsic misclassification ... is more likely to arise in practice with association analysis, particularly with regard to low-level groupings and all inverse analyses'.
3. 'The greatest single advantage of association analysis over

information-analysis is its greater speed.'

Information statistic

Following the definition of the information I and information-gain ΔI statistics (Sect. 2.2), it is natural that a monothetic divisive technique should be proposed to optimise I . In fact, it is surprising that Lambert and Williams (1966) choose to compare association analysis with the hierarchical fusion method which optimises I (Sect. 2.2): it would have been better if these two criteria could have been compared using the same strategy, viz. monothetic division. In any event, Macnaughton-Smith (1965) proposes that we should evaluate for each attribute k within a given group

$$\Delta I_k = I - I_{k+} - I_{k-}$$

where I is the present information content of the group, and I_{k+} and I_{k-} are the information content values for the two subsets of the group which are determined by the presence and absence of attribute k . We then split the group according to that attribute k' for which $\Delta I_{k'}$ is maximum. This technique is also proposed by Lance and Williams (1968), who suggest that $2\Delta I_{k'}$ is an estimate of χ^2 with $m'(n - 1)$ degrees of freedom, where n is the group size and m' the number of attributes which are not used to define the group. However, they do state that the number of degrees of freedom is usually 'very large, and the test of significance correspondingly weak'. Macnaughton-Smith exploits a connection

between $2\Delta I_k$ and $\sum_{j \neq k} \chi_{jk}^2$ (association analysis), stating that they are 'tolerably close', but in defining his stopping rule as determined by χ^2 ($m - 1$ d.f.) at the 5% level he adds the footnote:

"this stopping rule is completely arbitrary, and is not in any sense a significance test"

and, later:

"all stopping rules are highly arbitrary and their justification is empirical"

In any event, ΔI_k is probably a better divisive criterion than $\max \sum \chi^2$ because, firstly, it does not exhibit 'fragmentation' (see Crawford et al, 1970), and secondly, since I is monotonic decreasing by definition, 'reversals' will not appear on the subdivision tree obtained by plotting the I values.

Group analysis

Crawford and Wishart (1967 - Appendix Ia) propose an alternative to association analysis which is designed specifically for large ecological surveys. Since this paper is reproduced fully in Appendix Ia, the technique will be briefly summarized here using a slightly different approach.

We represent the i th individual by a point whose coordinates are:

$$(\alpha_{i1}, \dots, \alpha_{ij}, \dots, \alpha_{iM})$$

where $\alpha_{ij} = 1$ if attribute j is possessed by individual i , or 0

otherwise. The 'density' v_i is defined as the number of attributes possessed by the i th individual

$$v_i = \sum_{j=1}^M \alpha_{ij}$$

and we denote the mean group density by

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i$$

The mean density for the positive subset associated with attribute j (those individuals which possess j) is given by

$$V_j = \frac{1}{f_j} \sum_{i=1}^n v_i \alpha_{ij}$$

where $f_j = \sum_{i=1}^n \alpha_{ij}$, is the frequency of attribute j , and in the terminology of Crawford and Wishart, the positive subset associated with attribute j will be 'floristically rich' if $V_j \gg \bar{v}$, or 'floristically sparse' if $V_j \ll \bar{v}$.

Next we represent the group by a point whose coordinates, called the characteristic vector \underline{w} , are the weighted group centroid coordinates

$$w_j = (V_j/\bar{v}) p_j$$

where $p_j = f_j/n$ is the centroid coordinate or probability associated with attribute j . Firstly, we note that $0 \leq w_j \leq 1$, and secondly, when every v_i is constant $V_j/\bar{v} = 1$, and hence \underline{w} is identical with the group centroid $(p_1, \dots, p_j, \dots, p_M)$.

The similarity s_i between any one individual i and the group as a whole is computed from the normalised dot product statistic (Sect. 1.4)

$$s_i = \frac{\sum_j \alpha_{ij} w_j}{\sum_j w_j}$$

where the denominator $\sum_j w_j$ is constant for all i , and serves to transform the s_i 's such that $0 \leq s_i \leq 1$. Crawford and Wishart call s_i the 'set element potential', 'SEP', or the 'potential'; for this discussion, the latter term will be used.

Uniformity of group structure is measured by the group coefficient C , which is defined as the mean potential

$$C = \frac{1}{n} \sum_i s_i$$

which reduces, after expansion and manipulation, to

$$C = \frac{\sum_j w_j p_j}{\sum_j w_j}$$

We observe that $0 \leq C \leq 1$, and $C \rightarrow 1$ as every $p_j \rightarrow 1$.

The goodness of attribute j for division is measured by the interaction statistic μ_j^2 (also called $\mu_j'^2$ in the original paper), which is defined as the deviation from the maximum-likelihood estimate for the potential values within the positive (+ j) subset of the group, as follows:

Let $D_j = \sum_{i=1}^n s_i \alpha_{ij}$ be the sum of the potential values for the positive subset for attribute j , then the maximum-likelihood estimate for D_j is

$$E(D_j) = f_j C = p_j nC$$

(this is the same as the expected value for D_j under random sampling).

The interaction statistic is now defined as

$$\begin{aligned} \mu_j^2 &= [D_j - E(D_j)]^2 \\ &= \left[\sum_{i=1}^n s_i \alpha_{ij} - f_j C \right]^2 \end{aligned}$$

and we divide the group on the presence/absence of that attribute k for which μ_k^2 is maximum.

The main advantage of group analysis is its speed (computing time being proportional to $2n$) and storage requirements (maximum of $3M$). Consequently, the method can be used on those occasions when all other techniques are too expensive. However, in deriving a method whose computation time is linear, it is natural that some accuracy should be lost in the search for homogeneous subsets. This is demonstrated by figure 3.1.3, where $\max \mu_j^2$ is seen to be zero for data (a) and (b). Data (a) represents random sampling, and would be difficult to subclassify anyway, while data (b) represents a complete symmetry state for which two very definite subsets exist, although neither is a major subset. Data (c) is an example of the more usual situation, where $\max \mu_j^2$ indicates a division which removes the major subset.

Crawford and Wishart advocate the use of the group coefficient as a stopping rule, in the tradition of Williams and

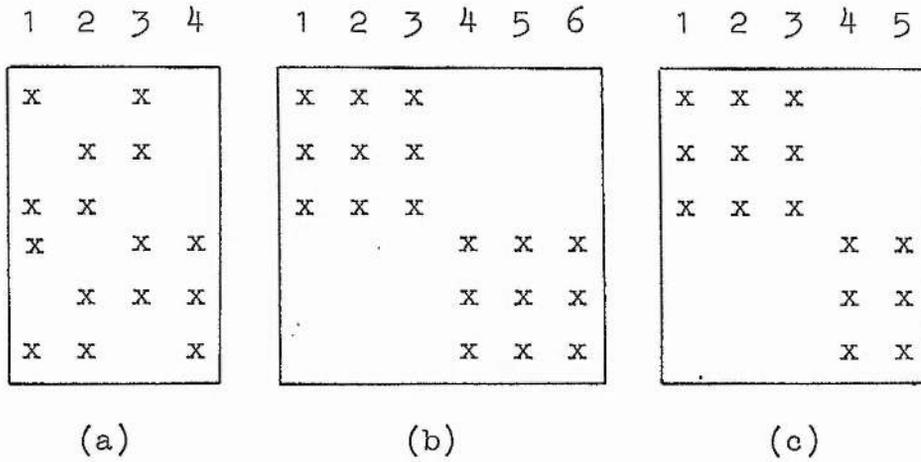
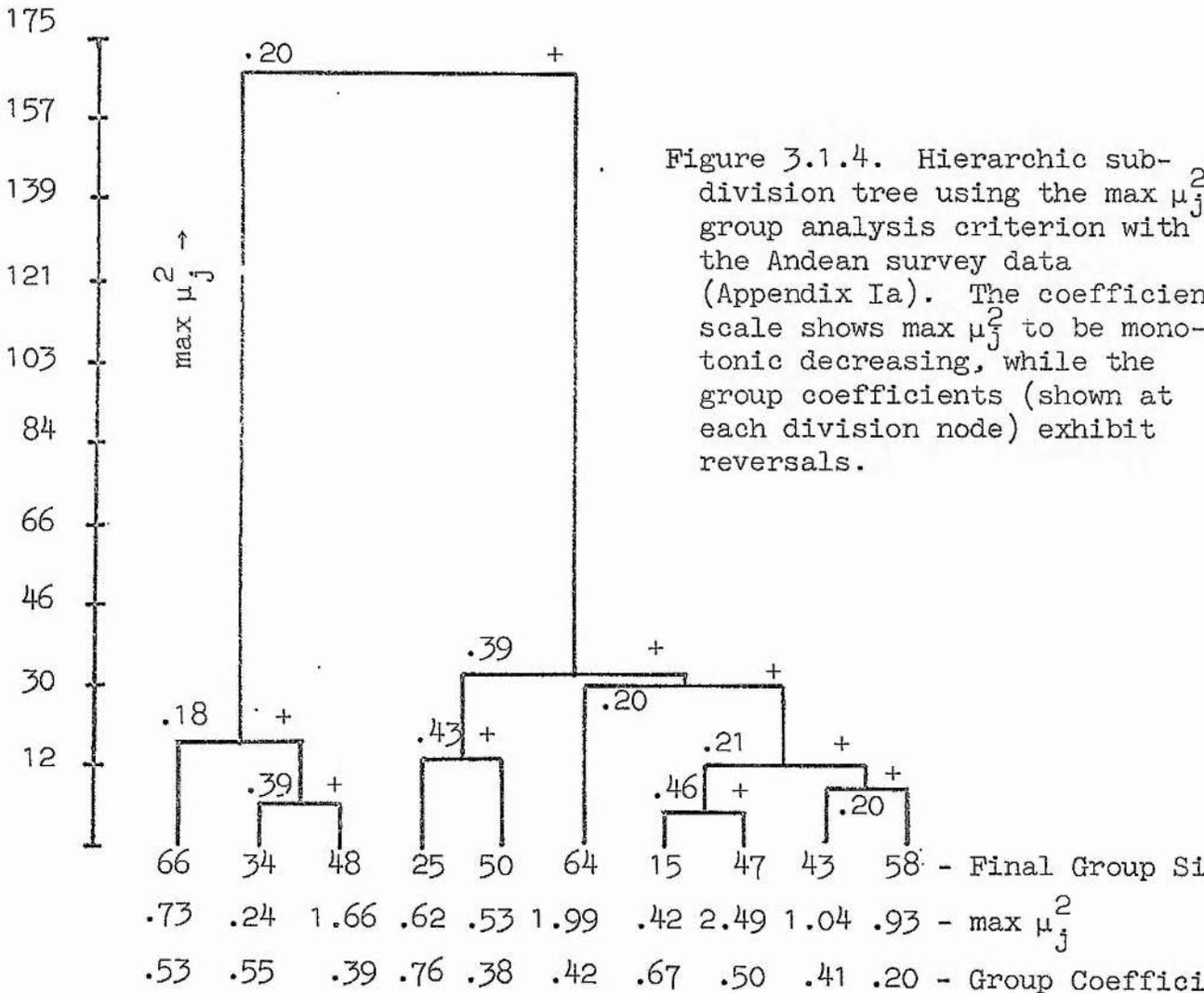


Figure 3.1.1. Three examples of data which illustrate the performance of μ_j^2 . (a) Random sampling: $\max \mu_j^2 = 0$; (b) complete symmetry: $\max \mu_j^2 = 0$; (c) major subset: $\max \mu_j^2 = \mu_1^2 = \mu_2^2 = \mu_3^2 > 0$



Lambert. However, it is easily seen from the linkage tree given in their original paper that C is liable to reversals, particularly on the negative branch. This is not so often the case with $\max \mu_j^2$, which is used as division criterion for the hierarchical subdivision analysis of the Andean survey data (Crawford et al, 1970). The division tree is shown in figure 3.1.4, and at each node, the C and $\max \mu_j^2$ values are indicated. Since the resulting group sizes suggest that $\max \mu_j^2$ at least produces an even division, it is now proposed that the group coefficient C should not be used as a stopping indicator. Instead, the $\max \mu_j^2$ value should be used as division criterion with the hierarchical subdivision algorithm.

Other methods

Gower (1967), when discussing association analysis, shows that the $\sum \chi_{jk}^2$ criterion is the same as the between-group sum of squares when the binary data are standardised. It should be noted, however, that Williams and his associates have never proposed that the binary data should be standardised - indeed, it is not possible to compute χ_{jk}^2 coefficients from standard scores, and it would be necessary to use product-moment correlation coefficients r_{jk} instead (the relationship between correlation r_{jk} and χ_{jk}^2 has been exploited by several writers, e.g. Lance and Williams, 1965; Lambert and Williams, 1966). Nevertheless, if we consider Gower's proposal (in a comparison of association

analysis with the Edwards and Cavalli-Sforza method, Sect. 3.2) as unique, then the between-group sum of squares is the exact complement of Ward's within-group error sum of squares E (Sect. 2.2). We therefore define a monothetic method which splits a group on that attribute k for which the resulting fall in the within-group sum of squares is maximum. That is, we evaluate the decrease in E for the division of each attribute k as

$$\Delta E_k = E - E_{k+} - E_{k-}$$

where E is the within-group sum of squares for the complete group, and E_{k+} and E_{k-} are the sums of squares for the two subsets obtained on division by k . The division attribute k is then chosen to maximise ΔE_k .

Gower next draws parallels between $\sum \sqrt{x_{jk}^2}$ and Kruskal's city-block metric (Kruskal, 1964), and concludes by suggesting two new divisive criteria:

$$\frac{1}{(f_k(n - f_k))^{\frac{1}{2}}} \sum_{j \neq k} |r_{jk}|$$

$$\frac{1}{f_k(n - f_k)} \sum_{j \neq k} r_{jk}^2 \quad (3.1.5)$$

where f_k is the frequency of attribute k . As Gower observes, (3.1.5) is equivalent to the distance between the centroids of the two clusters obtained on division by k , and is therefore analagous to the centroid criterion used with hierarchic fusion. That is,

we compute the distance d_k^2 between the centroids of the two subsets defined by the presence and absence of attribute k , and then choose for division that attribute which maximises d_k^2 .

Conclusions

The main disadvantage of all monothetic divisive methods is that certain individuals will be misclassified by the chance presence or absence of a key division attribute. The second important criticism, common to all divisive schemes (see also Sect. 3.2), is that early partitions can easily cut across natural groups, and by the hierarchic nature of the method, such blunders are unrecoverable. The only advantage of starting with one group is speed, in the sense that, for example, if 4 clusters are required only 3 division steps need be performed, whereas $N - 4$ steps are required for the hierarchical fusion method.

The monothetic divisive technique can be generalised so that we may define a criterion S of the similarity between two groups, and then choose for division that attribute which minimises the similarity between the two resulting subsets. That is, if $S(k+, k-)$ is the similarity between two subsets defined by the presence and absence of any attribute k , then we divide the group according to that attribute k' for which $S(k'+, k'-)$ is minimum.

Crawford et al (1970 - see Appendix Ia) found that Gower's centroid criterion and normal association analysis ($\sum \chi^2$), using the Andean survey data, suffered from fragmentation (however, the

$\sum \chi^2$ result was improved by applying the Crawford-Wishart frequency criterion). The following other divisive criteria were tested, and were found to be in reasonable agreement:

$$\max \sum \sqrt{\chi^2}$$

$$\max \sum (AD-BC)^2$$

$$\max 2\Delta I \quad (\text{information fall})$$

$$\max \Delta E_j \quad (\text{decrease in } E)$$

$$\max \mu_j^2 \quad (\text{group analysis})$$

Finally, it is recommended here that the monothetic divisive procedures should not be used exclusively for group-forming, unless only a rough partition of a very large survey is required. Perhaps the most useful function that the methods can perform is to obtain a fast initial part-optimum solution which can then be improved with the iterative relocation procedure (Chapter 4). For this purpose, group analysis is probably as efficient and economical as any of the others.

3.2 POLYTHETIC DIVISION

Edwards and Cavalli-Sforza (1965)

The error sum of squares E , although apparently first used by Ward (1963) as an 'objective function' optimised by hierarchic fusion (Sect. 2.2 and 8.2), is often attributed to Edwards and Cavalli-Sforza (1965) as a homogeneity indicator in the context

of cluster analysis (Orloci, 1967b; Gower, 1967; Calinski and Harabasz, 1970); in fact, Orloci says (personal communication) that he was "inspired by Edwards and Cavalli-Sforza" when he repropounded Ward's method (Orloci, 1967b).

That there exist one or more absolutely optimum solutions for the error sum of squares E for a given number of clusters has intrigued many writers (Forgey, 1964, 1965; Dagnelie, 1967; Bolshev, 1969; Calinski, 1969; Calinski and Harabasz, 1970), and Edwards and Cavalli-Sforza are attributed with the only method which guarantees to find the optimum partition. They do so by examining all $(2^{N-1} - 1)$ possible divisions of the population of N individuals into two classes, and compute the error sum of squares in every case. Having found the best two clusters, however, they then abandon the idea of finding the optimum division into 3 groups (because the examination of $(3^{N-1} - 2^N + 1)/2$ classifications is "impossible") and prefer instead to partition into two the first two groups obtained, using the same division procedure as before. They continue in this way, obtaining a division tree which is the same as the monothetic 'nested subdivision' (Sect. 3.1). Since it cannot be claimed that, in general, successive optimum solutions for E satisfy the imposed hierarchical structure, Edwards and Cavalli-Sforza cannot guarantee to find the optima for other than 2 groups; in fact, this drawback is mentioned by Edwards and Cavalli-Sforza, and demonstrated by

Calinski and Harabasz (1970).

It should also be noted that the Edwards and Cavalli-Sforza method is extremely inefficient, being computational "impossible" for more than about 20 individuals (Macnaughton-Smith, 1965; Lance and Williams, 1966b; Orloci, 1967b; Gower, 1967b) due to the enormous number ($2^{N-1} - 1$) divisions that have to be examined. The method is a classical example of the mis-use of computational facilities, and is of interest solely for its treatment of E.

Calinski-Harabasz Shortest Dendrite Method

The 'shortest dendrite' or minimum spanning tree (Florek et al, 1951; Gower and Ross, 1969) is the graph of $N - 1$ edges which connects all points in the sample space, has the least overall length and no circuits. It is analagous to the hierarchical fusion process for single linkage, where the pair of nearest neighbours at each step defines an edge of the graph. Calinski and Harabasz (1970; see also Calinski, 1969) reason intuitively that the optimum error sum of squares solution for k clusters may be obtainable by removing $k - 1$ edges from the shortest dendrite. That this is not always true, is demonstrated by figure 3.2.1 for which the optimum solution for E when $k = 2$ requires the removal of two edges (as indicated by the dotted partition line c). However, the method was shown by the authors to yield a better solution than the Edwards and Cavalli-Sforza method when $k = 5$ with a population of 12 Indian castes; in fact, the Calinski-

Harabasz result confirmed a previous finding of Rao (1952) who used an average distance criterion with principal components analysis.

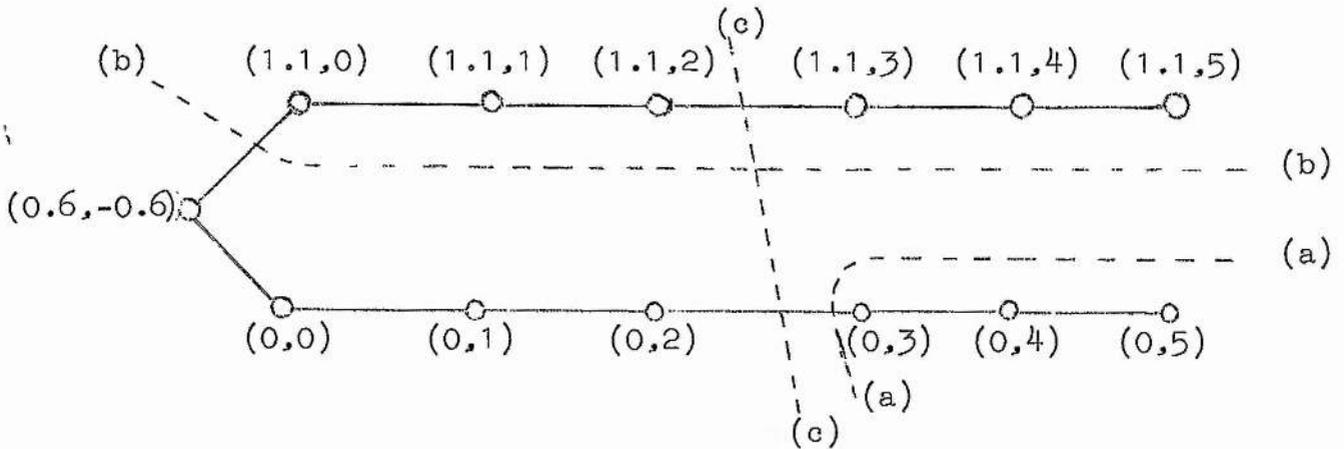


Figure 3.2.1. Example of a minimum spanning tree (solid lines) which cannot be partitioned to find the optimum error sum of squares for two clusters. The points' coordinates are given, and the distance matrix was computed without standardisation. Partitions are: (a) solution for 2 classes by Calinski-Harabasz method; (b) starting solution for the iterative relocation procedure (Chapter 4); (c) final solution after 6 relocations and 2 iterations. Solution (c) is evidently the optimum.

An important feature of the Calinski-Harabasz method is that it is non-hierarchical, as opposed to all other polythetic divisive schemes considered here, and hence it is not subject to the criticism that any one result is dependent on previous partitions for its efficiency (see below). The method is much faster than the Edwards and Cavalli-Sforza technique, requiring computation time proportional to $\binom{N-k}{k-1}$. However, although this permits

populations of order about 60 to be considered, the method is computationally slow when compared with the hierarchic fusion and iterative relocation procedures, the time factor being roughly N^k when $N \gg k$. It can also be argued that the method is inefficient because it considers some partitions of the dendrite which are highly unlikely to be profitable (viz. the removal of $k-1$ edges located together at an extreme vertex of the graph).

Dissimilarity Analysis

Macnaughton-Smith et al (1964; see also Macnaughton-Smith, 1965) propose a polythetic divisive scheme which determines a single partition of a cluster, and then derives a nested subdivision tree in the same fashion as Edwards and Cavalli-Sforza. The method works as follows:

1. Each individual is compared with the set of all the rest, and we choose that individual which is least similar to the rest. With the centroid criterion, we would select the point which is farthest from the group centroid.
2. We next consider all pairs of individuals, including the one chosen above, and select the 'best' pair.
3. Each triad, containing the pair selected at 2, is considered and the best such triad chosen.
4. In this way, we develop a partition of the set by moving each individual which belongs to the 'rest' into an accumulating

subset, and then examine the resulting similarity between the 'subset' and the 'rest'. At the end of each cycle, we move to the subset that single individual whose move results in the greatest dissimilarity between the subset and the rest.

5. The procedure stops when the 'best' individual is more alike to the 'rest' than to the 'subset', at which stage the move is deemed to be unprofitable.

Mathematically, we denote by $f(P,Q)$ the chosen function which measures the similarity between sets P and Q . Hence if x_i is an individual belonging to the 'rest' R , and G is the growing subset, then we evaluate for each $x_i \in R$

$$d(x_i) = f(G+x_i, x_i) - f(R-x_i, x_i)$$

and choose that individual X for which $d(X)$ is a maximum. Then, provided that $d(X) > 0$, we remove X from R and place X in G . If $d(X) \leq 0$ we stop, and the best partition of the set into subsets G and R has been found. Each subset thus found is considered separately for further division, thereby deriving the nested subdivision sequence (regrettably, no rule for the order of such divisions is suggested by the authors).

The important features of this analysis are as follows:

1. The computation is not fast, being of the order of

$$n + (n-1) + \dots + (n-g) = \frac{1}{2}(g+1)(2n-g)$$

when cluster G ends up with g members. If $g = n/2$ (and this is not necessarily the "worst" case, as suggested by Lance and Williams (1966b), because it is possible that $g > n/2$), then this reduces to $3n(n+2)/8$, which must be further multiplied by the factor corresponding to the evaluation of $d(x_i)$. In fact, the dissimilarity function used by Macnaughton-Smith et al is of the order M^2 , where M is the number of attributes (binary), so that the time for each division step is proportional to $3M^2n(n+2)/8$ in this case, where n is the size of the group being considered (Lance and Williams (1966b) incorrectly deduce the factor $3n^2/4$ for dissimilarity analysis).

2. The method, like all divisive schemes, suffers the drawback that inefficient early partitions cannot be corrected (see Gower, 1967; also Sect. 3.1, and below). For example, a natural 3-cluster grouping is unlikely to be found since one of the clusters will probably be split in two at the first step. Also, the initial direction of the partition is determined by the most remote individual (when the distance criterion is used), which is not particularly likely to indicate the direction of a natural density saddle. Furthermore, this likely peripheral misfit, regardless of its final relationships with R and G, is constrained to belong to G from the very start.

3. The method is evidently suggested for use with nested subdivision, and the disadvantages of this technique have already

been outlined (Sect. 3.1).

Conclusions

Many authors express a preference for divisive systems, using such arguments as:

"divisive methods, which start with the whole sample, are in general safer than agglomerative methods"

- Macnaughton-Smith et al (1964)

"since divisive methods are preferable to agglomerative, Similarity Analysis (meaning single linkage) is not considered in the present paper"

- Macnaughton-Smith (1965)

"when monothetic classification by attributes is acceptable or even desirable, a more powerful divisive system is possible, in that the function used for selection of attributes can be calculated over the entire population"

- Lance and Williams (1965)

"the single greatest advantage of a divisive system like association analysis is that the analysis begins at a high information level"

- Lambert and Williams (1966)

By contrast, Gower (1967) writes:

"It is held that divisive methods will not lead to any spurious groupings and although this is probably mostly true there appears to have been no formal investigation. For example, suppose we have three well-defined groups; then no harm is done if division is made as in figure 3.2.2(a), but can we guarantee that it will not occur as in figure 3.2.2(b)? We would, however, probably be happier if divisions were made as in figure 3.2.2(c), which is the type of clustering found by agglomerative methods."

In fairness to the previous advocates of divisive systems (who, incidentally, are all attributed with the authorship of divisive

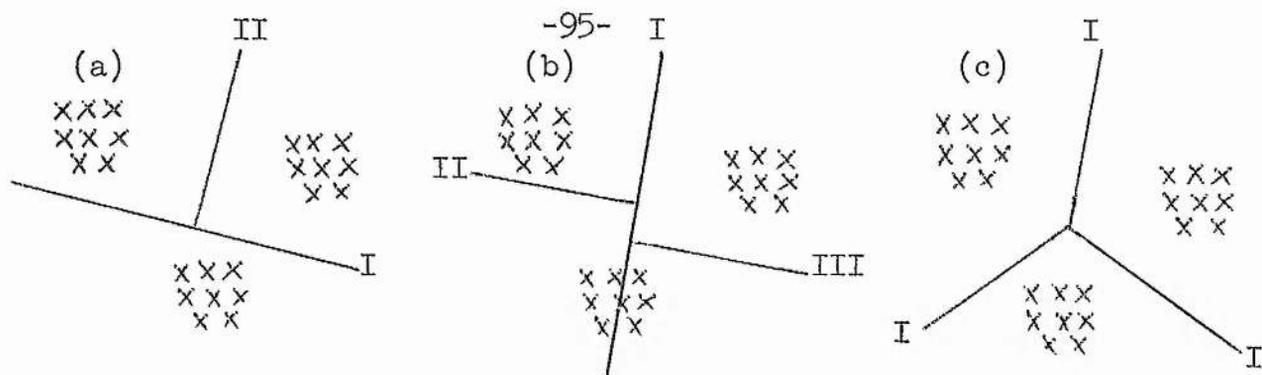


Figure 3.2.2. Possible divisive and agglomerative clustering results, (b) indicates the type of irreversible failure found in monothetic methods; (c) shows the sort of clustering produced by agglomerative methods.

methods), it should be noted that the agglomerative methods had not been fully exploited during the period 1964-66 (although most of those discussed in Chapter 2 had been published). In fact, Macnaughton-Smith (1965) mentions only single-linkage before disposing of agglomerative methods (in general), and Williams et al (1966) consider only centroid sorting and single-linkage: is it possible, therefore, that their dissatisfaction with agglomerative methods is merely the disguised practical experience of chaining? Faced with a dendrogram showing extensive chaining (e.g., Appendix Ic, or figures 2.3.1 and 2.3.2) it would be very easy to attribute the failing to "false decisions made in the early stages of the analysis" - Macnaughton-Smith et al (1964).

In any event, Gower's lucid assessment now seems more plausible than the previous unsubstantiated remarks, and it is therefore recommended here that agglomerative (Chapter 2) and iterative relocation (Chapter 4) methods should be used where possible, rather than divisive methods.

CHAPTER 4: ITERATIVE RELOCATION

4.1 GENERAL PROCESS

Needham (1962) describes a method for 'clumping' which, like the dissimilarity analysis of Macnaughton-Smith et al (1964; Sect. 3.2), finds a single division of the population into two classes by testing the movement of individuals across a varying partition. He first defines the cohesion (similarity) between a "subset" G and its complement the "rest" R, and then attempts to minimise the cohesion by iteratively scanning the population, moving individuals from R to G (or vice-versa) if the move yields a profitable decrease in the function value. Although this approach appears to be very alike to dissimilarity analysis, there is one significant difference: no assignment to the clump G by Needham is deemed irreversible. This therefore corrects the major fault of dissimilarity analysis (Sect. 3.2) by which inefficient early assignments cannot be corrected, for we observe that Needham's method permits an individual to initially belong to G, move into R as a result of a change in the membership of R, move back into G due to a further change of membership, and so on.

In a survey of iterative relocation methods, Lance and Williams (1967c) suggest five processes that they deem to be the necessary ingredients of a clustering (relocation) procedure. These processes may be summarized as follows:

- (a) A method of initiating clusters.
- (b) A method of allocating new members to existing clusters.
- (c) A method of fusing clusters.
- (d) A stopping rule to terminate procedures (b) and (c).
- (e) A method of reallocating individuals, in order to "redress any misclassification".

If we assume that such a procedure will be designed to look separately at several different classifications, the general case being that of k clusters, then although processes (a)-(d) will influence the result (Chapter 7) they are largely irrelevant. This is because the reassignment procedure (e) is essentially designed to improve a given classification, and the token of a good method is whether it can improve both good and bad classifications producing more or less the same final result. Hence, the way in which the initial classification is obtained should not be important.

As mentioned earlier, the significant feature of the iterative relocation procedure is that no allocation or reallocation is necessarily irreversible. We can therefore now define the iterative relocation procedure as any method which attempts to 'improve' a classification by altering cluster assignments during repeated scanning of the population, provided that no such reassignments are irreversible. It remains to define what is meant by

"improve 'a classification", and we shall do so, in general, by adopting a function $f(X,x)$ which measures the similarity between an individual x and a cluster of individuals X ($x \in X$ is permitted). Next, we consider a classification into k disjoint classes, where $x \in p$, and q is any other cluster. We find that all the instances of reassignment procedures in the literature can be generalised into one of two tests for relocation, which are that x is removed from parent cluster p and added to cluster q if:

either
$$f(p-x,x) < f(q,x) \quad (4.1.1)$$

or
$$f(p,x) < f(q,x) \quad (4.1.2)$$

(note that the inequality is reversed if f measures dissimilarity).

(4.1.1) implies that x is removed from its parent cluster p , and then compared with the residual members of p on an equal par with the members of any other cluster q . (4.1.2) allows x to be compared with the full set of objects p (including x), and therefore induces a slight bias in the LHS. In particular, it is possible that

$$f(p,x) < f(q,x)$$

while
$$f(p-x,x) > f(q+x,x)$$

in which case, (4.1.2) could initiate the oscillation of x between p and q . For these reasons, (4.1.1) is generally preferred. However, we note that if a large population is considered and all clusters have reasonable size, then both tests will behave alike when the function f takes into account group structure: the nearest-

neighbour criterion would be meaningless with (4.1.2) since the LHS usually becomes $f(x,x) = f_{\max}$.

In the particular event that cluster p comprises the single individual x , we shall adopt the arbitrary function value $f(x,x)$ for the LHS of (4.1.1). Hence, both tests reduce to

$$f(x,x) < f(q,x) \quad (4.1.3)$$

in this instance. Note that (4.1.3) can be satisfied with particular functions (e.g. the dot product coefficient). When this happens, x is relocated with cluster q so that cluster p is eliminated and the number of clusters reduces to $k-1$.

4.2 METHODS

Average linkage

Thorndike (1953) is probably the first to have proposed an iterative relocation procedure, using the following method:

"Generally speaking, a specimen is misassigned if it is closer to the members of another cluster than it is to the members of its own Cases of this sort are reassigned, one at a time, starting with the most obvious misfits, and the average distances are recomputed after each assignment ... Shifts are made until there is no further shift which will reduce the average of all the within-cluster distances."

If we denote the current total sum of the within-cluster distances by T , the sum of the distances between clusters p and q by S_{pq} , the size of cluster p by k_p , and the total number of within-cluster distances N_D , then the current average of within-cluster distances is

$$A = T/N_D$$

If an individual x belonging to cluster p is moved to cluster q , then the new sum of the within-cluster distances will be

$$T + Sqx - S(p-x)x$$

and since the number of these distances changes to

$$N_D + k_q - (k_p - 1)$$

Thorndike moves x from cluster p to cluster q if

$$\frac{T}{N_D} > \frac{T + Sqx - S(p-x)x}{N_D + k_q - (k_p - 1)}$$

This becomes, after manipulation

$$Tk_q - N_D Sqx > T(k_p - 1) - N_D S(p-x)x$$

or, on division by N_D

$$Ak_q - Sqx > A(k_p - 1) - S(p-x)x$$

Hence Thorndike requires relocation if (4.1.1) is satisfied for

$$f(X, x) = Ak_x - SXx.$$

k-mean System

Forgey (1965), Jancey (1966), Ball and Hall (1966; see also, Ball, 1965), and MacQueen (1967) all examine the distances from each individual to the centroids of the clusters. Any individual x whose distance from the centroid of its parent cluster p (d_{px}^2) exceeds its distance from the centroid of another cluster q (d_{qx}^2) is moved from p to q . Hence, in our terminology we write

$f(X,x) = d_{Xx}^2$, and use (4.1.2) for the relocation test.

It is worth noting that these three authors differ significantly in respect of (a) cluster initiation and (b) centroid computation. Forgey begins with the part-optimum solution for k clusters produced by Ward's method (which is not the "arbitrary" assignment of k groups reported by Lance and Williams, 1967c); Jancey selects k points at random from the euclidean space (not individuals), and therefore partitions the space; Ball and Hall, and MacQueen choose k individuals to act as primary centres (the rest of the population is assigned to a residue and then relocated during the first cycle).

MacQueen (like Thorndike) is the only writer who recomputes the cluster centroids at relocation time: that is, if x is moved from p to q then the centroids of p and q are immediately recomputed to account for the change. The other authors recompute centroids at the end of each scan of the population, and it seems probable that MacQueen's method will converge faster.

All of these methods evidently use (4.1.2) for the relocation test, and they are therefore liable to oscillate.

Error sum of squares E

Although Forgey (1965), Jancey (1966) and Ball and Hall (1966) use the nearest centroid to indicate misclassifications, they also adopt the error sum of squares E either as the optimised objective, or as indicating the goodness of the resulting classi-

fication. As previously mentioned, Forgey's method attempts to improve Ward's part-optimum result for E; Ball and Hall state as their objective "to minimise the error of fit", otherwise described as "the sum of the squared distances of the patterns (individuals) from their individual averages (cluster centroids)"; Jancey calls E the "total within-group variance" and uses it as an indicator of homogeneity with the k-mean system.

Beale (1969), however, proposes the following relocation test which directly optimises E:

$$\text{Let } \Delta E_x = (E_p + E_q) - (E_{p-x} + E_{q+x}) \quad (4.2.1)$$

be the fall in E resulting from the relocation of x from p to q, where E_x is the component of E attributed to cluster X. Hence, x is relocated with cluster q if

$$\Delta E_x \geq 0$$

i.e. if

$$E_p + E_q \geq E_{p-x} + E_{q+x}$$

$$E_p - E_{p-x} \geq E_{q+x} - E_q \quad (4.2.2)$$

Therefore, if we define $f(X,x) = E_{X+x} - E_x$, the increase in E_x caused by the addition of individual x to cluster X, then Beale's method is obtained with test (4.1.1).

The increase in E which results from the fusion of two clusters p and q is given by (Sect. 8.2)

$$I_{pq} = \frac{k_p k_q}{k_p + k_q} d_{pq}^2$$

where k_p, k_q are the cluster sizes. Hence, it is easy to show that

$$I_{(p-x)x} = \frac{k_p}{k_p - 1} d_{px}^2$$

where d_{px}^2 is the distance from x to the cluster p centroid computed without allowing for the removal of x . (4.2.2) now becomes

$$\frac{k_p}{k_p - 1} d_{px}^2 > \frac{k_q}{k_q + 1} d_{qx}^2 \quad (4.2.3)$$

This is, in fact, the relocation inequality used by Beale to minimise E .

One interesting aspect of Beale's method is that, having iterated to convergence for k clusters he proceeds to fuse those two clusters which cause the least increase in E (as for Ward's method), and then iterates to convergence for $(k-1)$ clusters. This alternate iteration-fusion process continues until a specified number of final clusters is achieved. Beale notes that "the program sometimes finds local optimum clusterings that are not global optima ... It has therefore been found advisable to start with at least 3 more clusters than one is really interested in. The solutions for all relevant numbers of clusters should then be good ones". This may well be due to his initial choice of k random points (not individuals) as first cluster centres, and would probably be corrected by using Ward's part-optimum solution instead.

Information Statistics

Lance and Williams (1967c), in discussing Hyvarinen's clumping method (Hyvarinen, 1962), suggest that the optimisation of I or ΔI would "undoubtedly repay further study". If we adopt the same definitions of information as those used with the hierarchic fusion (Sect. 2.2) and monothetic division (Sect. 3.1) methods, then an iterative relocation optimisation is easily formulated, as follows:

Denote by
$$\Delta I_x = (I_p + I_q) - (I_{p-x} + I_{q+x})$$

the information fall caused by relocation, then x is moved from p to q if $\Delta I_x > 0$. This is immediately seen to be the same as for E (4.2.1), so that we can define $f(X,x) = I_{X+x} - I_X$, and adopt inequality (4.1.1) for the relocation test.

Cohesion

In Section 4.1, the clumping theory of Needham (1962) was briefly described in relation to dissimilarity analysis, but the cohesion functions that he and his associates have proposed were not defined. Given that SAB is the sum of the between-group similarities for clusters A and B, then we obtain a partition of the entire population into group G and its complement R (the 'rest') in order to minimise the cohesion function C. The following definitions of cohesion have been published:

$$\text{SGR} \qquad \text{Needham, 1965b} \qquad (4.2.4)$$

$$\frac{\text{SGR}}{\text{SGG} + \text{SRR}} \qquad \text{Jones and Jackson, 1967} \qquad (4.2.5)$$

$$\frac{(SGR)^2}{SGG * SRR} \quad \text{Jones and Jackson 1967} \quad (4.2.6)$$

$$\frac{SGR}{SGG} * \frac{k_G(k_G-1)}{SGG} \quad \text{Jones and Jackson 1967} \quad (4.2.7)$$

$$\frac{SGR}{SGG} \left[\frac{k_G(k_G-1)}{SGG} - \frac{SGG}{\phi k_G(k_G-1)} \right] \quad \text{Parker-Rhodes and Jackson, 1969} \quad (4.2.8)$$

In (4.2.8), ϕ is a 'proportion' apparently chosen by the user. It should be noted that Needham also defined cohesion in 1962, but it is rather difficult to decipher precisely what he meant.

If, in general, we denote cohesion by $C(G,R)$, then Needham proposes that x should be moved from its parent group G to the complement R (or vice-versa) if

$$C(G,R) > C(G-x,R+x) \quad (4.2.9)$$

Let U represent the entire population, then $R = U - G$, and on substitution, (4.2.9) becomes

$$C(G,U-G) > C(G-x,U-G+x) \quad (4.2.10)$$

which is obtained with relocation test (4.1.1) when

$$f(X,x) = C(X+x,U-X-x) \quad (4.2.11)$$

However, (4.2.11) appears to serve no practical purpose, excepting to show that Needham's method can be expressed in terms of the generalised relocation test (4.1.1). If we now substitute the cohesion function of (4.2.5) in (4.2.10), the relocation test

becomes:

$$\frac{SGR}{SGG + SRR} > \frac{S(G-x)(R+x)}{S(G-x)(G-x) + S(R+x)(R+x)} \quad (4.2.12)$$

and since we may write

$$S(R+x)(R+x) - SRR = SRx$$

$$SGG - S(G-x)(G-x) = S(G-x)x$$

$$T - SRR - SGG = SGR$$

where T is the total sum of all the similarities (a constant),

then (4.2.12) is equivalent to

$$\frac{T - SGG - SRR}{SGG + SRR} > \frac{T - SGG + S(G-x)x - SRR - SRx}{SGG - S(G-x)x + SRR + SRx}$$

which, after manipulation, reduces to

$$T.SRx > T.S(G-x)x$$

Hence, when $T > 0$, x is moved from G to R if the sum of the similarities between x and the members of R exceeds the sum of the similarities between x and the other members of G. A similar result is obtained for (4.2.4), since the relocation test may be stated as: move x from G to R if

$$SGR > S(G-x)(R+x)$$

i.e. if $T - SGG - SRR > T - S(G-x)(G-x) - S(R+x)(R+x)$

which reduces, after substitution for the RHS, to

$$SRx > S(G-x)x$$

These two cohesion functions (4.2.4 and 4.2.5) may now be seen to behave differently according to the type of chosen similarity

measure. Suppose that the positive dissimilarity function d^2 is used, then $\max \text{SGR} (\min \text{SGG} + \text{SRR})$ has a stable optimum for which G and R are non-empty (since $\text{SGG} + \text{SRR}$ is maximum when $G = U$ and R is empty). On the other hand, for a positive similarity measure - such as $A/(A + B + C)$ used by Needham (1962), and Parker-Rhodes and Jackson (1969) - we require the $\min \text{SGR} (\max \text{SGG} + \text{SRR})$ partition: since $\text{SGG} + \text{SRR}$ is always greatest when G is the entire set (given a non-negative similarity measure) the function now becomes highly unstable.

It is easy to understand why Jones and Jackson write "classes on this definition (referring to 4.2.5) are unfortunately difficult to find", and Needham (1962) writes "the only hazard is that the algorithm may terminate at the absolute minimum cohesion - which occurs in the trivial case of the subset becoming the whole set". A re-examination of these functions (at present rejected by Needham and his associates) in relation to dissimilarity coefficients would probably prove to be valuable.

One interesting aspect of the method is that different 'clumps' are formed by using the iterative relocation procedure to convergence with different starting conditions, looking only for the subset G (R is ignored). However, precautions are evidently needed to ensure that neither R nor G becomes empty (Needham, 1965), and that oscillation is prevented (Jones and Jackson terminate the program, if necessary, after a specified number of

iterations). The method appears to find repeated clusters if C is well-defined, and overlapping clusters otherwise.

Agglomerative Group Analysis

Crawford and Wishart (1968 - see Appendix Ia, and Crawford, 1969) have attempted to improve their divisive group analysis (Sect. 3.1), using the potential values s_i as indicative of misfits. For any individual x ep such that $s_{px} < \beta$, where s_{px} is the potential of x with cluster p and β is a chosen threshold, Crawford and Wishart evaluate the potentials (s_{tx}) between x and all other clusters t , and move x to the 'best' cluster q if

$$s_{qx} > s_{px}$$

The relocation test is therefore obtained with (4.1.2) when $f(X,x) = s_{Xx}$. It is shown that different values of β yield remarkably different results (degrees of "accuracy"), and although the ecological example given by the authors is reasonably convincing, there remains the criticism that their iterative process is not stable, but dependent on user controls. This is probably due to the choice of function s_{Xx} , which is a directional coefficient (related to the dot product) and may well be ill-conditioned. In retrospect, the error sum of squares optimisation would be better employed for the plant ecology problem, possibly using the results of divisive group analysis as starting conditions.

4.3 DISCUSSION

It was argued earlier that the essential process of iterative relocation is the method by which individuals are reassigned in order to improve the selected similarity criterion, the token of a good criterion being that the same final result is obtained regardless of starting conditions. Needham's clumping method presents excellent examples of unsuitable similarity criteria: indeed, the method is designed to find a different clump from each starting condition, and the authors systematically select differing initial classifications to force unique clumps (Needham, 1965b). Forgey uses the k-mean system to improve Ward's part-optimum result, and then obtains a further reduction in E by "sliding the partitions back and forth between each pair of centroids". Beale notes that the random start does not always find a global optimum for E, and suggests the use of iterative relocation to find part-optimum solutions at the $(k+3) - (k+1)$ cluster levels so that a good result for k clusters may be achieved. It is now clear from these recommendations that no matter how good a measure of similarity, cohesion or disorder we define, the final classification will sometimes be dependent on the starting conditions, and it is therefore essential that a good part-optimum starting solution should be used. I therefore choose to generalise Beale's algorithm as follows:

- 1) Define a similarity function $f(X,x)$ - usually a measure of

intercluster similarity such as the increase in E (4.2.1). $f(X,x)$ may then be any measure $f(p,q)$ of the similarity between two clusters p and q , where we treat individual x as a 1-element cluster q .

2) Find a starting classification of k clusters: this could be the k cluster part-optimum solution obtained from a hierarchical method such as Ward's (ideally using the same similarity criterion f). Alternatively, if the population is too large and the hierarchical method too expensive, an initial solution such as k random individuals or k random groups could be generated. In this case, the value of k should exceed by 3 or more the required maximum number of clusters.

3) In one scan of the population, each individual is compared with all k clusters and moved from its parent cluster p to the 'best' other cluster q if

$$f(p-x,x) < f(q,x)$$

(the inequality should be reversed for a dissimilarity function f). All the necessary cluster characteristics (such as means, standard deviations, etc.) should be recomputed after each relocation, and not at the end of every scan (Sect. 5.2).

4) The population is repeatedly scanned until either no more individuals are relocated during one complete scan, or MAXIT scans have been finished. The parameter MAXIT (chosen by the user) is used to prevent oscillation, but a well-condition function

such as the increase in E would normally converge after no more than 10-12 iterations.

5) At convergence, the classification is made available, and then the two most similar clusters are fused. These are the clusters p and q for which $f(p,q)$ is maximum (or minimum if f measures dissimilarity). The procedure then returns to (3) and iterates until convergence for (k-1) clusters.

6) This cyclic process of iteration followed by fusion ends when an optimum solution has been found for k' clusters ($k \geq k'$).

The notable feature of the procedure is that computation time is more or less linear with population size (of the order $N \cdot \text{MAXIT}$) for each value of k. This means that if k is small, very large populations can be analysed efficiently and economically. Secondly, since cluster assignments are always reversible, and well-defined measures of disorder such as E appear to stabilise yielding meaningful results, the method seems to be logically better for group-finding than those previously discussed (however, the results are not as readily presentable as those of the hierarchic method: see, for example, Appendix Ie). Thirdly, the use of relocation test (4.1.2) is not recommended because it can initiate an indefinite oscillation: for example, with the k-mean system, the distance between x and the other members of its parent cluster p ($d_{p-x,x}^2$) would have been better employed than the distance d_{px}^2 used by those authors. Finally, the procedure does not force

clusters by means of constraints of methodology (such as the presence/absence requirement of all monothetic methods), and therefore iterative relocation provides an ideal model for the empirical evaluation of different similarity criteria (this is exploited in Chapter 7).

CHAPTER 5: GENERALISED TRIPARTITE PROCEDURE

5.1 INTERCLUSTER SIMILARITY FUNCTION

Let $(X_{i1}, \dots, X_{ij}, \dots, X_{iM})$ be the position vector for the i th individual, where X_{ij} is the value of variable j (continuous or binary). Suppose that cluster t comprises k_t individuals and has centroid $(U_{1t}, \dots, U_{jt}, \dots, U_{Mt})$; hence

$$U_{jt} = \frac{1}{k_t} \sum_{i \in t} X_{ij}$$

We now define the following parameters associated with cluster t :

$$\alpha_t = k_t \sum_{j=1}^M U_{jt} = \sum_j \sum_{i \in t} X_{ij}$$

$$\beta_t = k_t^2 \sum_j U_{jt}^2 = \sum_j \left[\sum_{i \in t} X_{ij} \right]^2$$

$$\gamma_t = \sum_j \sum_{i \in t} X_{ij}^2$$

and in the comparison of two disjoint clusters p and r , we define

$$\delta_{pr} = k_p k_r \sum_j U_{jp} U_{jr} = \sum_j \left[\sum_{i \in p} X_{ij} \right] \left[\sum_{i \in r} X_{ij} \right]$$

The similarity between two clusters p and q , in terms of several quantitative measures (see below), can now be expressed as a function of these parameters using the notation

$$S(p,q) = F(\alpha_p, \alpha_q, \beta_p, \beta_q, \gamma_p, \gamma_q, \delta_{pq}, k_p, k_q, M) \quad (5.1.1)$$

Fusion

Suppose that clusters p and q are combined, then the function parameters for the new cluster (p+q) can be evaluated in terms of the original parameters, as follows:

$$\begin{aligned} \alpha_{p+q} &= \sum_j \left[\sum_{i \in p} X_{ij} + \sum_{i \in q} X_{ij} \right] \\ &= \alpha_p + \alpha_q \end{aligned} \quad (5.1.2)$$

$$\begin{aligned} \beta_{p+q} &= (k_p + k_q)^2 \sum_j U_{j(p+q)}^2 \\ &= (k_p + k_q)^2 \sum_j \frac{1}{(k_p + k_q)^2} \left[\sum_{i \in p} X_{ij} + \sum_{i \in q} X_{ij} \right]^2 \\ &= \beta_p + \beta_q + 2\delta_{pq} \end{aligned} \quad (5.1.3)$$

$$\gamma_{p+q} = \sum_j \left[\sum_{i \in p} X_{ij}^2 + \sum_{i \in q} X_{ij}^2 \right] = \gamma_p + \gamma_q \quad (5.1.4)$$

$$\begin{aligned} \delta_{(p+q)r} &= (k_p + k_q) k_r \sum_j U_{j(p+q)} U_{jr} \\ &= (k_p + k_q) k_r \sum_j \left[\frac{k_p U_{jp} + k_q U_{jq}}{k_p + k_q} \right] U_{jr} \\ &= \delta_{pr} + \delta_{qr} \end{aligned} \quad (5.1.5)$$

Division

Suppose that a cluster p is subdivided into two groups p' and p". Let the function parameters and centroid \underline{U}_p for p be known, and let the function parameters and centroid $\underline{U}_{p'}$ for one subset p' be computed, then we obtain the similarity $S(p', p'')$

between the two subsets of p as follows:

$$U_{jp''} = \frac{k_p U_{jp} - k_{p'} U_{jp'}}{k_p - k_{p'}}$$

$$\begin{aligned} \text{Hence } \delta_{p'p''} &= k_{p'}(k_p - k_{p'}) \sum_j U_{jp'} \left[\frac{k_p U_{jp} - k_{p'} U_{jp'}}{k_p - k_{p'}} \right] \\ &= \delta_{pp'} - \beta_{p'} \end{aligned} \quad (5.1.6)$$

From (5.1.2)-(5.1.4) we have

$$\alpha_{p''} = \alpha_p - \alpha_{p'} \quad (5.1.7)$$

$$\gamma_{p''} = \gamma_p - \gamma_{p'} \quad (5.1.8)$$

$$\begin{aligned} \beta_{p''} &= \beta_p - \beta_{p'} - 2\delta_{p'p''} \\ &= \beta_p + \beta_{p'} - 2\delta_{pp'} \end{aligned} \quad (5.1.9)$$

Binary data

Since every X_{ij} is 0 or 1, then if we obtain the attribute frequency vector $\underline{f}_t = (f_{1t}, \dots, f_{jt}, \dots, f_{Mt})$ for cluster t , the parameters reduce to

$$\left. \begin{aligned} \alpha_t &= \gamma_t = \sum_j f_{jt} \\ \beta_t &= \sum_j f_{jt}^2 \\ \delta_{pr} &= \sum_j f_{jp} f_{jr} \end{aligned} \right\} \quad (5.1.10)$$

These results are particularly useful in the case of the monothetic divisive strategy (see below).

Similarities between individuals

Suppose that cluster t is the single individual X_i ($k_t = 1$), then the parameters simplify to

$$\left. \begin{aligned} \alpha_i &= \sum_j X_{ij} \\ \beta_i = \gamma_i &= \sum_j X_{ij}^2 \\ \delta_{ik} &= \sum_j X_{ij} X_{kj} \end{aligned} \right\} \quad (5.1.11)$$

These results are used in the iterative relocation model where an individual is compared with a cluster, and also when F is used to compute a similarity matrix. In the latter case, if the data are binary then we may write

$$\begin{aligned} A &= \delta_{ik} \\ B &= \alpha_i - \delta_{ik} \\ C &= \alpha_k - \delta_{ik} \\ D &= M + \delta_{ik} - \alpha_i - \alpha_k \end{aligned}$$

where A , B , C and D are the conventional 2×2 table cell counts (Sect. 1.4) for individuals i and k . Hence, for example, the binary distance measure $(B + C)/M$ can be computed within F from

$$(\alpha_i + \alpha_k - 2\delta_{ik})/M$$

Similarity formulae

Using the above notation, we can express several intercluster

similarity measures in terms of the parameters of F. For example, the product-moment correlation coefficient r_{pq} may be used to compare two clusters p and q represented by their centroids \bar{U}_p, \bar{U}_q , as follows:

$$\begin{aligned} r_{pq} &= \frac{M \sum_{jp} U_{jp} U_{jq} - \sum_{jp} U_{jp} \sum_{jq} U_{jq}}{\left\{ \left[M \sum_{jp} U_{jp}^2 - \left(\sum_{jp} U_{jp} \right)^2 \right] \left[M \sum_{jq} U_{jq}^2 - \left(\sum_{jq} U_{jq} \right)^2 \right] \right\}^{\frac{1}{2}}} \\ &= \frac{M k_p k_q \delta_{pq} - k_p \alpha_p k_q \alpha_q}{\left\{ \left[M k_p^2 \beta_p - k_p^2 \alpha_p^2 \right] \left[M k_q^2 \beta_q - k_q^2 \alpha_q^2 \right] \right\}^{\frac{1}{2}}} \\ &= \frac{M \delta_{pq} - \alpha_p \alpha_q}{\left\{ \left[M \beta_p - \alpha_p^2 \right] \left[M \beta_q - \alpha_q^2 \right] \right\}^{\frac{1}{2}}} \end{aligned}$$

Similarly, the distance d_{pq}^2 between the centroids of clusters p and q is given by

$$\begin{aligned} d_{pq}^2 &= \frac{1}{M} \sum (U_{jp} - U_{jq})^2 \\ &= \frac{1}{M} \left[\sum U_{jp}^2 - 2 \sum U_{jp} U_{jq} + \sum U_{jq}^2 \right] \\ &= (k_q^2 \beta_p + k_p^2 \beta_q - 2 k_p k_q \delta_{pq}) / (M k_p^2 k_q^2) \end{aligned}$$

and the increase I_{pq} in the error sum of squares E is immediately obtained from

$$\begin{aligned} I_{pq} &= \frac{k_p k_q}{k_p + k_q} d_{pq}^2 && \text{(equation 8.2.4)} \\ &= \frac{(k_q^2 \beta_p + k_p^2 \beta_q - 2 k_p k_q \delta_{pq})}{M k_p k_q (k_p + k_q)} \end{aligned}$$

The following additional functional definitions of the quantitative

similarity measures given in Sect. 1.4 may now be verified:

Similarity ratio:

$$\frac{\sum U_{jp} U_{jq}}{\sum U_{jp}^2 - \sum U_{jp} U_{jq} + \sum U_{jq}^2}$$

$$= k_p k_q \delta_{pq} / (k_{qp}^2 \beta_p + k_{pq}^2 \beta_q - k_p k_q \delta_{pq})$$

Dot product: $\frac{1}{M} \sum U_{jp} U_{jq} = \delta_{pq} / (M k_p k_q)$

Cosine: $\frac{\sum U_{jp} U_{jq}}{\left\{ \sum U_{jp}^2 \sum U_{jq}^2 \right\}^{1/2}} = \delta_{pq} / \sqrt{\beta_p \beta_q}$

Size difference: $\frac{1}{M^2} \left[\sum U_{jp} - \sum U_{jq} \right]^2$

$$= \left[\frac{k_{qp} \alpha_p - k_{pq} \alpha_q}{M k_p k_q} \right]^2$$

Shape difference:

$$\frac{1}{M} \sum (U_{jp} - U_{jq})^2 + \frac{1}{M^2} \left[\sum U_{jp} - \sum U_{jq} \right]^2$$

$$= \frac{M(k_{qp}^2 \beta_p + k_{pq}^2 \beta_q - 2k_p k_q \delta_{pq}) - (k_{qp} \alpha_p - k_{pq} \alpha_q)^2}{(M k_p k_q)^2}$$

Dispersion: $\frac{1}{M} \sum (U_{jp} - \bar{U}_p)(U_{jq} - \bar{U}_q)$

$$= \frac{M \delta_{pq} - \alpha_p \alpha_q}{M^2 k_p k_q}$$

In addition, we have the following intercluster similarity criteria:

Average distance:

$$\frac{1}{M} \left[\sum (U_{jp} - U_{jq})^2 + S_p^2 + S_q^2 \right]$$

$$= \frac{k_q \bar{X}_p + k_p \bar{X}_q - 2\delta_{pq}}{M k_p k_q}$$

where $S_p^2 = \frac{1}{M k_p} \sum_j \sum_{i \in p} (X_{ij} - U_{jp})^2$ is the variance of cluster p.

Average distance measures the average of all the between-group squared d_{ik}^2 distances ($i \in p, k \in q$). This formula may be found in Wishart (1969e).

$$\text{Variance: } \frac{1}{M} S_{p+q}^2 = \frac{(k_p + k_q)(\bar{X}_p + \bar{X}_q) - \beta_p - \beta_q - 2\delta_{pq}}{M(k_p + k_q)^2}$$

Exceptions

Two of the formulae that we have considered cannot be expressed as variants of (5.1.1). They are:

Nonmetric coefficient:

$$\frac{\sum |U_{jp} - U_{jq}|}{\sum (U_{jp} + U_{jq})} = \frac{k_p k_q \sum |U_{jp} - U_{jq}|}{k_q \alpha_p + k_p \alpha_q}$$

The numerator $\sum |U_{jp} - U_{jq}|$ must be evaluated using the two centroid vectors $\underline{U}_p, \underline{U}_q$.

Information gain (binary data): $2 \Delta I = I_{p+q} - I_p - I_q$ where

$$I_p = M k_p \log k_p - \sum \left[f_{jp} \log f_{jp} + (k_p - f_{jp}) \log (k_p - f_{jp}) \right]$$

This function has to be evaluated from the original frequency vectors $\underline{f}_p, \underline{f}_q$.

To incorporate these two measures within the generalised tripartite procedure, a special function F is defined which uses, in addition to the 10 parameters of (5.1.1), the two centroid vectors \underline{U}_p and \underline{U}_q .

5.2 METHODS

The above results may now be used to define the following generalised tripartite clustering procedure.

Hierarchic fusion

We adopt the fusion strategy given in Section 2.1, and define the similarity between two clusters $S(p,q)$ using an appropriate formula with function F . Each cluster initially comprises one individual, so that the first cluster centroids are obtained by copying the original N vectors of observation data. A similarity matrix corresponding to the similarity criterion F is evaluated, and parameters α_i and γ_i are computed for each individual i . We also set $\beta_i = \gamma_i$ (from 5.1.11) and $k_i = 1$, for each individual i .

Having determined the two most similar clusters p and q , the new centroid vector \underline{U}_{p+q} is computed and stored (supposing that $p < q$, we replace cluster p with the new cluster $p+q$ and make cluster q inactive). Next δ_{pq} is computed from \underline{U}_p and \underline{U}_q , and the parameters for cluster p are modified using (5.1.2)-(5.1.5) as follows:

$$\alpha_p = \alpha_p + \alpha_q$$

$$\beta_p = \beta_p + \beta_q + 2\delta_{pq}$$

$$\gamma_p = \gamma_p + \gamma_q$$

$$k_p = k_p + k_q$$

$$k_q = 0 \quad (\text{renders cluster } q \text{ inactive})$$

The similarities $S(p,r)$ between the new cluster p and all other clusters r may now be computed using F : this necessitates the reading of centroid vectors \underline{U}_r for active clusters ($k_r > 0$), and the evaluation of δ_{pr} (all the other parameters of F are stored). These new similarities replace the p th row and column of the similarity matrix, and then the next fusion cycle is entered.

Monothetic division

In the terminology of Section 3.1, we divide any group t on the presence or absence of that binary attribute k for which $S(k+,k-)$ is minimum (where $k+,k-$ denote the appropriate two subsets of cluster t). Since two non-empty subsets are required, we need only consider those attributes k for which $0 < f_{kt} < k_t$.

From the frequency vector \underline{f}_t for cluster t , we compute the function parameters (5.1.10)

$$\alpha_t = \gamma_t = \sum f_{jt}$$

$$\beta_t = \sum f_{jt}^2$$

Let $\underline{f}_{tk} = (f_{1tk}, \dots, f_{jtk}, \dots, f_{Mtk})$ be the frequency vector

for the subset of cluster t defined by the presence of k , then we evaluate the following parameters for this subset:

$$\alpha_{tk} = \gamma_{tk} = \sum_j f_{jtk}$$

$$\beta_{tk} = \sum_j f_{jtk}^2$$

and the cross-product:

$$\delta_{t(tk)} = \sum_j f_{jt} f_{jtk}$$

We may now use results (5.1.6)-(5.1.9) to write the divisive similarity criterion as

$$S(k+,k-) = F(\alpha_{tk}, \alpha_t - \alpha_{tk}, \beta_{tk}, \beta_t + \beta_{tk} - 2\delta_{t(tk)}, \alpha_{tk}, \alpha_t - \alpha_{tk}, \delta_{t(tk)} - \beta_{tk}, f_{kt}^{k+}, f_{kt}^{k-}, M)$$

When a subset is divided on the presence/absence of that attribute k for which $S(k+,k-)$ is minimum, the two new frequency vectors \underline{f}_{tk} and $(\underline{f}_t - \underline{f}_{tk})$ are stored. Each subset may then be considered separately for further division, either by the nested or hierarchic subdivisive methods.

Iterative relocation

We denote by $(x_1, \dots, x_j, \dots, x_M)$ the data for an individual \underline{x} belonging to cluster p . Hence, from (5.1.11)

$$\alpha_x = \sum_j x_j$$

$$\beta_x = \gamma_x = \sum_j x_j^2$$

$$\delta_{tx} = k_t \sum_{jt} U_{jt} x_j$$

For the generalised iterative relocation procedure (Sect. 4.1) we may write

$$S(t,x) = F(\alpha_t, \alpha_x, \beta_t, \beta_x, \bar{y}_t, \bar{y}_x, \delta_{tx}, k_t, 1, M) \quad (5.2.1)$$

and using results (5.1.6)-(5.1.9)

$$S(p-x,x) = F(\alpha_p - \alpha_x, \alpha_x, \beta_p + \beta_x - 2\delta_{px}, \beta_x, \bar{y}_p - \beta_x, \beta_x, \delta_{px} - \beta_x, k_p - 1, 1, M) \quad (5.2.2)$$

In the special case when cluster p is the single individual \underline{x} , we compute

$$S(x,x) = F(\alpha_x, \alpha_x, \beta_x, \beta_x, \beta_x, \beta_x, 1, 1, M) \quad (5.2.3)$$

The two relocation tests (4.1.1) and (4.1.2) require \underline{x} to be removed from parent cluster p and added to 'best' other cluster q if

either
$$S(p-x,x) < S(q,x) \quad (\text{test 4.1.1})$$

or
$$S(p,x) < S(q,x) \quad (\text{test 4.1.2})$$

respectively; when cluster p comprises the single individual \underline{x} both relocation tests reduce to (4.1.3): move \underline{x} if

$$S(x,x) < S(q,x) \quad (\text{test 4.1.3})$$

which, if satisfied, causes cluster p to be eliminated.

The iterative relocation method can therefore be obtained with suitable substitutions of (5.2.1)-(5.2.3), as follows:

- 1) For each initial cluster t , the centroid \underline{U}_t and function parameters α_t , β_t , \bar{y}_t , and k_t are computed and stored; also computed are the function parameters α_i and β_i for each individual i .

2) Individuals $\underline{x} = (x_1, \dots, x_M)$ are examined sequentially, and for each \underline{x} the cross-products with all clusters t -

$$\delta_{tx} = k_t \sum U_{jt} x_j$$

are evaluated. The appropriate similarity functions (5.2.1), (5.2.2) and (5.2.3) are then calculated, and \underline{x} is tested for relocation. If the chosen relocation test (4.1.1)-(4.1.3) is satisfied, then \underline{x} is moved from parent cluster p to the 'best' other cluster q , and the cluster parameters are immediately adjusted as follows:

$$\alpha_p = \alpha_p - \alpha_x$$

$$\beta_p = \beta_p + \beta_x - 2\delta_{px}$$

$$\gamma_p = \gamma_p - \beta_x$$

$$k_p = k_p - 1$$

$$\alpha_q = \alpha_q + \alpha_x$$

$$\beta_q = \beta_q + \beta_x + 2\delta_{qx}$$

$$\gamma_q = \gamma_q + \beta_x$$

$$k_q = k_q + 1$$

3) The procedure stops when no additional relocations occur during one complete population scan, or when MAXIT scans have been completed.

5.3 DISCUSSION

It is immediately obvious that the generalised tripartite

procedure described above by no means encompasses the full range of methods previously surveyed (Chapters 2-4). Perhaps the greatest omission is the group of hierarchical linkage techniques (Sect. 2.2) which could not be given a generalised treatment without allowing for the manipulation of the similarity matrix. Single, average and complete linkage can, however, be obtained from the reasonably efficient combinatorial algorithm (Chapter 8) so that their omission from the tripartite procedure is therefore not so severe. Also excluded are the polythetic divisive techniques, which, by the very nature of their complex methodology, cannot be incorporated within a general-purpose computer program. However, although these methods are excluded there is no reason why both the Edwards and Cavalli-Sforza method and dissimilarity analysis should not use F for the comparison of subset with complement. The inclusions and exclusions of the generalised tripartite procedure may now be summarized as follows:

(i) Hierarchic fusion

INCLUDED

- (a) All centroid sorting options
- (b) Ward's method (increase in E)
- (c) Information-analysis ($2\Delta I$)
- (d) Average linkage with distance, dot product and dispersion (see Sect. 8.3)
- (e) Median distance, provided that we may set

$$U_{j(p+q)} = \frac{1}{2}(U_{jp} + U_{jq}) \quad \text{at fusion}$$

EXCLUDED

- (a) Most linkage methods
- (b) Some combinatorial coefficients: e.g. flexible and McQuitty's similarity analysis (Sect. 8.1)

(ii) Monothetic division

INCLUDED

- (a) All centroid divisive options
- (b) Decrease in E
- (c) Information statistic ($2\Delta I$)

EXCLUDED

- (a) Association analysis
- (b) Divisive group analysis

(iii) Iterative relocation

INCLUDED

- (a) All k-mean variants
- (b) Decrease in E
- (c) Information statistic
- (d) Generalised centroid method

EXCLUDED

- (a) Most cohesion functions
- (b) Thorndike's average within-group distance criterion
- (c) Agglomerative group analysis

Perhaps the greatest single asset of the generalised tripar-

tite procedure is that it enables different similarity measures to be easily compared in empirical studies. A single function subprogram F may be written to evaluate all of the similarity formulae given in Section 5.1, and it is then available to be called by several different clustering programs. In particular, F may also call another function which can be reprogrammed by a user. For example, within the 'CLUSTAN' suite of computer programs (Chapter 10) there exists a 'USER' facility whereby a function subprogram may be rewritten to evaluate some new similarity measure with all the clustering programs. Thus, with the St. Andrews IBM 360/Model 44 version of CLUSTAN IA, the new similarity measure

$$\frac{\sum U_{jp} U_{jq}}{\sum (U_{jp} + U_{jq})}$$

would be fully incorporated within the generalised tripartite procedure using the following job step:

```
//SYSRDR ACCESS DW(USER), 191 = 'SA45V1'  
/*  
  
FUNCTION USER (AP,AQ,BP,BQ,GP,GQ,DPQ,KP,KQ,M)  
USER = DPQ/(KQ*AP+KP*AQ)  
  
RETURN  
  
END  
  
/*
```

(it is assumed that the denominator $KQ*AP+KP*AQ$ does not become

indeterminate - if this were possible, then an appropriate test should be included).

Finally, we observe that the generalisation of similarity measures within F yields reasonably efficient computer programs. This is because the α , β , γ and k parameters may be stored for all clusters (excepting the +ve subsets considered within monothetic division) and for individuals (in the case of iterative relocation). Hence, any similarity $S(p,q)$ can be evaluated from the appropriate formula, once the cross-product

$$\delta_{pq} = k_p k_q \sum U_{jp} U_{jq}$$

has been computed. The computation of a similarity coefficient is therefore reduced to roughly M multiplications and additions, plus the appropriate formula evaluation; this compares very favourably with the direct computation of such coefficients as product-moment correlation using the centroid vectors.

CHAPTER 6: THE PROBABILISTIC MODEL

6.1 MINIMUM-VARIANCE TECHNIQUES

In 1914, the astronomer H. N. Russell plotted the temperature against luminosity of visual stars on a scatter plot, which is now known as the H-R diagram, and classified the stars into two groups which he called "giants" and "dwarfs". The diagram in figure 6.1.1 is reproduced from the H-R diagram given in Struve and Zeberg's (1962) which shows the dwarf star sequence as an elongated swarm from bottom right to top left, and the giant sequence as the cluster at top right. Forgey (1964) applied Ward's method to the H-R diagram and obtained the final classification into two groupings which is shown by the partition line of figure 6.1.1. The classifications of Russell and Ward clearly do not coincide, and the conclusion must surely be that, for the astronomer's purpose anyway, Ward's method failed. This chapter is devoted to an examination of the reasons for that failure, a reappraisal of what a 'natural' grouping procedure should theoretically achieve, and the author's contribution of theory and method designed for taxonomic purposes.

The term 'minimum-variance' has been used by Forgey (1964, 1965) to describe the basis of those methods which attempt to minimise the within-group sum of squares. In this context, any method which imposes some form of constraint on the spread, or variance, of clusters of points is included in the category. The classical example of this concept is exhibited by complete linkage (Sørensen, 1948), and the minimum-variance approach is epitomised by Sørensen's statement "only one demand may justly be made on the nature of the vegetation in the limited area under investigation namely that it be homogeneous with as much approximation to that mathematical concept as nature can offer." To impose the requirement that a plant community should exhibit as near total homogeneity as is possible, that is, without any major factor of variation, is probably a perfectly valid constraint in the context of vegetation analyses.

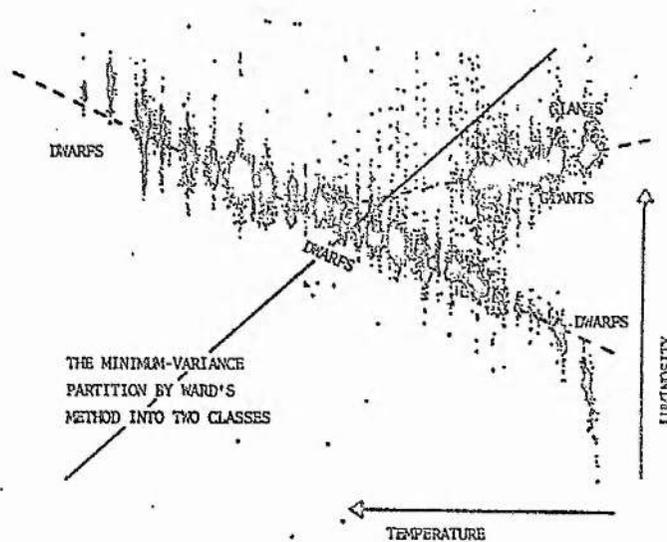


Figure 6.1.1. The Struve-Zeberg's H-R diagram for visible stars showing Russell's classification and the two class partition obtained by Ward's error sum method.

As Sørensen goes on to say "the various types of vegetation often are so insensibly merged as to form a sliding scale", and the use of a clustering method which searches for 'natural' or 'distinct' datum groupings would almost invariably fail to meet the plant ecologists' demands.

The consequence of such requirements has been that the major effort in the development of classification methods has been directed towards the definition of a satisfactory analysis which yields groupings that possess the minimum-variance property. Wishart (1969e) has shown that thirteen methods of cluster analysis, when compared on the basis of d^2 , possess a constraint of the minimum-variance type. These thirteen methods, which are merely a representative sample of perhaps a much larger list of attempts, have been discussed earlier and their constraints are summarized in Table 6.1.1.

The underlying axiom of variance constraint seems to have been developed intuitively by these writers from the idea that a resultant group of individuals should be homogeneous in relation to the total set of variables. That is, each individual should be relatively similar to every other individual in the same cluster for each variable. Expressed in geometric terms, the swarm of points which constitutes a minimum-variance cluster would be of spherical shape and should not possess any major axis of variation. Ideally, a principal components analysis of the cluster subset alone

| | <u>Author</u> | <u>Minimised Factor</u> |
|-----|---|-----------------------------|
| 1. | Sørensen: P-Q | r |
| 2. | MacNaughton-Smith: P-Q | r |
| 3. | Ward: P-Q | I_{PQ} |
| 4. | Sokal and Michener (centroid): P-Q | D_{PQ}^2 |
| 5. | Sokal and Michener (pair group): E-P | $d_{EP}^2 + S_P^2$ |
| 6. | Lance and Williams (group average): P-Q | $D_{PQ}^2 + S_P^2 + S_Q^2$ |
| 7. | Bonner: E-P | r |
| 8. | Hyvarinen: E-P | r |
| 9. | Ball and Hall: E-P, P-Q | $d_{EP}^2, S_P^2 \leq MS^2$ |
| 10. | MacQueen: E-P, P-Q | r |
| 11. | Sebestyen: E-P | r |
| 12. | Jancey: E-P | d_{EP}^2, I_{PQ} |
| 13. | Forgey: E-P | I_{PQ}, d_{EP}^2 |

- Notes:
- I_{PQ} - error sum of squares optimisation at fusion of clusters P and Q.
 - r - cluster subset diameter or radius constraint.
 - D_{PQ} - distance between centroids of clusters P and Q.
 - d_{EP} - distance from element E to centroid of cluster P.
 - S_P^2 - variance of cluster P.
 - M - number of variables.
 - S^2 - user variance threshold.
 - (P-Q) - relation between clusters P and Q.
 - (E-P) - relation between element E and cluster P.

Table 6.1.1. A comparison, based on Euclidean distance, of some clustering methods which exhibit variance constraints

should reveal no major difference between successive latent roots indicating that the dispersion is isotropic.

The classification of stars obtained by Russell (1914) from the H-R diagram does not possess the minimum-variance property; in fact, Russell writes "if we could put on it (the H-R diagram) some thousands of stars we would find that the points representing them clustered principally close to two lines." These lines, dotted in figure 6.1.1, would correspond to the grouping, suggested by Russell, of stars into the two categories "giants" and "dwarfs". The sequence of dwarf stars clearly has no minimum-variance property, and indeed the idea of splitting this band into sections obviously did not occur to Russell. What impressed him was that his scatter diagram revealed two distinct modes, and it is therefore to be expected that the classification of the H-R diagram obtained by Forgey (1965), using Ward's method (1963) for optimising the error sum of squares, would never coincide with Russell's conclusion.

In their book Numerical Taxonomy, Sokal and Sneath (1963) describe the traditional method by which taxa are defined as follows: "a search for characters reveals that within a subgroup A (of the population) certain characters appear constant, while varying in an uncorrelated manner in other subgroups. Hence a taxon is described and defined on the basis of this character complex X. It is assumed that this taxon is a monophyletic

or 'natural' taxon." The mathematical interpretation of the constant character complex would satisfy the minimum-variance criterion since a representation of the subgroup A in the character space X would yield a small variance spherical swarm of points. Those writers who have adopted the minimum-variance approach in numerical taxonomy have extrapolated this notion to the extent that numerically derived taxa are defined by its converse. That is, for a set of characters P, a 'natural' taxon is a subgroup of individuals for which P takes constant values. Notice that the traditional taxonomist, according to Sokal and Sneath, defines the taxon from a "search which reveals that certain characters appear constant within a subgroup A." This implies, as is usually the case, that while the character subspace defined by the complex X has the minimum-variance property, the complementary subspace, defined by those characters in P-X, certainly does not. It follows that the geometric properties of the swarm for taxon A in the subspace P-X would not satisfy the minimum-variance criterion, and consequently the same would be true of the total character space P. This can be easily verified by opening a flora at any page and selecting a species at random. Several of the characters would almost certainly be defined within wide limits indicating variation, while the distinctive characteristics of the species are probably indirect combinations of characters (for example leaf length and breadth might

have wide limits of variation, when the distinctive characteristic of the leaf is, in fact, its shape).

The following particular objections to the minimum-variance approach are now outlined and discussed with liberal reference to illustrations:

The minimum-variance solution produces clusters which are -

- (a) modified by changes in the character set,
- (b) transformation dependent,
- (c) destroyed by the introduction of non-relevant characters,
- and (d) sometimes partitioned by artificial and unsatisfactory boundaries.

In order to illustrate these points, consider the two artificial species A and B which have the following characters:

| | A | B |
|-------------------|---------|--------|
| LEAF LENGTH | 4-10 cm | 4-7 cm |
| LEAF BREADTH | 4-10 cm | 1-2 cm |
| NUMBER OF FLOWERS | 5-7 | 5-7 |

The discriminating feature present in this restricted character set is clearly the shape of the leaf; species A (which might well be *Nymphoides Peltata*) has spherical or orbicular leaves, while species B (perhaps *Myosotis Sylvatica*) possesses long, or ovate-spathulate leaves. The histogram of the ratio leaf length/breadth in figure 6.1.2(a) indicates two well defined modes corresponding to A and B such that the species would almost certainly be deter-

mined by a minimum-variance method using this single variable. On the other hand, the elongated swarms in figure 6.1.2(b), obtained from a scatterplot of length vs. breadth, do not possess this property, and the partition lines indicate the probable

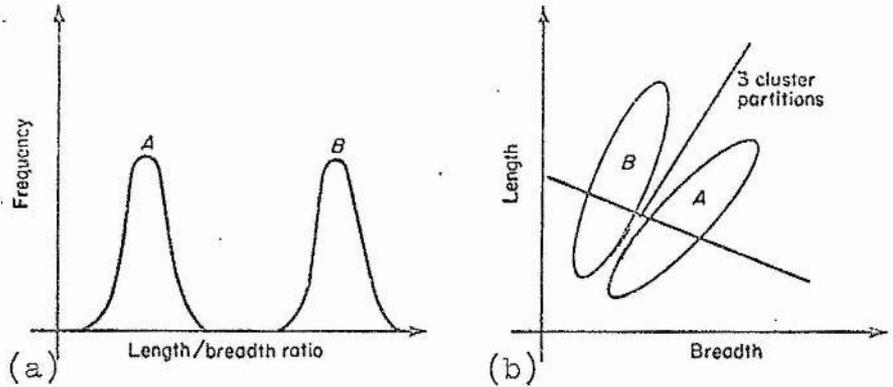


Figure 6.1.2(a) Histogram of the length/breadth leaf ratio showing two well-defined modes associated with the artificial species A and B.

Figure 6.1.2(b) Scatterplot of leaf length versus breadth showing elongated swarms for species A and B, and the probable division into 3 classes by a minimum-variance method.

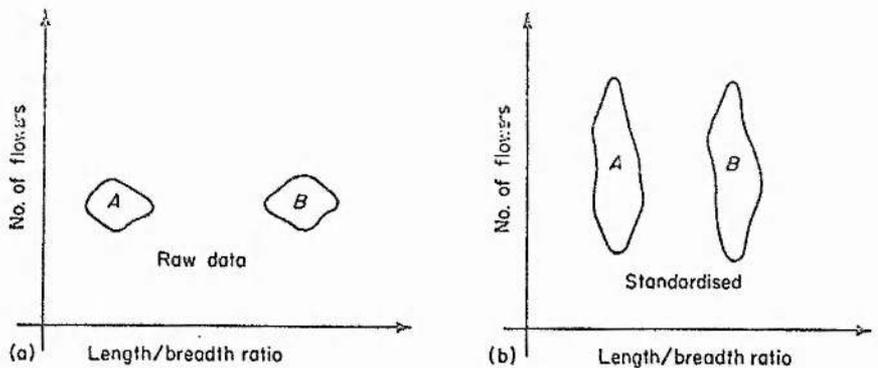


Figure 6.1.3. Two scatterplots of 'no. of flowers' versus 'leaf length/breadth ratio'. Prior to standardisation, (a), the swarms are spherical; after standardisation, (b), the swarms are elongated and fail the minimum-variance criterion for separation.

division, when three clusters are requested, which a minimum-variance method would derive. It follows that the classifications obtained from such methods depend on the original choice or manipulation of the character set.

In figure 6.1.3(a), the length/breadth ratio is plotted against the number of flowers per plant. Since the latter character is constant and non-diagnostic for both species, the swarms cluster well and satisfy minimum-variance conditions. However, the overall horizontal variance is considerably greater than the variance in the vertical direction, and consequently after the usual standardisation of variables to unit variance, the swarms would be elongated, figure 6.1.3(b), and no longer possess the minimum-variance property. Clearly, when non-diagnostic (or non-discriminate) variables are included in the character set, such transformations cause the elongation of the clusters in the subspace defined by the non-diagnostic character complex. The result might well be that the swarms are not separated in their entirety, but partitioned by arbitrary boundaries (this is particularly true of hierarchical systems where the final fusions are often inefficient).

The same situation arises when irrelevant characters are introduced. Suppose that the number of dogs within a ten-mile radius of each specimen is plotted against leaf length/breadth. If we can assume that this variable is normally distributed and

independent of the other, then the resultant scatter diagram, after standardisation, would be very similar to figure 6.1.3(b). An irrelevant character, simply by its non-relevance, can be taken to be non-diagnostic, and therefore the objections are the same as for the previous category. This possibility is less likely to arise in the context of the classification of plants where the characters are fairly well defined, but it would be a problem in a situation such as the classification of diseases. A patient's height might have absolutely nothing to do with his likelihood of contracting one of a group of diseases, and the inclusion of this variable, which might seem reasonable at sampling time, would result in a similar disease-swarm elongation effect.

Finally, a major objection to the partitions obtained by a minimum-variance method is that they may easily cut across a dense swarm of points (e.g. figure 6.1.1 and 6.1.2(b)) with the result that on either side of the partition there will be a fairly large number of individuals which are practically identical. This defeats the objective of the analysis.

6.2 NATURAL-CLASS METHODS

Forgey (1964, 1965) states a case for describing natural phenomena, taxa etc., in terms of disjoint data modes. He writes "when we see a frequency distribution on a continuous variable, we generally expect it to have a single mode. If there are definitely several modes, we are likely to consider our sample a

mixture of several distinct types of cases ... A scatterplot showing two distinct 'clusters' of data points suggests that the sample is a mixture of two more distinct classes of individuals ... On the other hand, when the typical single cloud of data points is observed, it would seem arbitrary to divide the sample into any number of discrete classes." In each of the examples, figures 6.1.2 and 6.1.3, the data swarms have different shapes, but one feature in common, namely, the cluster swarms for species A and B are always separated. It would seem, therefore, that the ideal classification method for taxonomic purposes is one which can firstly tell us if there exists more than one data mode, and secondly, resolve distinct data modes regardless of their shape or variance. Classification methods which would be suitable for this purpose, that is, those which do not possess some form of variance constraint, are rare. Perhaps the most widely known is Sneath's method (1957) of single linkage (Sect. 2.2). A distance threshold r is determined by the user, and any two data points separated by a distance not greater than r are connected by a 'bond'. In this way, a cluster is described by a lattice of linked points, having the property that each element is 'similar' to at least one other element, while two disjoint clusters are separated by a distance which exceeds r . This method has been severely criticised (Lance and Williams, 1967a; Williams, Lambert and Lance, 1966; Jardine and Sibson, 1968) for its so-called

'chaining-effect', a phenomenon which is most easily explained by reference to a diagram such as figure 6.2.1. Two distinct modes (containing points 1, 2, 3, 4, 5 and 8, 9, 10, 11, 12) are joined by the 'noise' or 'chaining' points 6 and 7 which, by their crucial siting cause the lattice of links to be extended between the modes and results in their fusion. As Forgey has pointed out, noise is a perfectly natural phenomenon of biological data where continuous variables are often normally distributed. One expects a cluster in the multidimensional space to exhibit a dense centre, or mode, which is surrounded by a cloud or noise. When attempting to classify the H-R diagram, figure 6.1.1, Forgey found that single linkage failed due to the chaining effect which occurred in the noise data that forms the 'saddle' region between the giant and dwarf star sequences. From a series of empirical trials using artificial normally-distributed data, Forgey concluded "that the method (single-linkage) performed well with very distinct clusters of any shape, but as soon as a moderate amount of noise was added the results quickly became quite erratic." Other writers have criticised the method for its chaining effect, notably Lance and Williams (1967a) who write "we submit that nearest-neighbour sorting should be regarded as obsolete." However, the evidence is totally empirical and few attempts seem to have been made to correct this failing.

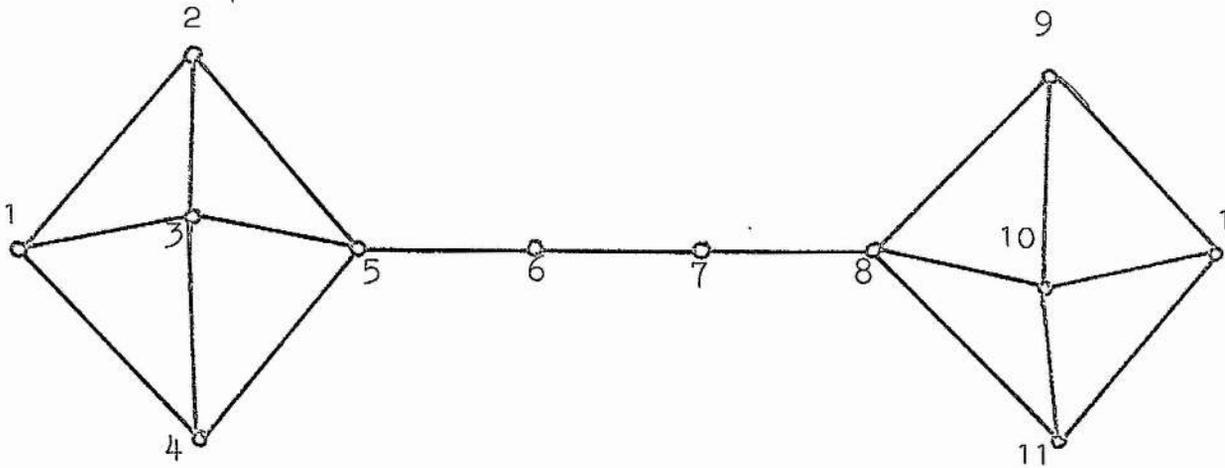


Figure 6.2.1. An illustration of the 'chaining' effect - the lines join points which are 'similar' at some critical distance threshold. Single-linkage fails to resolve two distinct clusters due to the siting of 'noise' points 6 and 7.

Reducing the chaining-effect

The obvious approach to the reduction of chaining-effects is to remove all noise data and then apply single linkage to the dense regions that are left. Forgey describes a method by which, he claims, the points are subjected to "those physical events that would occur if data points actually had mass, exerted gravitational pulls upon each other, and moved, but were not able to gather momentum." The idea is that the noise surrounding a mode would contract towards the dense centre causing the cluster to become more distinct. Unfortunately, the actual algorithm for this method does not appear to be documented, so a more detailed discussion is prohibited. Forgey does, however, concede that the method failed a test to resolve "pairs of elongated parallel clusters, even when they were made quite distinct". This may well be due to the fact that the clusters will inevitably attract each other causing them to collapse together at their mutual centre of gravity, but since the method is not known, this is pure speculation.

Sneath (1966), however, has documented a gravitation-simulation algorithm for a method developed essentially to discover the 'shape' of clusters and extrapolate a hierarchical pattern. The method seems to be theoretically satisfactory, but the complexity of its programming and user control appears to have prohibited its wide usage.

6.3 SINGLE-LEVEL MODE ANALYSIS

Wishart (1968, 1969e) proposes a method called 'Mode analysis' which is applied directly to the problem of removing noise data prior to single-link clustering of the denser data modes. The traditional statistical method for detecting the modes of a single continuous variable is to construct a histogram. A frequency threshold k is chosen, and the saddle regions (corresponding to class intervals which have a frequency that is less than k) are provisionally removed. The modes, if there are more than one, will now appear as groupings of the remaining class intervals (those which have a significant density) which are adjacent. Finally, the saddle regions are re-entered and associated with their nearest modes. For a scatterplot in two or more dimensions, this method is generalised to the contingency table technique whereby a grid of rectangular cells is constructed over the distribution, and the frequency of each cell is computed. Unfortunately, in order to extend the idea to M dimensions (where M is large) a contingency table of p^M cells is required when each variable is divided into p class intervals. Clearly, this would have its limitations. An alternative might be to retain only those cells which contain data points. Since each datum can only lie inside one cell, a maximum of N cells would have to be retained. The use of rectangular cells does, however, introduce a certain inefficiency, since to compare a data point with one

cell would require M computed tests (one for each dimension). The ideal solution is to use spherical cells, since the comparison is achieved by simply measuring the distance from the point to the cell centre and comparing this with the cell radius. To ensure the maximum accuracy of mode detection when N is small, it is proposed that a cell should be located about each point, and the one-level algorithm can be stated as follows:

- (a) Select a distance threshold r , and a frequency (or density) threshold k .
- (b) Compute the triangular similarity matrix of all inter-point distances.
- (c) Evaluate the frequency k_i of each data point, defined as the number of other points which lie within a distance r of point i (that is, those points inside a spherical cell of radius r centred at point i).
- (d) Remove the 'noise' or non-dense points, those for which $k_i < k$.
- (e) Cluster the remaining dense points ($k_i \geq k$) by single linkage at threshold r , forming the mode nuclei.
- (f) Reallocate each non-dense point to a suitable cluster according to some criterion. For the present program, each non-dense point is included in the cluster containing its nearest dense point.

6.4 HIERARCHICAL METHOD

A major criticism of the one-level test is that two thresholds r and k must be chosen by the user. This external control can be reduced by defining a hierarchical algorithm which is based on the order in which points become dense. The method can be summarized as follows:

- (a) Select the density threshold k , compute the inter-point distance matrix and the distances PD from each point to its k th nearest point.
- (b) Order the distances PD so that the smallest is first using the array KP as an index. Thus KP defines the order in which the data points become dense: point $KP(1)$ has the smallest k th distance $PD(1)$ and is first to become dense when $r = PD(1)$, point $KP(2)$ is second at $PD(2)$, and so on.
- (c) Select distance thresholds $PMIN$ from successive PD values, initialising a new dense point at each cycle. As the second and each subsequent dense point is introduced, the method tests the new point to determine one of three possible fusion phases:
 - either (i) the new point does not lie within $PMIN$ of another dense point, in which case it initialises a new cluster mode,
 - (ii) the point lies within $PMIN$ of dense points from one cluster only, and therefore the point is directly fused to that cluster,

- or (iii) the point falls in the saddle region, lying within PMIN of dense points from separate clusters, and the clusters concerned are fused.
- (d) Finally, a note must be kept of the nearest-neighbour distance DMIN between dense points of different clusters. When PMIN exceeds DMIN, the direct fusion of the two clusters separated by DMIN is indicated.

This algorithm is concisely represented by the flow chart in figure 6.4.1.

Output of classifications

It is conceivable that, at each cycle of the algorithm, all the non-dense points could be reallocated to the cluster nuclei and the cluster groupings made available. However, this leads to a vast collection of results which are confusing, and it is therefore desirable to restrict the output in some way. The fusion of a new dense point to an existing cluster ($c(i)$ of the algorithm) is probably the least significant step. This can be interpreted as the growth of a mode and corresponds to an information-gain for the cluster concerned, thus the previous grouping has a lower information content and can be considered of less value. Similarly, at the introduction of a new cluster nucleus ($c(i)$ of the algorithm), the groupings become outdated when the cluster subsequently 'grows' and increases in information-content. The really critical phases are therefore those at which existing

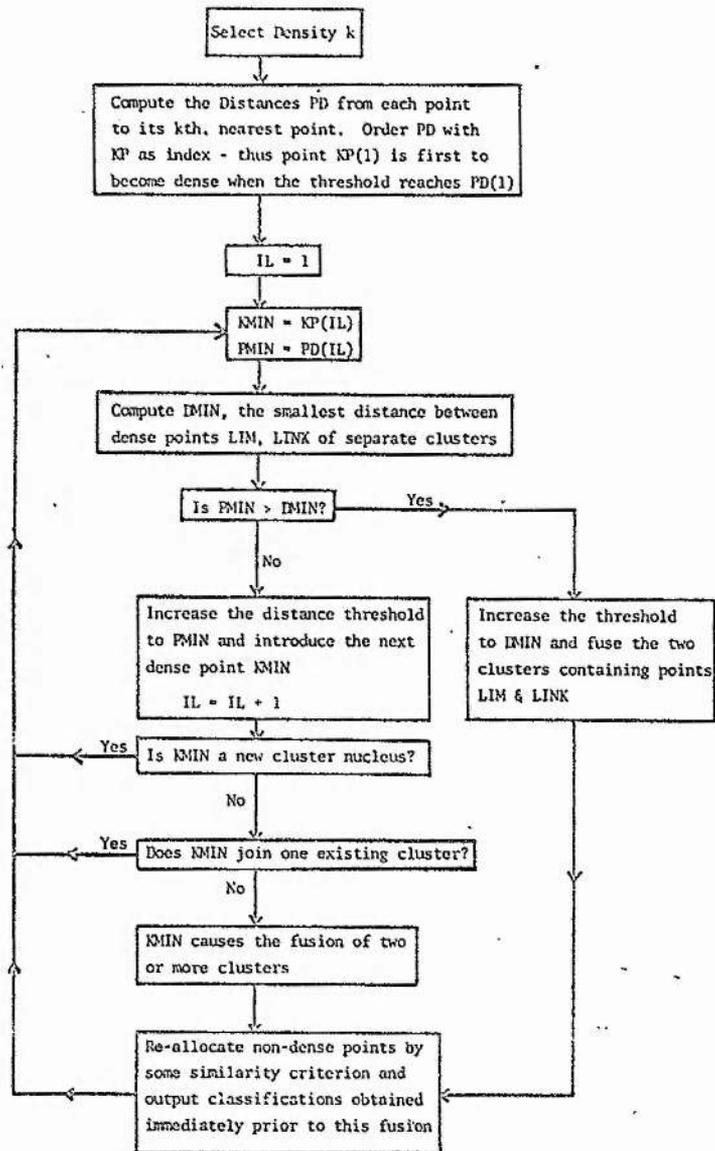


Figure 6.4.1. Flow Chart for the Hierarchical Mode Analysis Computer Program.

clusters are fused (c(iii) and d), and output is restricted to those groupings which are obtained immediately before such a fusion. Two alternative levels of classification are offered to the user: the nuclei level groups only those data points (including any which are non-dense) that lie within those spheres that correspond to dense points, while at the complete classification level, each non-dense point is allocated to the cluster containing its nearest dense point. Non-dense points which lie outside dense spheres are denoted unclassifiable at the nuclei level, while, for those users who demand a best-possible fit for all their cases, the complete level of classification allocates the entire population to the cluster modes.

Unique Features of Mode Analysis

1. For the first and last cycles of the analysis only one cluster is defined. Thus, at some intermediate stage, the number of clusters reaches a maximum that can be interpreted as the widest classification which is 'natural' or 'taxonomically significant'. It is possible that an analysis will never reveal more than one cluster, indicating that the data swarm is unimodal. In a large study of several real data matrices (population sizes ranging from 30 to 350), the method never defined a grouping of more than nine clusters, and the average analysis maximum was about six.

2. The useful range of the density threshold k is about 1 to

6 depending on the population size. For large data sets ($N > 200$), empirical trials indicate that values of k in the range 3 to 5 yield practically identical results. Thus the user control is severely restricted, and seldom critical.

3. When k takes the value 1, the algorithm degenerates, by definition, to nearest neighbour, making available this additional method as an option for very small data sets.

4. The number of separate classifications is severely limited, by the output control, to those groupings obtained prior to cluster fusions. During the trials, the largest number of groupings obtained was 24, while the average was about 11. In one case, the method generated only six groupings for a population of size 310.

6.5 DENSITY FUNCTION AND LARGE POPULATIONS

The histogram is usually a device for describing the shape of the probability density function of a single continuous variable. If the sample is sufficiently large, statistical tests such as chi-square can be employed to estimate the likelihood that the sampled distribution conforms to a theoretical probability density distribution, e.g. Normal, Poisson, etc., and when such tests prove positive we say that the sampled variable is distributed according to a particular probability function. By using contingency tables, this process is generalised to provide a means of estimating bivariate (e.g. binomial) and multivariate (e.g. multinomial) distributions. For Mode analysis, spherical

cells replace the rectanguloid contingency table cells, but in all other respects the process of estimating the modes of a probability function is identical. We can say that, for large sample size, a particular spherical interval radius r and a density threshold k , those spheres which are 'dense' are sample estimators of the regions for which a complex continuous probability density function $P(u)$ takes probability values in excess of some unknown limit p . In other words, the space defined by $P(u) \geq p$ is estimated by a covering of dense spherical intervals, and if P has two or more modes at the level of probability p , then the covering will be partitioned into two or more disjoint connected subsets of points. Furthermore, if it is the case that P has more than one mode, then we can reason, by Forgey's argument, that the population is a complex mixture of several more homogeneous subpopulations which can be isolated, using Mode analysis, by partitioning the covering of dense spheres into its constituent disjoint subsets. This evaluation holds only when the sample is of sufficient size, and of course, the larger the number of dimensions (variables) which are used, the larger is the sample space and consequently the sample size must be suitably increased. At present, the program (Wishart, 1968, 1969d) developed for Mode analysis can accommodate 999 cases, and uses each datum point to define a spherical interval. When really large populations are to be analysed, the theory would be satisfied if density spheres

are chosen about a selected subset of points. In fact the traditional histogram can be thought of as a sequence of spherical (one-dimensional) density intervals selected systematically through the range of variable values. This technique would be equally valid if the intervals were chosen about actual data points selected at random. Thus, in the multidimensional space of a sample of size 3000, 400 points selected at random could be used to define spherical density intervals in order to locate the population modes. The hierarchical method for a large population could therefore be obtained with the following algorithm:

1. Select a subset of q key points either systematically or at random, and compute the distances PD from each key point to its k th nearest point (from the entire population).
2. Compute the distances from each non-key point to its nearest key point.
3. Classify the key points alone using PD to define the order in which they become dense.
4. Using the complete classifications derived at each level of output, the grouping of the entire population is obtained by classifying each non-key point with the parent cluster of its nearest key point.

Finally, when particular accuracy is required, it is proposed to make a single movement of the centres of the spherical intervals in the direction of increasing density. One can imagine

that spherical intervals selected at random on the fringes of two close parallel clusters might cause chaining effects. If, however, each sphere centre is moved once to the centroid of those points which it initially contains, those which lie in the saddle regions would tend to separate and become disjoint.

6.6 AVERAGE DISTANCE AS DENSITY ESTIMATE

In the hierarchical algorithm, the array of k th distances PD is used to estimate the density of the space in the immediate vicinity of the points. This definition of PD allows us to compare Mode analysis with histogram and contingency table techniques, because $PD(i)$ defines precisely the radius of a spherical cell, surrounding the i th point, which is required to enclose k other points. It has also been claimed (Sect. 6.4) that, for large populations, values of k in the range 3 to 5 yield practically identical results. For this to be satisfied theoretically, PD must be a good estimate of local density, and also successive PD vectors derived for different (small) values of k should be well correlated. In practice, this is not always the case, usually because $PD(i)$ is a single distance between i and another point, and does not therefore take into account the configuration of points in the vicinity of i . Where the points are sparse, this will produce discontinuities in the vector of successive $PD(i)$ values, for any point i , obtained using different values of k . For example, in figure 6.7.2, which shows the principal

components diagram for a 9 variable survey of Poland, it is easily seen that successive PD(i) values for members of cluster 10 will not increase smoothly.

A better solution is to use the average A_i of the $(2k+1)$ least distances from point i as an estimate of the k th distance, and hence inverse estimate of local density. These values will not only increase smoothly with k , but also take into account the configuration of several points in a small region of the space when estimating density.

6.7 IMPROVED ALGORITHM

The following criticisms and proposed improvements to the probabilistic mode-seeking algorithm were prompted by a recent geographical application (Dawson, 1970) of hierarchical mode analysis. The country of Poland, described in terms of 318 urban and rural administrative units, was classified on the basis of 9 variables associated with population movements and indices of industrialisation for the period 1949-1965. These data were dominated by the rural units, with the result that when hierarchical mode analysis was used with the full population the large rural and semi-rural clusters emerged and fused before the relatively small, but important, regions of high industrialisation had appeared. The first analysis is shown by the dendrogram of figure 6.7.1(A), where clusters 3, 6, 7 and 8 are seen to emerge at threshold level 0.03; clusters 3, 7 and 8 are then fused

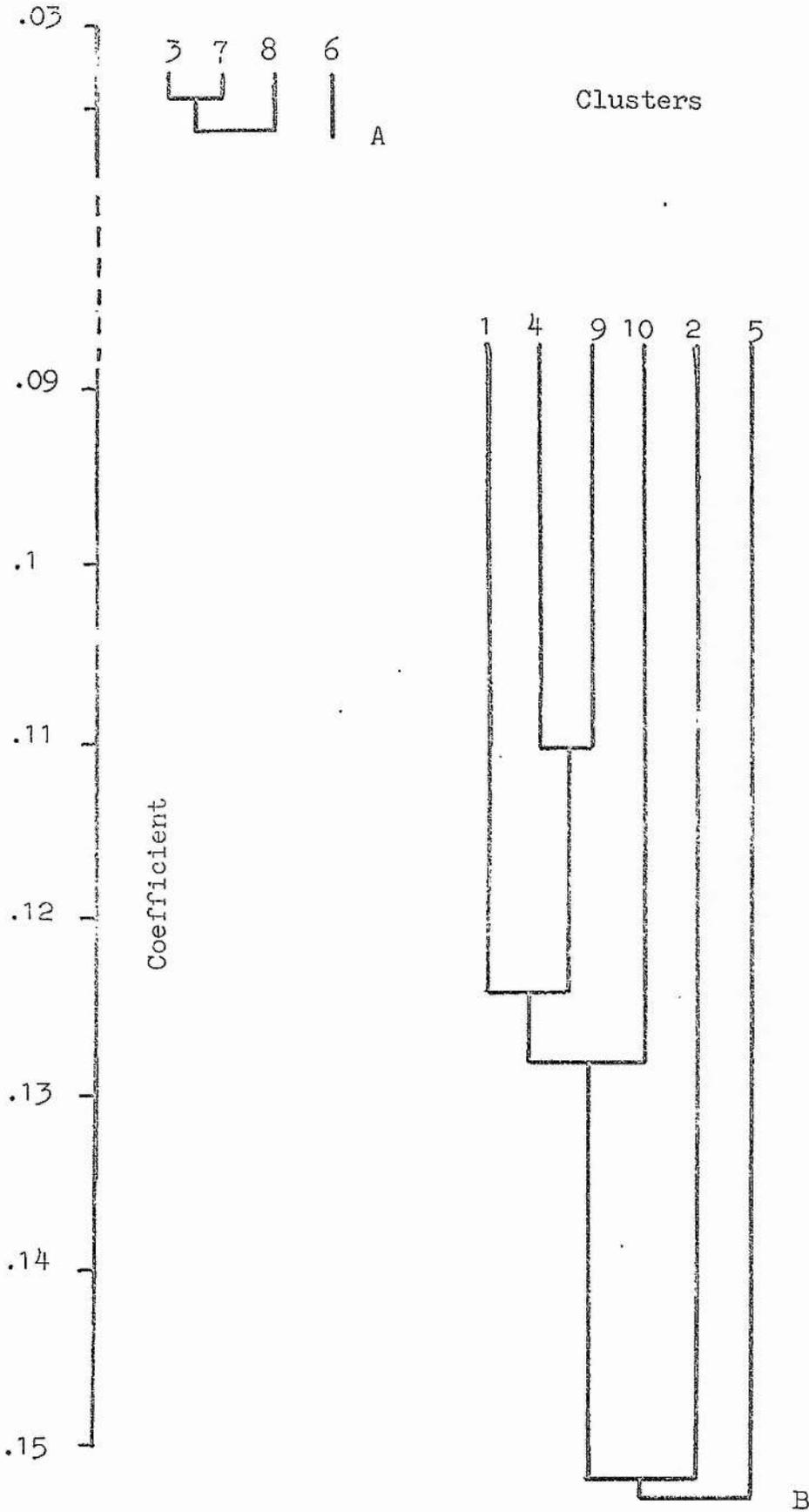


Figure 6.7.1. Dendrograms for both phases of the classification of Poland by hierarchical mode analysis: (A) First classification which finds the rural and semi-rural clusters; (B) classification of residual population which finds the industrial regions, and the tails of the rural clusters.

at level 0.04, losing their independent identities, before the industrial clusters (e.g. 9 and 10) emerge. In order to retain these four clusters it was found necessary to stop the analysis at level 0.03, remove the basic¹ classifications and then reclassify the residual population. Figure 6.7.1(B) shows the result of this second analysis, and the cluster characteristics for the ten groups thus obtained are given in Table 6.7.1. A very rough guide to the characteristics of six of these ten clusters is shown in Table 6.7.2, demonstrating that it was important from the geographers' point of view that clusters 3, 7 and 8 should be separated. Figure 6.7.2 shows the 3-dimensional display of points associated with the three principal components (which together account for 81.4% of the total variance) resulting from principal components analysis of the 9 x 318 data matrix. The inset of Figure 6.7.2 is Dawson's attempt to draw the regions of principal 3-space occupied by each of the 10 clusters.

The first important feature of this analysis is that

¹ 'basic' classifications were introduced to mode analysis (Pocock and Wishart, 1969) as an intermediate level between 'nuclei' and 'complete' to provide the user with a finer guide to peripheral objects. This level has now been abandoned in favour of an ordering of cluster members according to their goodness-of-fit (determined from their A_i values, and discussed later).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | All Poland |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------------|
| Number of constituent counties | 27 | 8 | 80 | 20 | 10 | 16 | 44 | 30 | 59 | 8 | 318 |
| Index of industrialisation 1949 | 8 | 9 | 15 | 24 | 24 | 39 | 42 | 75 | 109 | 258 | 75 |
| Index 1965 | 42 | 39 | 34 | 48 | 30 | 47 | 75 | 127 | 164 | 262 | 119 |
| Increase in industrial employment 1949-1965 (1949 = 100) | 669 | 574 | 299 | 337 | 189 | 191 | 240 | 220 | 256 | 128 | 203 |
| of which 1949-1960 (1949 = 100) | 518 | 236 | 232 | 236 | 233 | 170 | 193 | 178 | 203 | 119 | 169 |
| of which 1960-1965 (1960 = 100) | 132 | 246 | 128 | 147 | 85 | 113 | 124 | 124 | 123 | 108 | 120 |
| Increase in population 1950-1965 (1950 = 100) | 114 | 115 | 114 | 141 | 126 | 143 | 124 | 126 | 146 | 125 | 126 |

Table 6.7.1. Characteristics of the 10 clusters of administrative units obtained in the classification of Poland by hierarchical mode analysis.
Note: the 'index of industrialisation' is the number of jobs per 1000 population.

| CLUSTER | 3 | 7 | 8 | 6 | 9 | 10 |
|---------------------------|---|---|---|----|----|----|
| INITIAL INDUSTRIALISATION | L | M | A | M | H | VH |
| GROWTH OF INDUSTRY | L | M | A | L | H | L |
| GROWTH OF POPULATION | L | A | A | VH | VH | A |

Table 6.7.2. Rough guide to the characteristics of six of the 10 clusters obtained from the classification of Poland by mode analysis.
Key to symbols: L - LOW; M - MODERATE; A - AVERAGE; H - HIGH; VH - VERY HIGH.

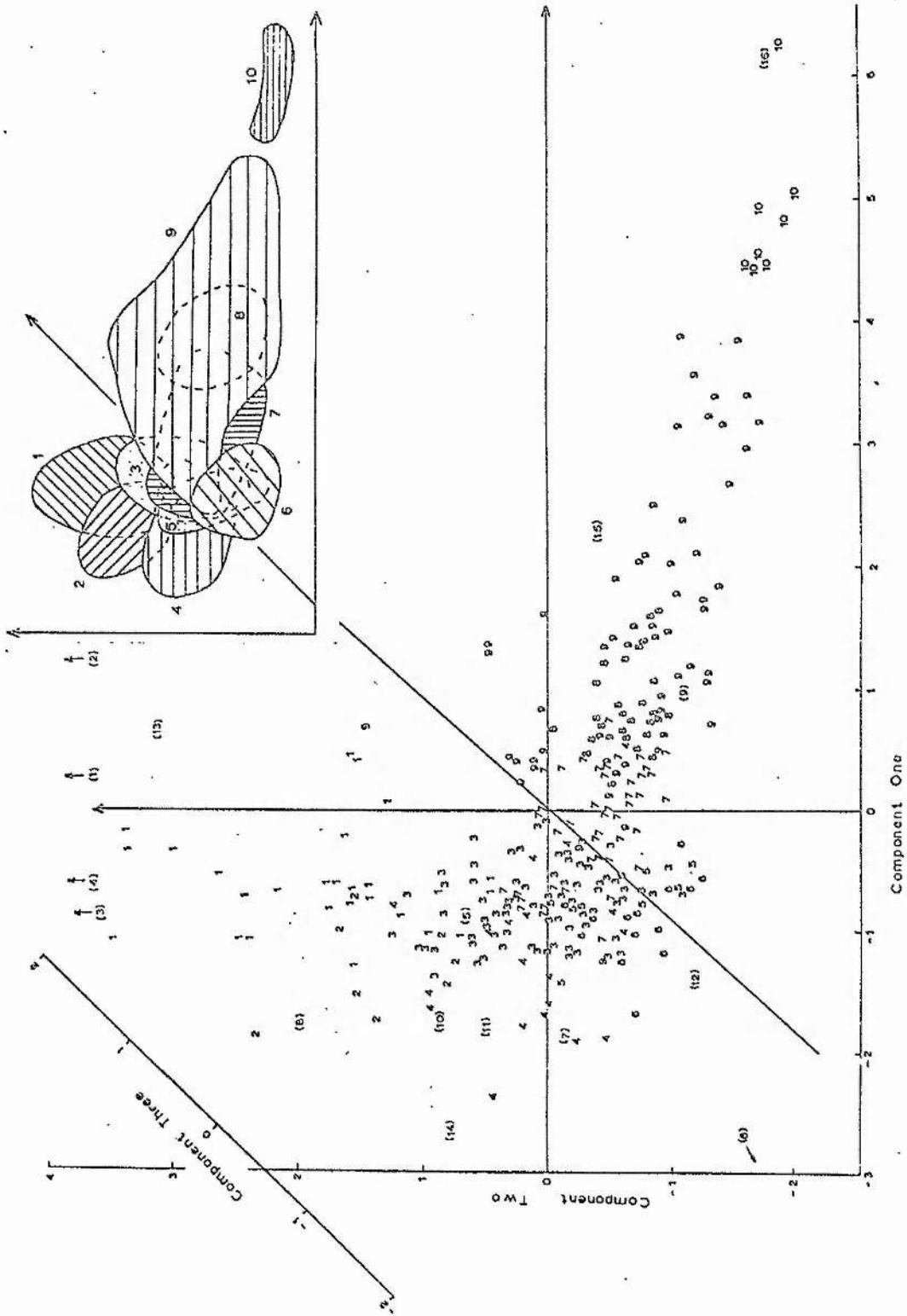


Figure 6.7.2. Scatter diagram for the first three principal components obtained with the Polish data, showing the relative positions of the 10 clusters. Object codes in parentheses are 16 counties unclassified by hierarchical mode analysis. 81.4% of the total variance is explained by the principal 3-space. Inset: a rough representation of the extent of the 10 clusters.

clusters 1, 2, 4 and 5 constitute variations of the dominating rural groups: in fact, these four groups, which were obtained during the second phase of the analysis, can be regarded as the 'tails' of the distribution for cluster 3.

The second point that should be made is that, in this instance, the method of analysis is rather unsatisfactory. Despite the fact that the groups so obtained were meaningful to the geographer (see Dawson, 1970), the division of the population into two parts in order to obtain reasonable groups demonstrates not only a fault of mode analysis, but suggests that the distribution of points formed a continuum (shown by Figure 6.7.2) which would have been better partitioned using a minimum-variance method² such as Ward's. However, the fault of mode analysis shown by this result applies equally well to studies which require the probabilistic solution, and therefore an attempt is now made to explain and correct the procedure for cluster fusion in the hierarchical algorithm.

²regrettably, a computer program for such a method was not readily available at the time of this analysis.

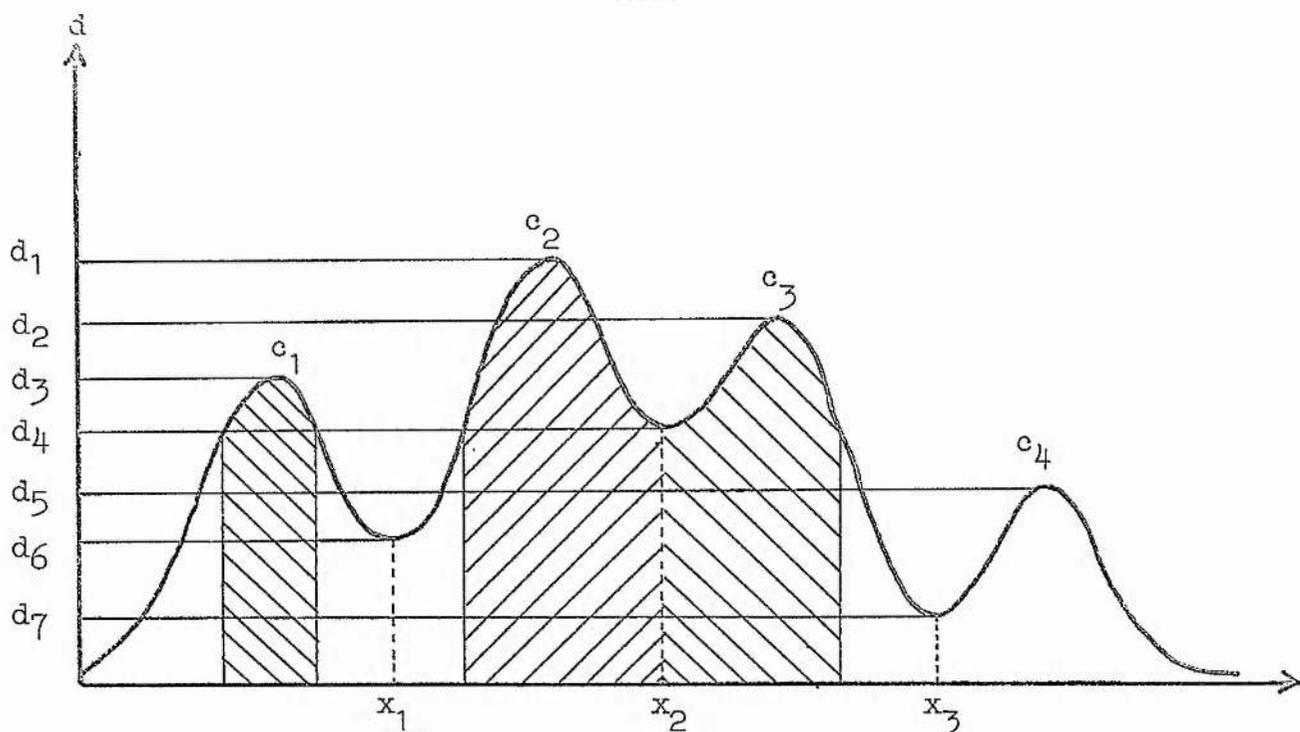


Figure 6.7.3. Histogram showing four modes which illustrates the failure of hierarchical mode analysis to resolve all four clusters simultaneously. c_2 and c_3 are fused at level d_4 , before cluster c_4 has been recognised.

Suppose that in a biological experiment, the unlikely histogram of figure 6.7.3 is obtained for a single variable x . In what way can we expect the biologist to respond when presenting his results? It is likely that, having checked his working and repeated the experiment, he would conclude that his observations can be grouped into four classes which are determined by the partition values x_1 , x_2 and x_3 . He might also

observe that the two classes denoted c_2 and c_3 , which comprise the bulk of the observations, could be grouped together leaving classes c_1 and c_4 as unusual entities. If it is required that a 'natural' mode-seeking method should repeat this intuitive analysis, then hierarchical mode analysis fails the test for the following reasons:

It is assumed that the k th least distances PD, or the average $(2k+1)$ distances A_i , are inversely proportional to the density ordinate d . If this is the case, we can expect hierarchical mode analysis to function as follows:

1. $A_i \propto 1/d_1$ Initiate cluster c_2
2. $A_i \propto 1/d_2$ Initiate cluster c_3
3. $A_i \propto 1/d_3$ Initiate cluster c_1
4. $A_i \propto 1/d_4$ Fuse c_2 with c_3
5. $A_i \propto 1/d_5$ Initiate c_4
6. $A_i \propto 1/d_6$ Fuse c_1 with $(c_2 + c_3)$
7. $A_i \propto 1/d_7$ Fuse c_4 with $(c_1 + c_2 + c_3)$

The failure of hierarchical mode analysis to reproduce the biologist's intuitive result is illustrated by stage 4, where two of the clusters are combined before the fourth (c_4) is initiated. In figure 6.7.3, the shaded areas underneath the histogram correspond to those observations which are classified at the nuclei level for threshold d_4 . We observe that, if complete groupings are obtained, then clusters c_1 and c_2 are extended roughly to the point x_1 (the median of the

unshaded region at the x_1 saddle), and the entire unshaded region $x \geq x_3$ will be clumped with cluster c_3 . This result is in complete contrast to that obtained at stage 7, where c_4 is recognised as independent and c_1 is grouped with $(c_2 + c_3)$. Furthermore, the nuclei classifications at stage d_4 comprise large regions of clusters c_2 and c_3 , and a relatively small region of cluster c_1 : hence c_1 is poorly represented at level d_4 .

Finally, as already suggested (Wishart, 1969a), the present method of reallocation of non-nuclei points to obtain the complete classifications is rather unsatisfactory. Each such individual is grouped with the cluster containing its nearest dense point. This means that all peripheral objects are allocated, regardless of how distant they are from the cluster centres, and the final cluster partitions do not necessarily follow density saddles: with hierarchical mode analysis, a partition surface bisects the space which separates the surfaces defining two cluster nuclei.

Towards the end of hierarchical mode analysis, each peripheral object is allocated as it becomes dense, by order determined from its A_i value, to its nearest cluster nucleus; since it is assumed that A_i is inversely proportional to density, this method of 'growing' cluster nuclei on a density basis conforms better to the concept of density surfaces defining cluster spaces

ordinates d_i are not known, but the estimates A_i observed at steps 4, 6 and 7 are used.

The full algorithm is given in the flow chart of figure 6.7.5, where \underline{C} is the classification array that defines the final cluster membership; \underline{T} is a temporary classification array in which fusions are effected so that the 'density level' parameter A_{i-1} for fusion sUt is not modified; \underline{H} is a vector of triples containing the hierarchy data from which the dendrogram is obtained.

If, in addition to \underline{H} and \underline{C} , the \underline{A} values are also retained, then it is possible to edit from the final clusters their misfit members. Suppose that a cluster comprises objects (c_1, \dots, c_n) which are associated with ordered A_i values (a_1, \dots, a_n) , that is, $a_i \leq a_{i+1}$, then the user will be able to edit members (c_{i+1}, \dots, c_n) if he considers there to be a sufficiently large discontinuity in the density estimates: $a_{i+1} \gg a_i$. Although this manipulation of cluster membership is purely subjective, these ideas may yield promising avenues of investigation and improvement of the probabilistic model. It is already the case that both of the levels of classification provided with hierarchical mode analysis (nuclei and complete) have their uses (Pocock and Wishart, 1969; Kelly, 1969; Dawson, 1970).

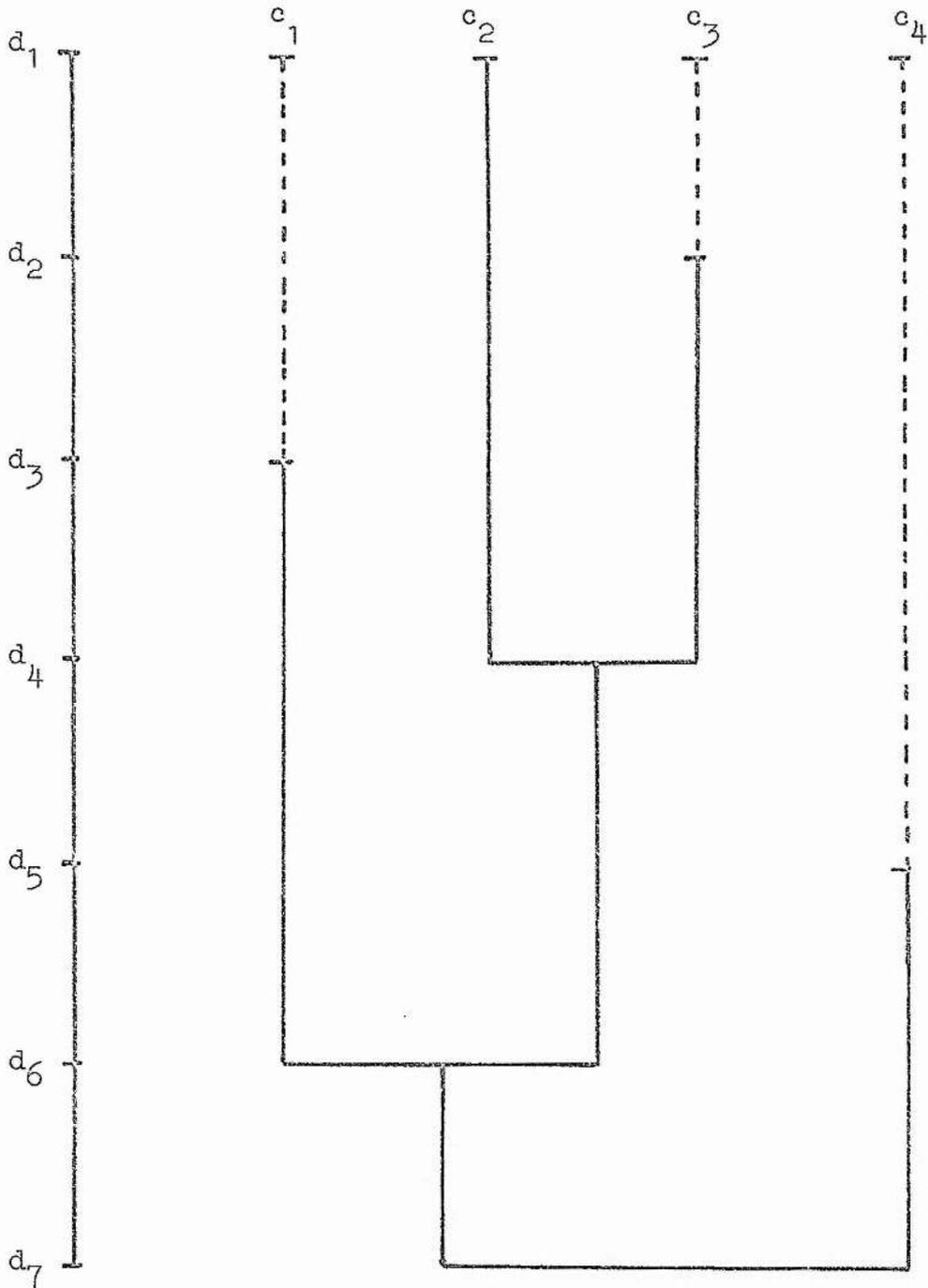


Figure 6.7.4. Dendrogram for the probabilistic methods using the data of figure 6.7.3. Solid lines correspond to hierarchical mode analysis; dotted lines denote the extensions of the dendrogram for the improved method. Hence all four clusters are recognised at levels d_1 to d_4 with the improved algorithm.

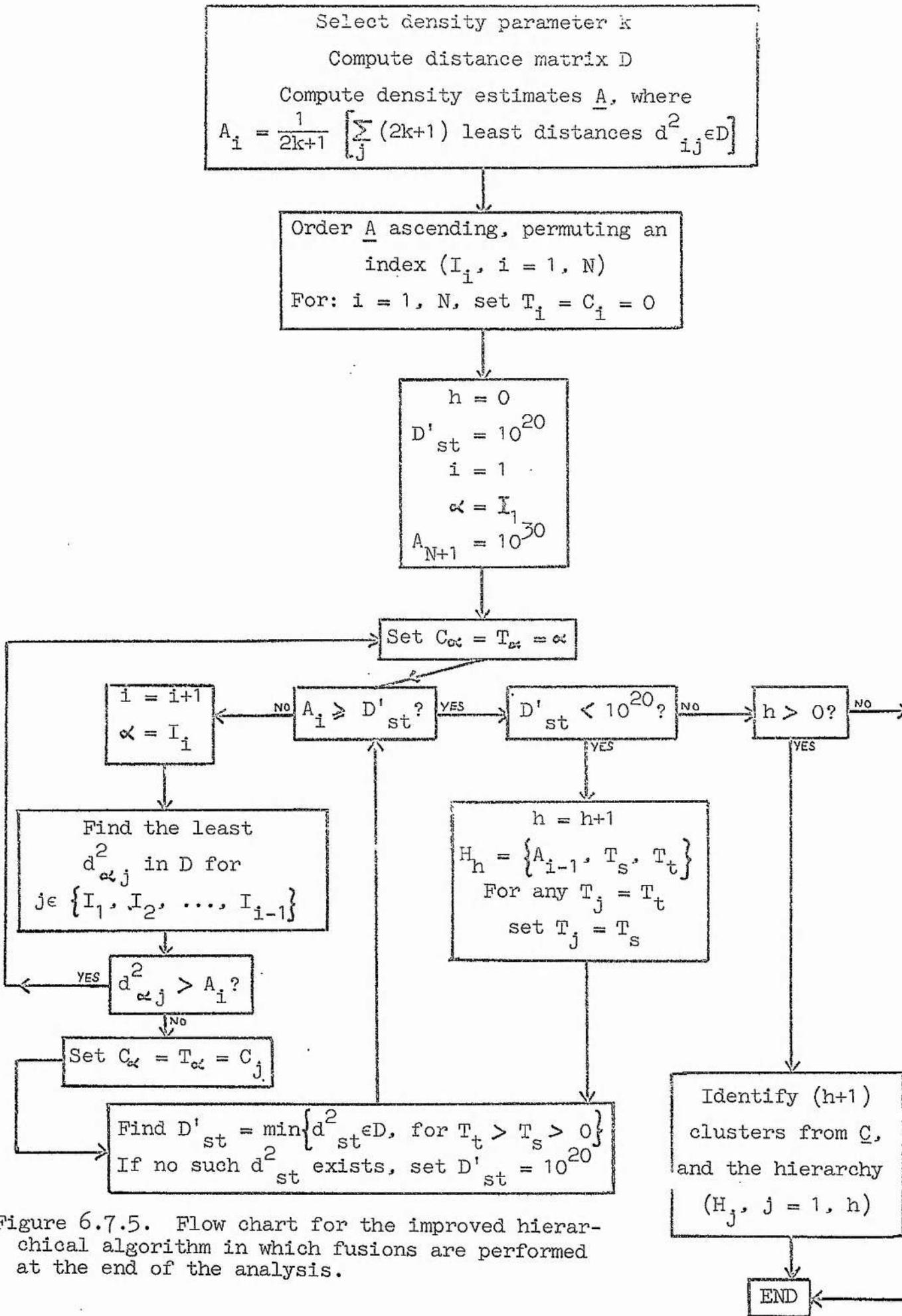


Figure 6.7.5. Flow chart for the improved hierarchical algorithm in which fusions are performed at the end of the analysis.

CHAPTER 7: EXPERIMENTAL TESTS

The purpose of this chapter is to report and interpret some experimental results obtained from testing the generalised iterative relocation procedure (Sect. 4.3 and Sect. 5.2) and hierarchical Mode analysis (Chapter 6). The general objectives are threefold: firstly, twelve different similarity criteria (Sect. 1.4 and Sect. 5.5) are compared using the flexible iterative relocation technique to detect unusual results and isolate ill-conditioned similarity measures. Secondly, results of iterative relocation are obtained using both small and large populations with different starting conditions in order to determine the consistency of the method and the reliability of the similarity measures. Lastly, hierarchical Mode analysis is compared with iterative relocation using all of the test populations, including an instance of parallel elongated clusters, in an attempt to throw more light on the need for resolving natural classes.

In order to construct an adequate test population it must be assumed that a spherical multivariate normal distribution comprises a single indivisible entity which should be resolved, at least at some stage of the analysis, as a single cluster. This assumption may be a little restrictive, but it is difficult to think of an occasion when a well-defined spherical swarm which possesses no major axis of variation should be either subdivided

or grouped with some other distinct swarm. We can obtain a multivariate normal distribution by applying the central limit theorem to a sample generated from a rectangular distribution using a standard random number routine and the following method.

Random normal number generator

The rectangular distribution $f(x) = 1/w$ for $0 \leq x \leq w$ has mean μ and variance σ^2 which are obtained as follows:

$$\mu = E(x) = \int_0^w \frac{1}{w} x dx = \frac{1}{2} w$$

$$\mu'_2 = E(x^2) = \int_0^w \frac{1}{w} x^2 dx = \frac{1}{3} w^2$$

$$\sigma^2 = E((x-\mu)^2) = \mu'_2 - \mu^2 = \frac{1}{3} w^2 - \left(\frac{w}{2}\right)^2 = w^2/12$$

Using the central limit theorem, the sample (x_1, \dots, x_n) taken from $f(x)$ has mean $\bar{x} = \sum x_i/n$ which approaches the normal distribution $N(\mu, \sigma^2/n)$. It follows that the distribution function for the random variable

$$(\bar{x} - \mu) \sqrt{n}/\sigma = \left(\bar{x} - \frac{w}{2}\right) \sqrt{12n}/w$$

approaches the standardised normal distribution function $N(0,1)$, and therefore

$$\mu_c + \left(\bar{x} - \frac{w}{2}\right) \sigma_c \sqrt{12n}/w$$

is approximately distributed as $N(\mu_c, \sigma_c^2)$. By choosing $w = 1$ and $n = 12$, this simplifies to

$$x' = \mu_c + \left(\sum x_i - 6\right) \sigma_c$$

where $\sum x_i$ denotes the sum of twelve random values from the rectangular distribution $f(x) = 1$ for $0 \leq x \leq 1$. Since most modern computers are equipped with fast routines for generating random values from $f(x)$, the normally distributed variate x' is easy to obtain.

For our purposes, descriptive solutions are required and we shall therefore restrict the distribution to two dimensions so that it can be plotted on a scatter diagram. The above formula is used to generate coordinate pairs (x'_1, x'_2) for points having distribution mean (μ_{c1}, μ_{c2}) and joint variance $(\sigma_{c1}^2 + \sigma_{c2}^2)$, where μ_{c1} , μ_{c2} , σ_{c1} and σ_{c2} are chosen parameters.

Iterative Relocation Method

The method chosen for the comparison of the similarity criteria is that defined in Section 4.3, using relocation test (4.1.1), and test (4.1.3) in the case of the comparison of a single individual x with itself (a 1-element cluster). It should be mentioned that some similarity measures satisfy (4.1.3) so that a cluster may occasionally be eliminated through the relocation of all its members. Since normally distributed data are necessarily continuous, the binary information statistic will not be included in the tests.

7.1 ITERATIVE RELOCATION TESTS

Figure 7.1.1 shows a population of 100 points comprising four bivariate normal distributions having means $(\overset{+}{-}3, \overset{+}{-}3)$ and unit

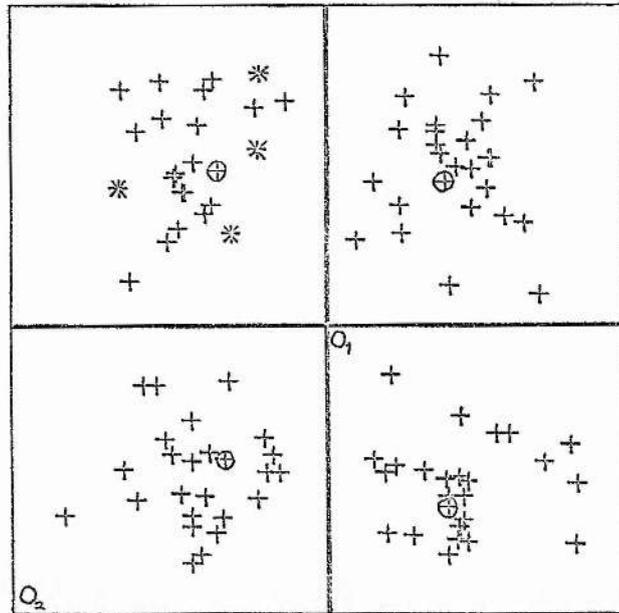


Figure 7.1.1. A 100 point 4-cluster population generated from 4 bivariate normal distributions. Two origins of coordinates were used, these being sited at O_1 and O_2 respectively. The starting classifications chosen for iterative relocation are indicated by:

START 1 * (four bad random points);

START 2 ⊕ (four good points);

START 3 (the four optimum clusters, partitioned by the coordinate axes through O_1).

variances (for both variables within each group); since the variances on both axes were equal, the data were effectively standardised. Two origins of coordinates (O_1 and O_2 in figure 7.1.1) were chosen to demonstrate those similarity measures which are origin dependent. In both cases, three initial classifications were used to start the iterative relocation procedure. These were

four clusters, described as follows:

START 1: 4 bad points, shown by * in figure 7.1.1, which were selected from the same bivariate normal distribution.

START 2: 4 good points, shown by @ in figure 7.1.1, each selected from a different distribution. This can be described as a part-optimum initial solution.

START 3: The four optimum clusters, partitioned by the coordinate axes through O_1 in figure 7.1.1. The intention of supplying the expected final result as starting solution is to expose unstable or badly defined similarity criteria.

Each of the twelve similarity criteria shown in Table 7.1.1 were submitted to iterative relocation using these six combinations of origin and starting solution; naturally, some final classifications were duplicated, and of the total of 72 tests 17 unique results were obtained. Twelve of these are shown in figure 7.1.2 using partition lines to demark cluster boundaries; four of the other five results were sufficiently random to preclude the drawing of partition lines, and the fifth comprised one cluster being the entire population. The 72 results are identified in Table 7.1.1, which also shows the number of iterations required before stability was reached. In two tests the maximum of 15 iterations was completed, so that the procedure terminated without reaching stability.

The conclusions that can be drawn from these tests are, to a

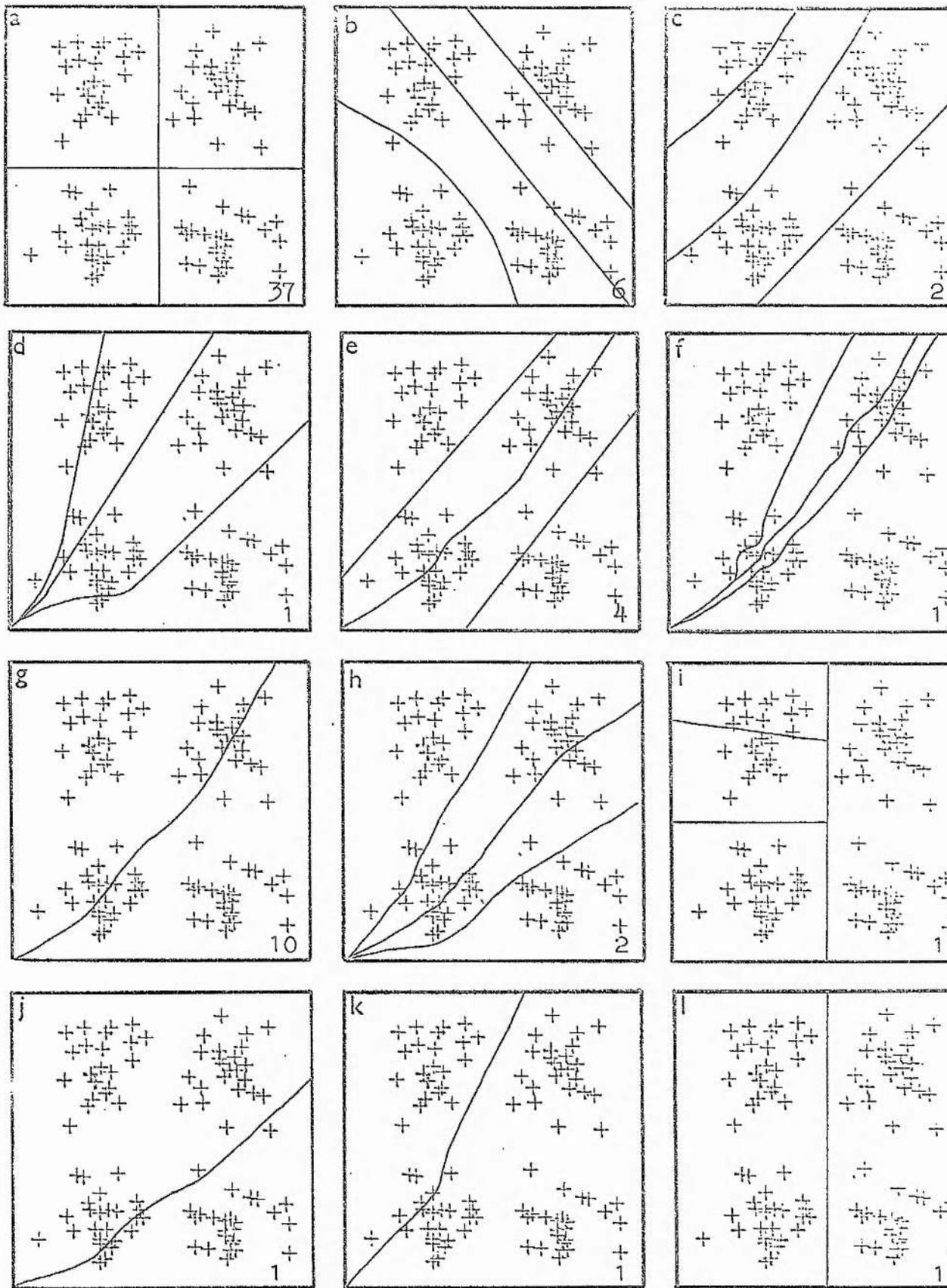


Figure 7.1.2. 12 different partitions of the data of figure 7.1.1 obtained during 6 tests to optimise 12 similarity criteria by iterative relocation. The number against each square is the frequency of the result in Table 7.1.1.

| SIMILARITY CRITERION | 0 ₁ | | | 0 ₂ | | | 0 ₁ | | | 0 ₂ | | |
|----------------------|----------------|---|---|----------------|---|---|----------------|---|---|----------------|---|---|
| | START | | | START | | | START | | | START | | |
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Distance | a | a | a | a | a | a | 6 | 1 | 0 | 6 | 1 | 0 |
| Average Distance | a | a | a | a | a | a | 4 | 1 | 0 | 4 | 1 | 0 |
| Similarity Ratio | a | a | a | i | a | a | 4 | 1 | 0 | 3 | 1 | 0 |
| Error Sum of Squares | a | a | a | a | a | a | 5 | 2 | 0 | 5 | 2 | 0 |
| Variance | a | a | a | a | a | a | 11 | 6 | 0 | 12 | 6 | 0 |
| Cosine | a | a | a | d | h | h | 4 | 1 | 0 | 3 | 2 | 2 |
| Nonmetric | R | R | R | a | a | a | U | 8 | U | 6 | 1 | 0 |
| Size Difference | b | b | b | b | b | b | 2 | 3 | 2 | 2 | 3 | 2 |
| Shape Difference | c | e | e | c | e | e | 4 | 3 | 3 | 4 | 3 | 3 |
| Dispersion | g | g | g | g | g | g | 2 | 1 | 1 | 2 | 1 | 1 |
| Correlation | g | g | R | g | g | f | 4 | 4 | 4 | 2 | 2 | 5 |
| Dot Product | l | a | a | - | k | j | 3 | 1 | 0 | 2 | 3 | 4 |

Table 7.1.1. Summary of the results obtained using iterative relocation to optimise twelve different similarity criteria with the data of figure 7.1.1. The two columns of letters correspond to the twelve final partitions shown in figure 7.1.2, and on the right are the numbers of iterations required for convergence in each test. The data were clustered using three different starting solutions with reference to two origins of coordinates 0₁ and 0₂ (figure 7.1.1), making a total of six tests against each criterion. R denotes a grouping which was sufficiently random to preclude the drawing of partition lines; - denotes the dot product test in which only one final cluster was obtained; U specifies the two tests for which the nonmetric coefficient failed to converge within 15 iterations.

certain extent, self-evident; they may be summarized as follows:

- 1) Distance, Average Distance, Similarity Ratio, Error Sum of Squares and Variance all perform satisfactorily, although Variance exhibits a certain instability.
- 2) Cosine and Nonmetric are origin dependent. It is worth noting that the Nonmetric coefficient is often used with unstandardised all-positive scores (e.g. binary 1/0 data, or species frequencies in stands) - the satisfactory performance of this coefficient using origin O_2 therefore accounts for its successful usage (Lance and Williams, 1966b).
- 3) Dispersion, Correlation and Dot Product are very unsatisfactory, and their further use is not recommended.
- 4) Size Difference and Shape Difference produce interesting results, although their value can probably be questioned. Both coefficients are origin independent, and the resulting elongated clusters could be regarded as symptomatic of the need to eliminate such internal factors of variation as shape and size, respectively.
- 5) As expected, the part-optimum starting solution (START 2) yields faster convergence than the random initial classification (START 1), especially with the first five coefficients in Table 7.1.1.

Large Populations

If the good results of the first trials are to be reliable,

then they must also be duplicated with large populations. Figure 7.1.3 shows a population of 800 points generated from the same model as figure 7.1.1; thus each cluster increases in size by a factor of 8. Figure 7.1.3 also shows the four points which constitute the random starting solution (START 1) with these data. The part-optimum starting classification (START 2) was replaced in this case with the worst population-partition which could be devised (shown in figure 7.1.4); every fourth point was allocated to the same cluster, and since the distributions were generated in blocks of 200 points numbered sequentially, each starting cluster contained $\frac{1}{4}$ of each final cluster. Table 7.1.2 shows the results of iterative relocation using the three starting solutions with the origin at the intersection of the coordinate axes.

These tests confirm that Variance is unreliable (failing to converge within 15 iterations, excepting with the optimum result as starting solution). Distance, Average Distance, Similarity Ratio and the Error Sum of Squares all perform satisfactorily, although the Error Sum of Squares finds an unstable local optimum result (i) with the random starting solution. A rather unexpected finding is that the 'worst possible' population-partition of figure 7.1.4 yields a much faster convergence than the four random points (figure 7.1.3). It seems, therefore, that in the absence of a suitable part-optimum starting solution, a random population-partition is probably a better initial classification than k random points.

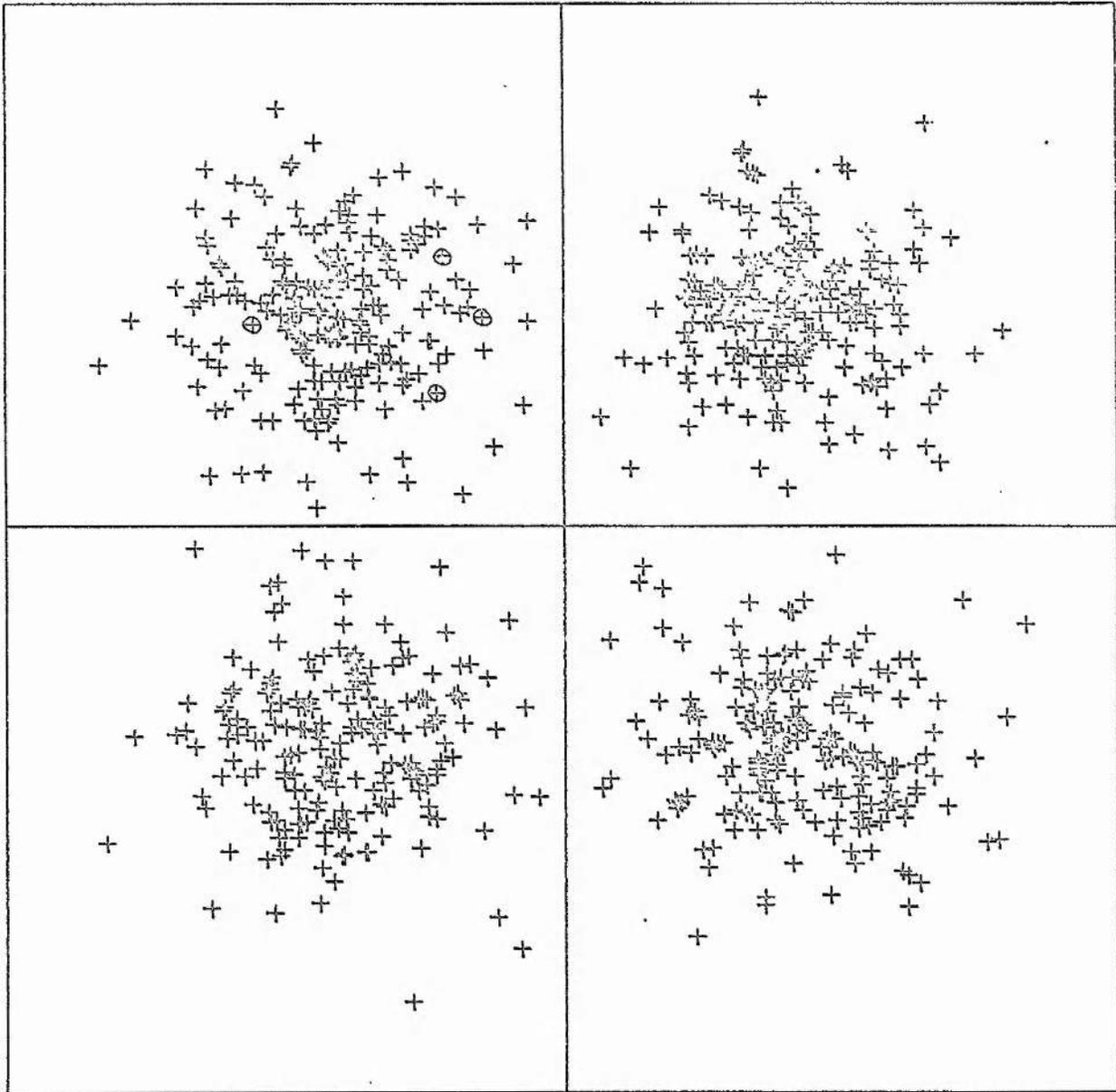


Figure 7.1.3. An 800 4-cluster distribution generated from the same distribution as figure 7.1.1. ⊕ denotes the 4 bad random points used as starting solution with iterative relocation to optimise the 5 well-conditioned similarity criteria listed in Table 7.1.2. The origin of coordinates is at the intersection of the axes, which serve to partition the population into the optimum 4-cluster solution.

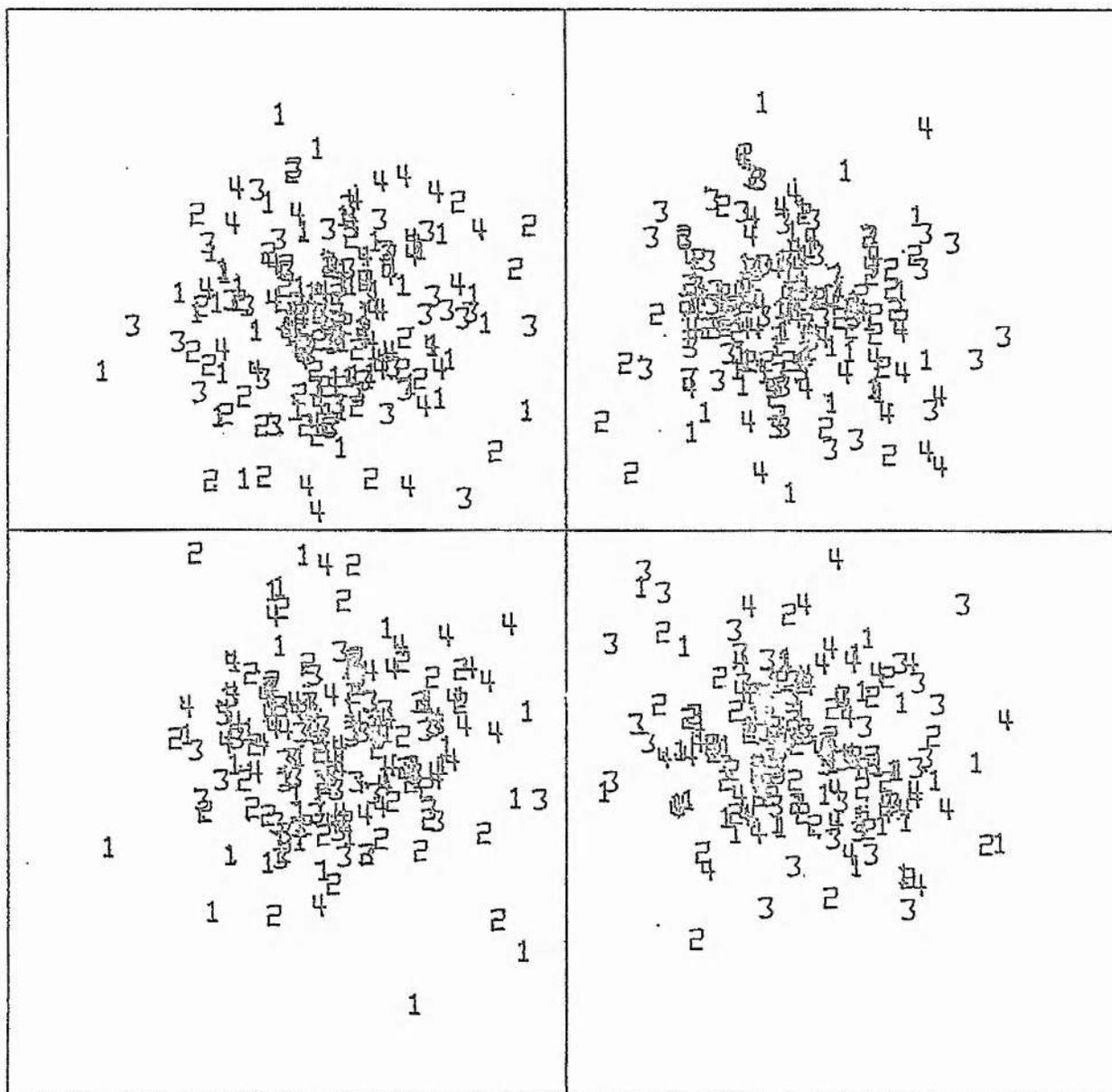


Figure 7.1.4. A bad random population-partition of the data of figure 7.1.3 into four clusters. Each digit is the code of the cluster which has been allocated the corresponding point of figure 7.1.3.

| <u>SIMILARITY CRITERION</u> | <u>STARTING CLASSIFICATION</u> | | |
|-----------------------------|--------------------------------|------------------------------|------------------------------|
| | 4 POINTS (7.1.3) | 4 BAD CLUSTERS (7.1.4) | OPTIMUM RESULT (7.1.3) |
| Distance | a 7 | a 2 | a 0 |
| Average Distance | a 5 | a 2 | a 0 |
| Similarity Ratio | a 5 | a 2 | a 0 |
| Error Sum of Squares | i 4 | a 2 | a 0 |
| Variance | a U | a U | a 1 |

Table 7.1.2. Summary of the results for iterative relocation using the data of figure 7.1.3 to optimise 5 similarity criteria for 4 clusters. Each similarity measure was tested using three starting conditions: 4 BAD RANDOM POINTS (shown in figure 7.1.3); 4 BAD CLUSTERS (shown in figure 7.1.4); and the OPTIMUM RESULT (indicated by the partition of the coordinate axes in figure 7.1.3). The type (a or i) of result is illustrated in figure 7.1.2 - note the single convergence of the error sum of squares to a sub-optimum solution (type i). The figures are the numbers of iterations required for convergence (U denotes no convergence after 15 iterations).

Consistency of Results

One criticism of the tests so far used is that the populations contain four very distinct clusters, a situation seldom found in 'real' data. It is quite possible that if these clusters had been overlapping to a greater extent, then the results for the four successful similarity measures would not have been so good. To check the consistency of the similarity measures in finding an obscure classification, the iterative relocation procedure was used to partition a unimodal bivariate normal distribution, shown in

figure 7.1.5. The starting solution of two points (ringed in figure 7.1.5) generated the final partition shown by the partition line of figure 7.1.5, for which the error sum of squares had the value 140.83. Figure 7.1.6 shows a random population-partition used as starting solution, together with its final 2-cluster partition; in the latter case, the value of the error sum of squares was 137.18 - a slight improvement on the previous result. These two very different results were duplicated in each test of all four similarity criteria, shown below. The difference in the partitions is very distinct, and it is probable that other unique results could have been obtained by carefully choosing pairs of points to act as initial cluster centres. The number of iterations required for convergence of the starting solution of figure 7.1.5 (two random points), and the population-partition (figure 7.1.6) with each test were as follows:

| | <u>7.1.5 (2 points)</u> | <u>7.1.6 (population-partition)</u> |
|----------------------|-------------------------|-------------------------------------|
| Distance | 3 | 7 |
| Similarity Ratio | 3 | 3 |
| Average Distance | 4 | 8 |
| Error Sum of Squares | 3 | 3 |

It is noticeable that the population-partition takes longer to converge than the 2 random points. This finding is exactly the reverse of the 4 cluster test (previous paragraph), but since the population-partition appears to yield a slightly better final classification the

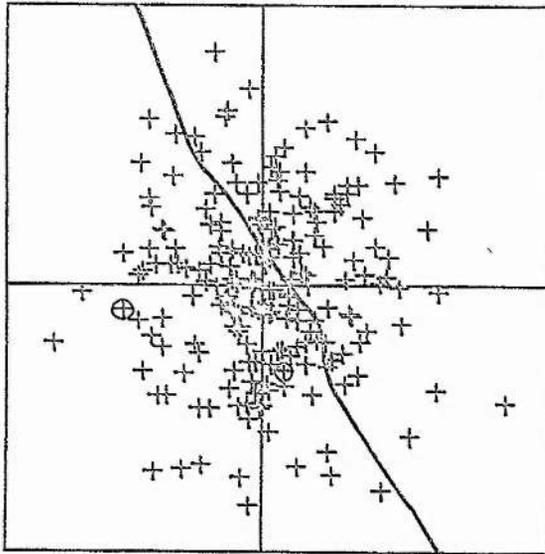


Figure 7.1.5. A standardised unimodal bivariate normal distribution showing two points (indicated by \oplus) which were used as initial centres for iterative relocation to optimise four similarity criteria (the first four in Table 7.1.1) at the 2 cluster level. The partitioning line indicates the final classification obtained with all four measures, and the error sum of squares for this grouping was 140.83.

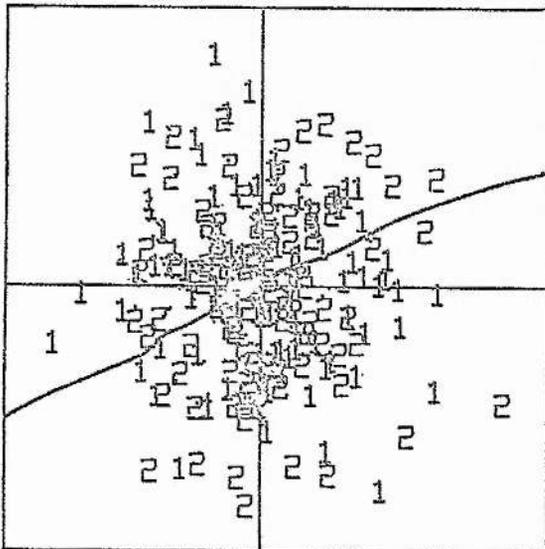


Figure 7.1.6. The random population-partition of the distribution in figure 7.1.5 as starting classification with iterative relocation at the 2 cluster level. The same final partition was obtained with all four tested similarity measures and the error sum of squares for this grouping was 137.18, suggesting that it improved the previous result shown in figure 7.1.5.

conclusion and preference still hold. The tests also point to a certain lack of consistency in the iterative relocation method under different initial classifications.

Elongated Clusters

One general demand that can reasonably be made of a classification method is that if there exists a data set having, say, 4 well-defined clusters (such as those in figure 7.1.3) which the method finds, and if we remove two of these clusters then the method should also successfully find the two remaining clusters. To test the iterative relocation procedure in this way, two 100-point bivariate normal distributions were generated with means $(\pm 3, 0)$ and unit variances; that is, from the same model as that used to generate the two upper clusters in figure 7.1.3. After standardisation, the clusters are seen to be elongated parallel to the y-axis (figure 7.1.7). At this stage it was thought sufficient to test the distance criterion with iterative relocation (i.e. the k-mean system), since the other three satisfactory similarity measures have previously behaved almost identically. Ten 2-cluster starting solutions were devised, as follows: five variants of two initial bad points were obtained by selecting two random individuals from the same elongated cluster; four variants of two good points were defined by two individuals, one from each of the elongated clusters; the tenth starting solution was the random population-partition derived by allocating every other point in both of the

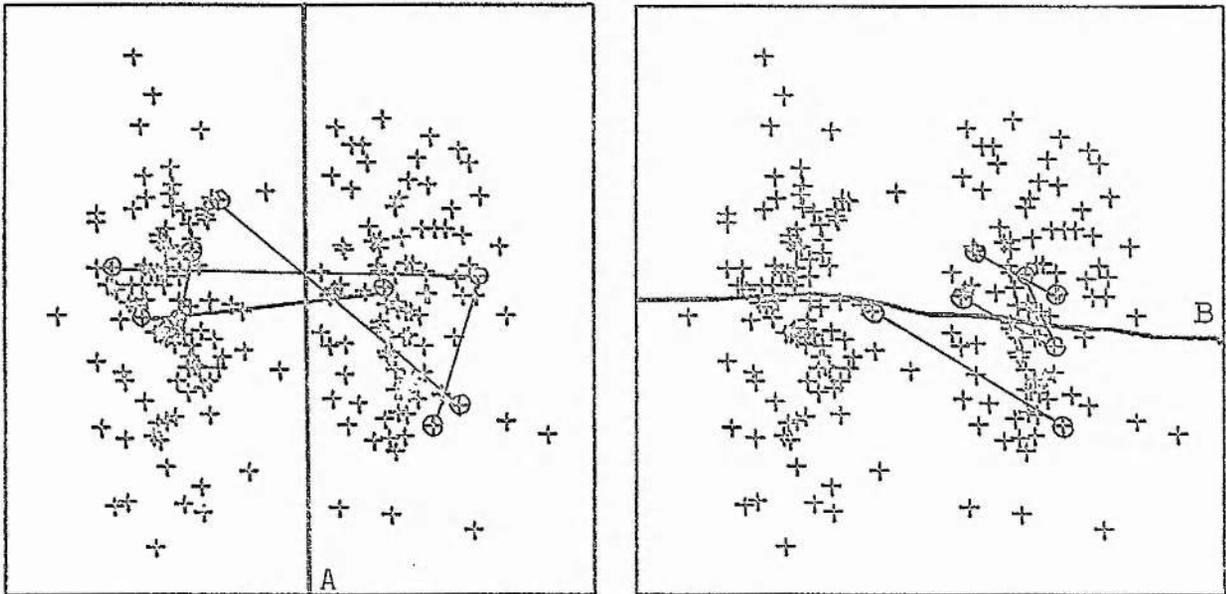


Figure 7.1.7. An example of two parallel clusters elongated through standardisation, and the two final partitions obtained using iterative relocation to optimise distance (the k-mean system). Lines join pairs of points chosen as starting solutions for which the two stable partitions A and B were derived. The error sum of squares in each case was: A(113.9) and B(138.4).

elongated clusters to the same starting group.

Every test converged to one of two stable solutions, shown by the partition lines of figures 7.1.7A and 7.1.7B, respectively. The random population-partition produced the preferred result (figure 7.1.7A), and the final solutions are shown with their corresponding starting solutions indicated on each diagram of figure 7.1.7 by joining the two initial cluster centres with lines.

These results suggest that the iterative relocation procedure does not reliably find elongated clusters, and they support the previous criticisms (Sect. 6.1) that 'minimum-variance' techniques are inclined to force spherical clusters and may derive partitions which cut across dense swarms of points (viz, figure 7.1.7B). Once again, the random population-partition appears to be preferable to the choice of random points or individuals for the initial classification.

7.2 HIERARCHICAL MODE ANALYSIS

When the distribution of figure 7.1.7 was subjected to hierarchical Mode analysis it was found that the preferred partition (A) was obtained during each of 8 trials. However, this encouraging result was slightly offset by the generation of other lower level classifications. Using the average distance as density estimate (Sect. 6.6), values of k (the density parameter) from 3 to 10 were tried, and it was found that the method recognised six basic clusters (shown in figure 7.2.1). Generally speaking, the higher the value of

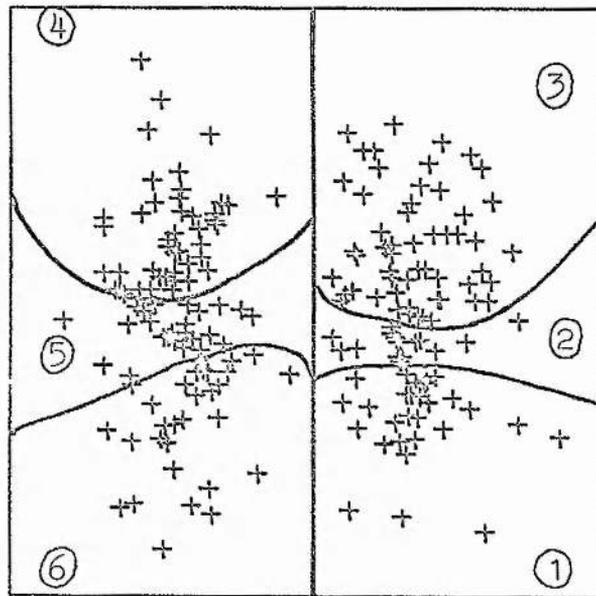


Figure 7.2.1. The six-cluster partition of the data of figure 7.1.7 obtained using hierarchical Mode analysis. The optimum 2-cluster solution (1+2+3) and (4+5+6) was found eventually during all of the tests, and different combinations of these six sectional groups were obtained by varying the density parameter k (see also figure 7.2.2).

k the fewer were the number of clusters (indeed, for $k > 8$ only the 3 and 2 cluster levels were obtained). All of the unique results are shown by the dendrograms of figure 7.2.2, where the cluster codes (1-6) correspond to the partitions in figure 7.2.1. It is noticeable that a very large jump occurs prior to the 2-cluster grouping in each case, indicating stable separation of the two elongated clusters.

The question now arises: should hierarchical Mode analysis only recognise the two elongated clusters, or is it reasonable to

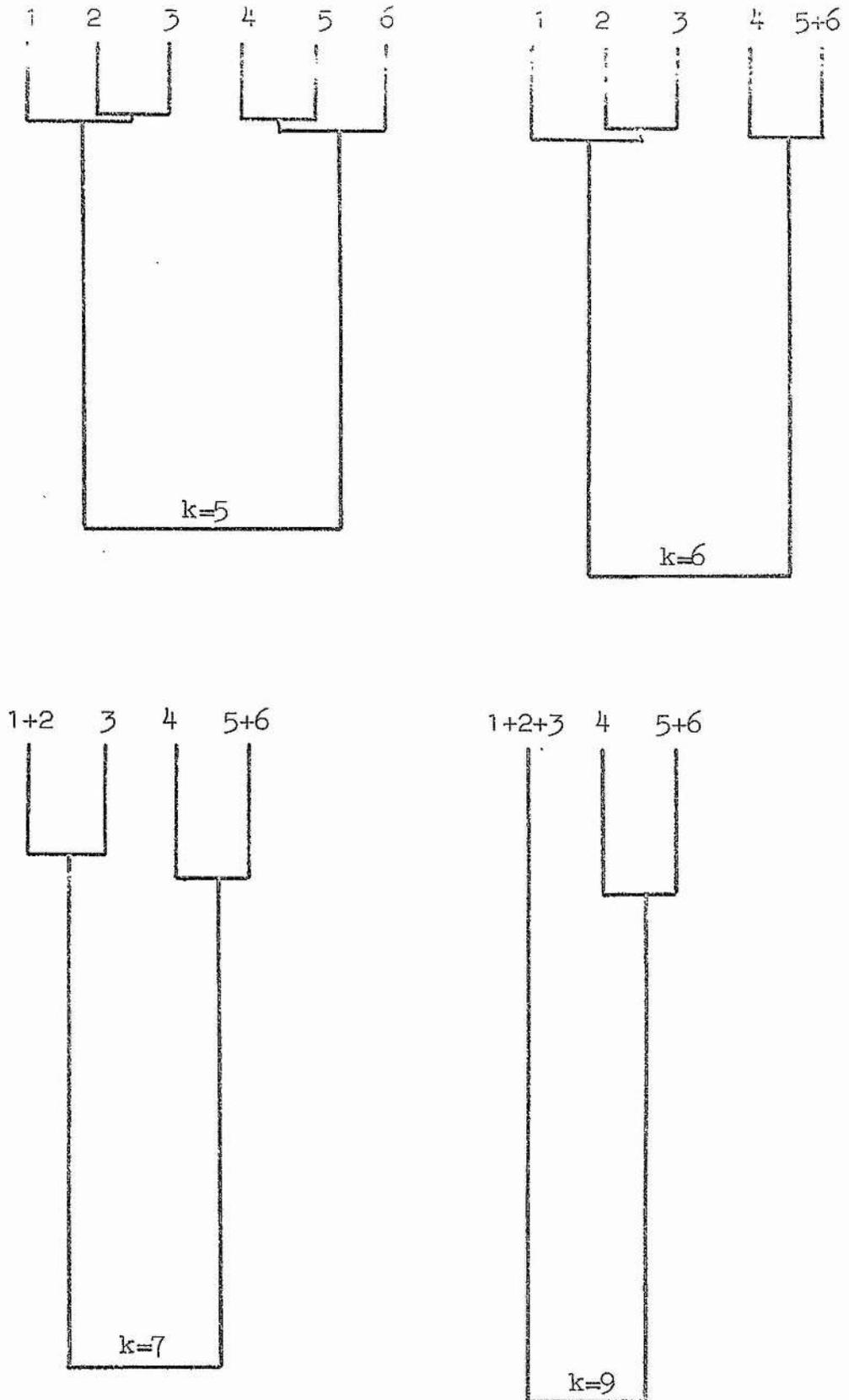


Figure 7.2.2. Four dendrograms showing the various classifications derived by hierarchical Mode analysis with the data of figure 7.2.1; the cluster codes 1-6 are identified by partitions on figure 7.2.1. Note the large increase in the density threshold prior to the 2-cluster grouping in each case.

further subdivide a bivariate normal distribution into compact sections (as shown in figure 7.2.1)? It may be that the example contains too few points to satisfactorily estimate the underlying density function, in which case the method is unlikely to apply its theoretical design in practice since one would expect 200 points to be a sufficiently large population for two dimensions. Alternatively, a confidence test based on the number of points used to determine modes at the 6-3 cluster levels could be devised, and this might then render 'not significant' all results excepting the 2-cluster level. Such a test could be based on the density threshold values for these classifications, shown on the dendrograms (figure 7.2.2).

One other possibility is that the method will always recognise small concentrations of points in a discrete population, and is therefore very likely to find small spherical clusters during the early stages of analysis. This would suggest that the density criterion (average of the $2k+1$ least distances for each point) is inadequate, and might be improved by a smoothing technique, or something similar.

Minimum Cluster Size Criterion f

To test the hypothesis that compact spherical clusters are generated by Mode analysis around small unusually dense centres of a discrete distribution, a size criterion f was introduced to the program. At fusion cycles, output of classifications is restricted to those groupings obtained prior to the fusion of two or more

clusters which each comprise more than f 'dense points'.

All combinations of k and f in the range 1 to 10 were then tried using the unimodal population of figure 7.1.5 and the 4-cluster distribution of figure 7.1.1. It was found that a line on the 10 x 10 grid of k against f could be drawn such that all combinations of k and f to the left of the line failed, while combinations of k and f on the line and to the right of the line succeeded. By 'succeeded' we mean that the program produced no classifications of the unimodal population, and a maximum of 4 groups (partition a of figure 7.1.2) in the 4-cluster case. These findings are shown in figure 7.2.3, and it is evident that successful combinations of k and f are related to population size: it would be nice to repeat the tests with different sizes of population, but this would require a considerable amount of computing time.

It is clear from figure 7.2.3 that the original hypothesis is justified. The method does tend to find more than the expected number of clusters at the start of the analysis, when localised dense regions defined by a very few points are being merged. This effect is controlled to a certain extent by increasing the value of the density threshold k , which itself acts as a cluster size criterion. That is, a small isolated group of 4 points (say) might not be recognised when $k > 3$, since the density estimate for each point will be based on the distance from the point to its k th nearest neighbour (which is outside the group of 4). The value of k there-

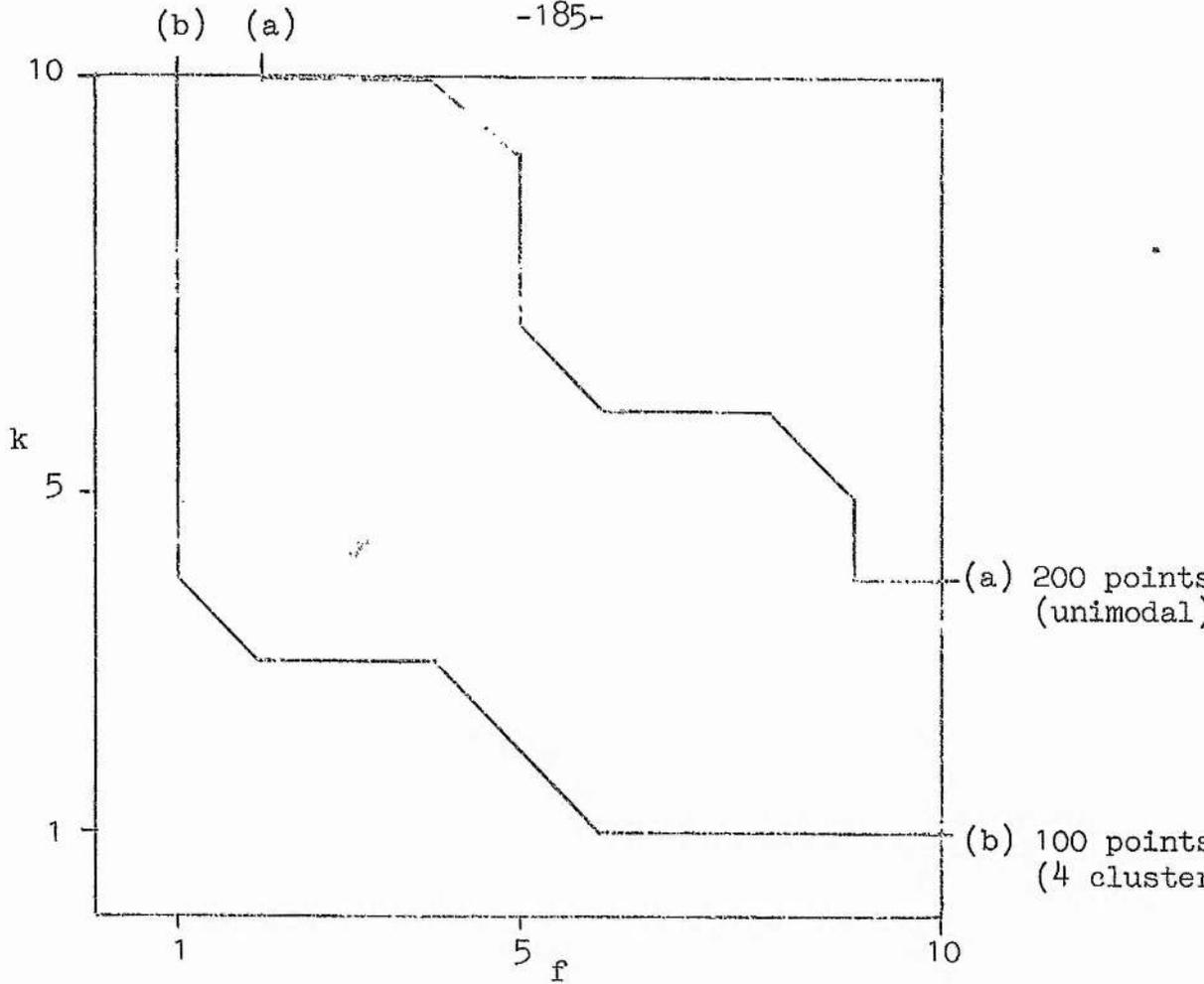


Figure 7.2.3. Combinations of parameter k with the cluster size criterion f (see text) are shown which successfully recognise: (a) the unimodal distribution of figure 7.1.5 (200 points), and (b) the 4-cluster distribution of figure 7.1.1 (100 points). Combinations of k and f below the lines (a) and (b) found more than the expected number of clusters (1 and 4 respectively). The inference is that successful values of k are related to population size.

fore cannot be increased without limit, for that would preclude the recognition of small clusters.

7.3 CONCLUSIONS

The results of the chapter show that the different similarity criteria behave in very different ways. Some of the angular measure-

ments derive elongated clusters, suggesting that some workers (e.g. Boyce, 1969) have found it necessary to depart from the more standard minimum-variance approach. Such application can be taken as symptomatic of the need to recognise elongated clusters, and confirms the previous reasoning (Sect. 6.1) that good spherical clusters will often be stretched by the introduction of irrelevant variables, the presence of internal factors of variation such as size and shape, or the use of transformations such as standardisation.

If tight clusters are required, then the iterative relocation method seems to be fairly reliable when used with the distance, average distance, error sum of squares and similarity ratio criteria. However, the results often vary with different starting classifications, particularly when the optimum solution is not well-defined. There is also no guarantee that the method will not cut across a dense swarm of points, as predicted in Section 6.1. It seems that the best starting solution is probably a part-optimum result (perhaps that obtained with hierarchic fusion); failing that, a random partition of the population to form k groups is preferable to k random points or individuals. Where possible, different starting solutions should be used to test the stability of the final result.

Hierarchical Mode analysis has been shown to consistently find elongated clusters, but the method also tends to split such

clusters into compact sections, particularly at the start of the analysis. That the two elongated clusters in figure 7.1.7 were always eventually separated by the method is sufficiently encouraging to stimulate further investigation of the properties of 'natural classes'. Perhaps the most rewarding area of study is the development of significance tests based on the numbers of points used to define cluster kernels. It seems likely that a size threshold could be defined which would enable the early classifications of Mode analysis to be deemed 'not significant'. Alternatively, some type of smoothing technique could be applied to the density estimates for each point, and this might reduce the effect of small density saddles so that localised density modes would tend to disappear.

CHAPTER 8: COMBINATORIAL COEFFICIENTS

8.1 LANCE-WILLIAMS' 4-PARAMETER MODEL

Some of the hierarchical fusion methods can be evaluated solely from transformations of the similarity matrix. Suppose that the triangular matrix $D = (d_{ij}^2; i = 1, j - 1; j = 2, N)$ of distances is computed, then it has been shown (Lance and Williams, 1966a, 1967a) that the fusion sequences for five hierarchical methods are obtained by replacing distances as follows:

- 1) At any stage of the fusion sequence, find the ¹least 'distance' d_{pq}^2 in D for 'active' clusters p and q ($p < q$).
- 2) Denote by r the new cluster ($p + q$), and compute new 'distances' d_{ir}^2 between all clusters i ($i \neq p, q$) and cluster r , using the general formula:

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma |d_{ip}^2 - d_{iq}^2| \quad (8.1.1)$$

(the parameters α_p , α_q , β and γ are given below).

- 3) Render cluster q inactive, and replace cluster p with the new cluster ($p + q$) by substituting d_{ir}^2 for the elements d_{ip}^2 of the p th row and column of D .
 - 4) Return to (1) and find the next fusion in the sequence.
- The method ends when all ($N - 1$) fusions have been completed.

¹If D comprises similarity coefficients, then the greatest element 'should' be found.

The five methods examined by Lance and Williams are obtained for the values of the parameters α_p , α_q , β and γ given below, where k_p , k_q and k_i are cluster sizes, and $k_r = k_p + k_q$.

Single linkage: $\alpha_p = \alpha_q = \frac{1}{2}$; $\beta = 0$; $\gamma = -\frac{1}{2}$

These values hold for all dissimilarity (i,j) coefficients. However, Lance and Williams do not consider the case of similarity coefficients, for which the sign of γ must be reversed. That is, if D contains similarity coefficients then the parameters should be $\alpha_p = \alpha_q = \frac{1}{2}$; $\beta = 0$; $\gamma = \frac{1}{2}$; and step 1 of the algorithm should find the greatest 'similarity' in D.

Complete linkage: $\alpha_p = \alpha_q = \frac{1}{2}$; $\beta = 0$; $\gamma = \frac{1}{2}$; for dissimilarities,
or $\alpha_p = \alpha_q = \frac{1}{2}$; $\beta = 0$; $\gamma = -\frac{1}{2}$; for similarities.

Average linkage: $\alpha_p = k_p/k_r$; $\alpha_q = k_q/k_r$; $\beta = \gamma = 0$.

These parameters hold for all coefficients whether of type similarity or dissimilarity. The form of the combinatorial algorithm for average linkage is also given by Ray and Berry (1965) and McQuitty (1966b).

Centroid (using distances): $\alpha_p = k_p/k_r$; $\alpha_q = k_q/k_r$; $\beta = -\alpha_p\alpha_q$;
 $\gamma = 0$; Lance and Williams propose

these parameters for use only with distances. However, it can be shown (Sect. 8.3) that centroid sorting with some other coefficients can be obtained using the combinatorial algorithm. The above parameter values are also given, in a slightly modified form, by Gower (1967) and Proctor (1966).

Median (Gower's method): $\alpha_p = \alpha_q = \frac{1}{2}; \beta = -\frac{1}{4}; \gamma = 0;$

This method, as previously described (Sect. 2.2), can only be given a geometric interpretation if D contains distances: that is, when p and q are fused the new cluster (p + q) is located at the midpoint of the line which joins the points representing clusters p and q. However, Gower (1967) seems to propose the use of the combinatorial parameter values given above for all similarity measures.

In addition to these five methods considered by Lance and Williams, we can add the following two fusion techniques which are defined purely as combinatorial transformations of D:

McQuitty's similarity analysis: $\alpha_p = \alpha_q = \frac{1}{2}; \beta = \gamma = 0;$

McQuitty (1966b) proposes this combinatorial transformation of D in a preamble to his introduction (evidently independently) of average linkage. These parameters are suggested for all (i,j) measures.

Lance-Williams' flexible: $\alpha_p = \alpha_q = \frac{1}{2}(1 - \beta); \gamma = 0; \beta = \text{variable}.$

This method is defined as a family of fusion strategies determined by the variable parameter β , which Lance and Williams suggest should take the value $\beta = -\frac{1}{4}$. It has been shown empirically (Lance and Williams, 1967a) that as β varies from 1 to -1 the method changes from extreme 'space-conserving' to 'space-dilating' (the authors' terms for methods which chain and find tight minimum-variance clusters, respectively). However, I submit that the pro-

vision of yet another control which is to be 'chosen' by the user reduces the objectivity and practicality of the method (it is almost suggested that one can vary β to obtain any desired classification).

8.2 WARD'S METHOD

Ward (1963) proposes that at any stage of the hierarchical fusion analysis, the 'loss of information' which results from the grouping of points into clusters can be measured by the total sum of the squared deviations of every point from the mean of the cluster to which it belongs. At each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined. Initially, each of the individuals is regarded as a single-point cluster, and the first fusion clearly involves those two points which are closest. At subsequent steps, however, the fusion of multi-point clusters must be considered, and the combinatorial transformation for Ward's method (Wishart, 1969c) is obtained as follows:

In general, if X_{ijt} is the value of the j th variable for the i th point of cluster t , containing k_t points, and U_{jt} is the mean of the j th variable for t , then the error sum of squares for t is defined as

$$E_t = \sum_{i=1}^{k_t} \sum_{j=1}^M (X_{ijt} - U_{jt})^2$$

which, when expanded, becomes

$$E_t = \sum_{i=1}^{k_t} \sum_{j=1}^M X_{ijt}^2 - k_t U_t^2, \quad (8.2.1)$$

where $U_t = (U_{1t}, U_{2t}, \dots, U_{Mt})$ is the position vector of the mean for cluster t .

The value of the objective function E is the sum of the error sum of squares for each of the T clusters,

$$E = \sum_{t=1}^T E_t$$

and at the suggested fusion of clusters p and q , the increase in E is given by

$$I_{pq} = E_r - E_p - E_q,$$

where E_r is the error sum of squares for the union set $r = p + q$.

Thus, from (8.2.1),

$$I_{pq} = \sum_{i=1}^{k_r} \sum_{j=1}^M X_{ijr}^2 - k_r U_r^2 - \sum_{i=1}^{k_p} \sum_{j=1}^M X_{ijp}^2 + k_p U_p^2 - \sum_{i=1}^{k_q} \sum_{j=1}^M X_{ijq}^2 + k_q U_q^2.$$

The sums of the squares X_{ijt}^2 cancel, and hence

$$I_{pq} = k_p U_p^2 + k_q U_q^2 - k_r U_r^2. \quad (8.2.2)$$

But

$$\begin{aligned}
 k_{r-r}^2 U^2 &= (k_{p-p} U_p + k_{q-q} U_q)^2 \\
 &= k_{p-p}^2 U_p^2 + k_{q-q}^2 U_q^2 + 2k_p k_q U_p U_q \\
 &= k_{p-p}^2 U_p^2 + k_{q-q}^2 U_q^2 + k_p k_q (U_p^2 + U_q^2 - (U_p - U_q)^2) ,
 \end{aligned}$$

which reduces to

$$U_r^2 = \frac{k_p}{k_r} U_p^2 + \frac{k_q}{k_r} U_q^2 - \frac{k_p k_q}{k_r^2} (U_p - U_q)^2 . \quad (8.2.3)$$

On substitution for U_r^2 , equation (8.2.2) becomes

$$I_{pq} = (k_p k_q / k_r) (U_p - U_q)^2$$

but since $(U_p - U_q)^2 = d_{pq}^2$, the distance between the means of clusters p and q,

$$I_{pq} = (k_p k_q / k_r) d_{pq}^2 \quad (8.2.4)$$

and fusion occurs when I_{pq} is a minimum.

At the fusion $r = p + q$, the suggested fusion of any other cluster i with the new union cluster r will result in an increase in the objective function of

$$I_{ir} = \left[k_i k_r / (k_i + k_r) \right] d_{ir}^2 . \quad (8.2.5)$$

The distance between the means of i and r is given by

$$d_{ir}^2 = (U_i - U_r)^2$$

which reduces, in the same manner as equation (8.2.3) to

$$d_{ir}^2 = \frac{k_p}{k_r} d_{ip}^2 + \frac{k_q}{k_r} d_{iq}^2 - \frac{k_p k_q}{k_r^2} d_{pq}^2 . \quad (8.2.6)$$

On substitution for each d_{ij}^2 in terms of k_i , k_j , and I_{ij} from equation (8.2.5), equation (8.2.6) becomes, after manipulation

$$I_{ir} = \frac{1}{(k_r + k_i)} \left[(k_i + k_p)I_{ip} + (k_i + k_q)I_{iq} - k_r I_{pq} \right] . \quad (8.2.7)$$

If the triangular matrix of all inter-point squared Euclidean distances $D = (d_{ij}^2; i = 1, j - 1; j = 2, N)$ is calculated and stored, then the increase in the objective function which results from the fusion of any two single-element clusters, p, q is, from (8.2.5)

$$I_{pq} = \frac{1}{2} d_{pq}^2 .$$

The first fusion therefore concerns those two points p and q for which d_{pq}^2 is a minimum. If, at the union $p = p + q$, the elements

$(d_{ip}^2; i = 1, N; i \neq p, q)$, of the matrix D , are replaced by

$$d_{ip}^2 = 2I_{ip} \quad (8.2.8)$$

then these new values are consistent with the original distances in D ; that is, equation (8.2.8) holds for all $(d_{ij}^2; i, j \neq q)$.

By replacing the cluster p with the union $p + q$, cluster q becomes inactive, and therefore the elements of the q th column and row of D are redundant.

Equation (8.2.8) becomes, after substitution for I_{ip} from equation (8.2.7),

$$\begin{aligned}
 d_{ip}^2 &= \frac{2}{(k_i + k_r)} \left[(k_i + k_p)I_{ip} + (k_i + k_q)I_{iq} - k_i I_{pq} \right] \\
 &= \frac{1}{(k_i + k_r)} \left[(k_i + k_p)d_{ip}^2 + (k_i + k_q)d_{iq}^2 - k_i d_{pq}^2 \right]
 \end{aligned}
 \tag{8.2.9}$$

and if, at every fusion step, the elements of the pth column and row of D are modified by (8.2.9), then equation (8.2.8) will hold for all 'distances' d_{ij}^2 , for active sets i,j. The term 'distance' d_{ij}^2 no longer applies in the Euclidean sense, but may be thought of as 'objective distance' or $2I_{ij}$. It follows that Ward's method is evaluated by the combinatorial algorithm for parameters:

$$\alpha_p = \frac{k_i + k_p}{k_i + k_r} ; \alpha_q = \frac{k_i + k_q}{k_i + k_r} ; \beta = - \frac{k_i}{k_i + k_r} ; \gamma = 0 .$$

8.3 CENTROID SORTING

Centroid sorting (Sect. 2.2) can be applied to any coefficient which measures the similarity between two clusters. That is, if we can measure the similarities S_{pq} , S_{ip} and S_{iq} , then we can also measure the similarity $S_{i(p+q)}$ between cluster i and the cluster produced by fusion of p with q. For there to exist a combinatorial solution for centroid sorting using the similarity criterion S, it must be possible to express $S_{i(p+q)}$ as a function of the known parameters S_{pq} , S_{ip} , S_{iq} , k_p , k_q and k_i . We now derive such transformations for four additional similarity criteria, where U_{pj} = jth. coordinate of the centroid vector for cluster p.

Dot product:

$$D_{pq} = \frac{1}{M} \sum_j U_{pj} U_{qj}$$

$$\begin{aligned} \text{Hence } D_{i(p+q)} &= \frac{1}{M} \sum_j U_{ij} \left[\frac{k_p U_{pj} + k_q U_{qj}}{k_p + k_q} \right] \\ &= \frac{k_p}{k_p + k_q} D_{ip} + \frac{k_q}{k_p + k_q} D_{iq} \end{aligned}$$

It follows that centroid sorting using a matrix of dot product coefficients is obtained for the combinatorial parameters:

$\alpha_p = k_p/k_r$; $\alpha_q = k_q/k_r$; $\beta = \gamma = 0$; that is, using the average linkage transformation. It is interesting to note that this result proves that average linkage and centroid sorting produce identical results with dot product.

Dispersion:

$$R_{pq} = \frac{1}{M} \sum_j (U_{pj} - \bar{U}_p)(U_{qj} - \bar{U}_q)$$

$$\text{where } \bar{U}_p = \frac{1}{M} \sum U_{pj}$$

$$\text{Hence } R_{pq} = \frac{1}{M} \sum U_{pj} U_{qj} - \frac{1}{M^2} \sum U_{pj} \sum U_{qj}$$

For the combinatorial solution we have:

$$\begin{aligned} R_{i(p+q)} &= D_{i(p+q)} - \frac{1}{M^2} \sum \left[\frac{k_p U_{pj} + k_q U_{qj}}{k_p + k_q} \right] \sum U_{ij} \\ &= \frac{k_p}{k_p + k_q} \left[D_{ip} - \frac{1}{M^2} \sum U_{pj} \sum U_{ij} \right] + \\ &\quad \frac{k_q}{k_p + k_q} \left[D_{iq} - \frac{1}{M^2} \sum U_{qj} \sum U_{ij} \right] \\ &= \frac{k_p}{k_p + k_q} R_{ip} + \frac{k_q}{k_p + k_q} R_{iq} \end{aligned}$$

It follows that average linkage and centroid sorting produce identical results for dispersion coefficients, the combinatorial transformation parameters being: $\alpha_p = k_p/k_r$; $\alpha_q = k_q/k_r$; $\beta = \gamma = 0$.

Size difference: $Z_{pq} = \frac{1}{M^2} \left[\sum U_{pj} - \sum U_{qj} \right]^2$

Therefore

$$Z_{i(p+q)} = \frac{1}{M^2} \left[\frac{k_p \sum U_{pj} + k_q \sum U_{qj}}{k_p + k_q} - \sum U_{ij} \right]^2$$

This form is almost identical to the combinatorial solution for distances (see Lance and Williams, 1967a), and it is easily shown that $Z_{i(p+q)}$ expands to

$$\begin{aligned} Z_{i(p+q)} &= \frac{k_p}{M^2(k_p + k_q)} \left[(\sum U_{pj})^2 - 2 \sum U_{pj} \sum U_{ij} + (\sum U_{ij})^2 \right] \\ &+ \frac{k_q}{M^2(k_p + k_q)} \left[(\sum U_{qj})^2 - 2 \sum U_{qj} \sum U_{ij} + (\sum U_{ij})^2 \right] \\ &- \frac{k_p k_q}{M^2(k_p + k_q)^2} \left[(\sum U_{pj})^2 - 2 \sum U_{pj} \sum U_{qj} + (\sum U_{qj})^2 \right] \\ &= \frac{k_p}{k_p + k_q} Z_{ip} + \frac{k_q}{k_p + k_q} Z_{iq} - \frac{k_p k_q}{(k_p + k_q)^2} Z_{pq} \end{aligned}$$

Shape difference:

$$\begin{aligned} S_{pq} &= \frac{1}{M} \sum (U_{pj} - U_{qj})^2 - \frac{1}{M^2} \left[\sum U_{pj} - \sum U_{qj} \right]^2 \\ &= d_{pq}^2 - Z_{pq} \end{aligned}$$

where d_{pq}^2 is the euclidean distance coefficient, and Z_{pq} is the size difference coefficient. Hence

$$\begin{aligned}
 S_{i(p+q)} &= d_{i(p+q)}^2 - Z_{i(p+q)} \\
 &= \frac{k_p}{k_p + k_q} d_{ip}^2 + \frac{k_q}{k_p + k_q} d_{iq}^2 - \frac{k_p k_q}{(k_p + k_q)^2} d_{pq}^2 \\
 &\quad - \frac{k_p}{k_p + k_q} Z_{ip} - \frac{k_q}{k_p + k_q} Z_{iq} + \frac{k_p k_q}{(k_p + k_q)^2} Z_{pq} \\
 &= \frac{k_p}{k_p + k_q} S_{ip} + \frac{k_q}{k_p + k_q} S_{iq} - \frac{k_p k_q}{(k_p + k_q)^2} S_{pq}
 \end{aligned}$$

It follows that the combinatorial transformation parameters for centroid sorting are valid for both size and shape difference coefficients, as well as d^2 .

8.4 COMBINATORIAL ALGORITHM

Wishart (1969c) describes briefly the following computer program for the combinatorial algorithm. Let $(K(i), i = 1, N)$ be the vector of cluster sizes, where $K(i) = 0$ if cluster i is inactive, and $(C(i), i = 1, N)$ be the classification array such that $C(i)$ is the code of the cluster containing the i th. individual. then:

- (a) Find the least 'distance' in D for active sets

$$d_{pq}^2 = \min (d_{ij}^2; i = 1, j - 1; j = 2, N; K(i) > 0; K(j) > 0).$$
 If D comprises similarities, then search for the greatest element.
- (b) Print the clusters p and q being fused, together with the fusion coefficient d_{pq}^2 .
- (c) Replace cluster p with cluster $(p+q)$ by modifying the row

(d_{ip}^2 ; $i = 1, p - 1$; $K(i) > 0$) and column (d_{pj}^2 ; $j = p + 1, N$; $j \neq q$; $K(j) > 0$) of D using the appropriate transformation (at the sth. fusion cycle there will be $N-s-1$ such modifications to D).

(d) Set $K(p) = K(p) + K(q)$, and $K(q) = 0$ to render cluster q inactive.

(e) Reclassify the elements of cluster q with cluster p . This is achieved by scanning vector C and replacing all elements $C(i) = q$ with $C(i) = p$. The procedure returns to (a) and is repeated for $(N - 1)$ fusion cycles.

This algorithm, as used for Ward's method, is concisely represented by the flow chart given in Figure 8.4.1. However, steps (a) and (c) require some further clarification. Suppose that for a population of 7 objects, whose distance matrix is given in Table 8.4.1, we have the following first two fusion cycles:

| CYCLE | p | q | d_{pq}^2 |
|-------|---|---|------------|
| 1 | 3 | 4 | 0.2 |
| 2 | 6 | 7 | 0.4 |

so that clusters 4 and 7 are inactive, and clusters 3 and 6 comprise 2 objects each. Then we define the following arrays: $DX(i) = \min(d_{ij}^2; j = 1, i - 1; K(j) > 0)$, provided that $K(i) > 0$, $KD(i) = J$ such that $d_{iJ}^2 = DX(i)$; or zero if no active d_{iJ}^2 exists. That is, $DX(i)$ is the least distance with active cluster J in the row of D associated with active cluster i , and $KD(i)$ is the code of cluster J .

Values for DX and KD are given in the example in Table 8.4.1, and the least elements of matrix rows for active clusters are underlined in the distance matrix D. We further set $KD(i) = 0$ if there is no active cluster j amongst $j = 1, i - 1$.

| | | | | | | | |
|------------|--------------|--------------|-----|-----|------------|-----|---|
| R_2 | <u>5.4</u> * | | | | | | |
| R_3 | 3.2 | <u>2.1</u> * | | | | | |
| 4 | 2.5 | 2.4 | 0.2 | | | | |
| 5 | 3.4 | <u>1.1</u> | 2.0 | 2.4 | | | |
| 6 | 5.1 | <u>9.2</u> * | 8.3 | 4.2 | <u>4.0</u> | | |
| 7 | 2.2 | 1.2 | 3.1 | 4.5 | 6.2 | 0.4 | |
| Object no. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| C | 1 | 2 | 3 | 3 | 5 | 6 | 6 |
| K | 1 | 1 | 2 | 0 | 1 | 2 | 0 |
| DX | - | 5.4 | 2.1 | - | 1.1 | 4.0 | - |
| KD | 0 | 1 | 2 | 0 | 2 | 5 | 0 |

Table 8.4.1: Explanation of the modification phase of the combinatorial algorithm

* Least row element for active clusters which are to be changed

R Rows which must be treated for new minimum DX values

Observe that the least d_{pq}^2 in D is given by the least $DX(q)$ value, where $p = KD(q)$. Step (a) of the combinatorial algorithm merely searches DX for the least element $DX(i)$ for which $KD(i) > 0$. This finds p, q and d_{pq}^2 for the current fusion cycle.

Step (c) of the algorithm requires an efficient method of updating the DX and KD values in order that step (a) may be repeated in the next cycle.

Firstly, we observe that no row of D associated with an object coded less than p ($p < q$) need be examined. This is because

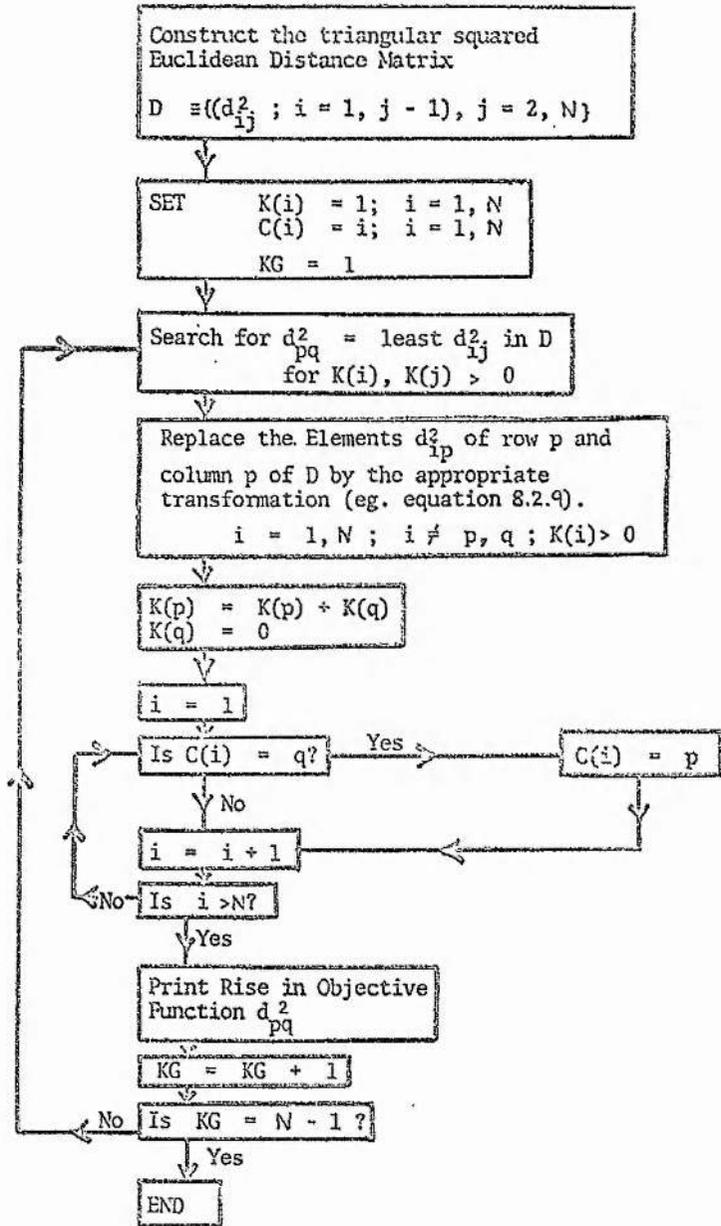


Figure 8.4.1. Flow diagram for the combinatorial algorithm using transformation 8.2.9 for Ward's method

the elements of such rows cannot belong to the p th. row and column of D (which are to be changed) and therefore the previous DX and KD values will be unaltered.

Secondly, any row i for which $KD(i) = 0$ need not be examined because this means that cluster i cannot have a least $DX(i)$ value, either because there is no active cluster j amongst $j = 1, i - 1$, or because cluster i is inactive.

Thirdly, no full examination of a row i for which $KD(i) > 0$ and $KD(i) \neq p, q$ is required. This condition implies that $DX(i)$ is associated with a cluster $KD(i) = s$ other than p or q . Hence, either the new distance d_{pi}^2 is less than the current $DX(i)$, in which case we replace $DX(i)$ with d_{pi}^2 and set $KD(i) = p$. Alternatively, $DX(i) \geq d_{pi}^2$, in which case $DX(i)$ and $KD(i)$ remain unchanged.

Finally, any row i for which $KD(i) = p$ or q may have to be examined in full in order that $DX(i)$ and $KD(i)$ can be updated. This is because $DX(i)$ refers to either d_{pi}^2 or d_{qi}^2 which are currently to be modified. If the resulting coefficient d_{pi}^2 is greater than $DX(i)$ (in the case of dissimilarities), then it is possible that there exists another active cluster s for which $d_{si}^2 < d_{pi}^2$. In this case, the distance d_{pi}^2 must be evaluated by the transformation, and the entire i th row of D must be examined for a new minimum distance $DX(i)$.

In the example, the least $DX(i)$ value is $DX(5) = 1.1$ for which $KD(5) = 2$. Hence $p = 2$, $q = 5$ and $d_{pq}^2 = 1.1$, and step (c) of the

algorithm requires the following action:

1. Compute the new element d_{12}^2 using the appropriate transformation, and insert this value into D and DX(2). Suppose that $d_{12}^2 = 2.9$.

2. Since $KD(3) > 0$, it is necessary to replace d_{23}^2 in row 3 of D with the new distance computed by the transformation. Suppose

| | | | | | | | |
|------------|------------|------------|-----|-----|-----|-----|---|
| 2 | <u>2.9</u> | | | | | | |
| 3 | <u>3.2</u> | 3.8 | | | | | |
| 4 | 2.5 | 2.4 | 0.2 | | | | |
| 5 | 3.4 | 1.1 | 2.0 | 2.4 | | | |
| 6 | 5.1 | <u>3.8</u> | 8.3 | 4.2 | 4.0 | | |
| 7 | 2.2 | 1.2 | 3.1 | 4.5 | 6.2 | 0.4 | |
| Object no. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| C | 1 | 2 | 3 | 3 | 2 | 6 | 6 |
| K | 1 | 2 | 2 | 0 | 0 | 2 | 0 |
| DX | - | 2.9 | 3.2 | - | - | 3.8 | - |
| KD | 0 | 1 | 1 | 0 | 0 | 2 | 0 |

Table 8.4.2: Values for D, C, K, DX and KD following the 3rd fusion cycle of the combinatorial algorithm for the worked example (see text for explanation)

that this value is 3.8 which exceeds 2.1 (the previous DX(3) value), then the full row 3 of D must be searched for the new minimum, since

$KD(3) = 2 = p$. It happens that $d_{13}^2 = 3.2$ is the new minimum; hence $d_{23}^2 = 3.8$, $DX(3) = 3.2$ and $KD(3) = 1$.

3. Row 4 is ignored because $KD(4) = 0$ (cluster 4 became inactive at fusion step 1).

4. Row 5 is ignored because cluster 5 is currently being joined to cluster 2 and will become inactive.

5. Compute the new element d_{26}^2 and insert this value into the 6th row of D. Suppose that $d_{26}^2 = 3.8$, which happens to be less than the previous least value $DX(6) = 4.0$, then we substitute $DX(6) = 3.8$ and $KD(6) = 2$. If d_{26}^2 had exceeded $DX(6)$, then the entire 6th row would have had to be searched for a new minimum because $KD(6) = q = 5$.

6. Row 7 is ignored because $KD(7) = 0$ (cluster 7 is inactive). Set $KD(5) = 0$, and proceed to step (d) of the algorithm. Table 8.4.2 shows the current state of the matrix D and the four vectors at the end of the 3rd cycle, and the next fusion will therefore combine clusters 1 and 2 at coefficient 2.9.

Note: The above algorithm has been described in the form suitable for a dissimilarity matrix. If D contains similarities, then all the coefficient tests and inequalities should be reversed, and the terms 'minimum' and 'least' should be replaced with 'maximum' and 'greatest'.

CHAPTER 9: K-PARTITION

9.1 JARDINE-SIBSON ALGORITHM

Suppose that for a population P , comprising N individuals, a similarity matrix D is computed using a suitable coefficient of association, and a linkage parameter k and similarity threshold h are chosen. Each individual may be represented by a vertex on a graph, and all pairs of vertices which correspond to pairs of individuals having a similarity of at least h are connected. All maximal complete subgraphs are found and all pairs of such subgraphs that intersect in at least k vertices are further connected. This leads to the concept of a chain of connected subgraphs. In particular, since a sufficiently dense set of points in Euclidean space could form such a chain cluster having any shape or variance, the method when applied to metric data is of the 'natural class' type (see Chapter 6). The method induces overlapping clusters since two intersecting maximal complete subgraphs which have less than k overlap vertices are distinguished as separate clusters. When $k = 1$ no overlap occurs and the procedure is identical to single linkage.

Method I below is a summary of the method proposed by Jardine and Sibson (1968) for deriving a numerical representation of this 'k-partition' clustering from the similarity matrix D . The modified matrix U that results from these operations on D expresses the connections completed within all the chain clusters. A method of con-

structuring the clusters from U is discussed later.

Method I

1. In one 'scan' of D, all possible subsets of P comprising exactly $k+2$ objects are considered.
2. For each subset determine the least and second least similarities h' and h'' respectively. If $h' < h''$ replace h' by h'' in D. If $h' = h''$ leave D unchanged.
3. Step 2 is repeated on the reduced similarity matrix until every subset contains a non-unique minimum similarity (if dissimilarities are used, the expressions 'least', $h' < h''$ and minimum are replaced by 'greatest', $h' > h''$ and maximum respectively: see, for example, Jardine and Sibson, 1968). This condition is detected when no further transformations of the reduced matrix occur during one complete scan. The reduced matrix is then the modified matrix U.

Similar algorithms are used by Johnson (1967) and Roux (1968) for the ultrametric transformation in the special case $k = 1$.

Method I evidently requires a considerable amount of unproductive work. During the second and subsequent scans of D, the examination of the majority of the $\binom{N}{k+2}$ subsets of P will yield no additional transformation, and the last scan of D is completely unproductive since the matrix is merely checked for the lack of any further modification. Also, the work required for the algorithm is proportional to

$$W_I = \frac{p}{2} (k+2)(k+1) \times \frac{N!}{(N-k-2)!(k+2)!}$$

where p is the number of scans required. Clearly, when $N \gg k$

$$W_I \approx \frac{p}{2k!} N^{(k+2)}$$

and the population will be restricted by most computers to $N \leq 20$.

9.2 COLE-WISHART ALGORITHM

One approach (Cole and Wishart, 1970)¹ to the reduction of W_I is to eliminate as much of the unproductive work as possible. Williams, Lambert and Lance (1966) define a single linkage algorithm which uses an ordering of the similarities in D , tagged by their corresponding object pairs. The method works progressively through this ordered similarity list deriving the single linkage hierarchic dendrogram in the process, and since the similarities associated with each fusion on the dendrogram correspond precisely to the ultrametric distances when $k = 1$, the method incidentally develops the elements of U . We choose therefore to adapt this approach to the construction of U for all k , and propose the following basic algorithm together with its subsequent improvements as a means of inserting similarities directly from the ordered similarity list into U .

¹The remainder of this chapter is extracted directly from our paper, which at the time of the preparation of this thesis was still in press. I acknowledge Professor Cole's idea to form U using a method which inserts similarities directly into an empty matrix, and I express my gratitude for the many helpful sessions of discussion which resulted in the formulation of Method II.

Method II

1. Tag the elements of the similarity matrix D and order them by descending similarity in an array Q. Initialise an empty matrix U which is to receive the reduced matrix associated with D.

2. Sequentially remove similarities from Q. If we are considering the pth similarity Q_p corresponding to the object pair (i,j) and if U_{ij} has been filled, then we ignore Q_p and proceed to Q_{p+1} . If U_{ij} is not filled, then we set $R = Q_p$, put the object pair (i,j) at the head of a new 'insertion list' L and then proceed to 3.

3. From the insertion list L, select object pairs (g,h) corresponding to entries that are to be made in U. Initially L contains one pair only, but the procedure below may cause additional object pairs to be added. If U_{gh} is filled, then we consider the next pair in L. If U_{gh} is not filled, we set $U_{gh} = R$ and proceed to 4.

4. Now consider all (k+2) element subsets of P containing both g and h. Associated with each subset $S = (\alpha_1, \alpha_2, \dots, \alpha_k, g, h)$ we have the set η of similarities $(U_{ij}; i, j \in S)$, which can be partitioned into two subsets η', η'' where η' contains the elements U_{ij} of η which are not filled in U, and η'' is the complement of η' in η , containing those elements U_{ij} which have been filled in U. We observe that the least similarity in η'' is R and no element of η' is greater than R by virtue of the ordering in Q. If, for some subset

S, η' is a single unfilled element d' corresponding to $U_{g'h'}$, then $d' \in R$. Hence, either d' is a unique minimum in which case Method I requires that we replace $U_{g'h'}$, by R , or else $d' = R$. We therefore set $U_{g'h'} = R$ and add the pair (g', h') to L .

5. When all subsets containing objects g and h have been considered, we return to 3 and select a new pair from the insertion list L .

6. When L is finally exhausted, we check U for unfilled elements. If U is completely filled then we exit; otherwise we return to 2 and consider new similarities from Q for insertion.

A comparison between Methods I and II

We shall omit the rather tedious formal proof that Method II derives the correct reduced matrix U . However, it is not immediately clear that the final matrix U obtained by Method II cannot be further modified by a Method I-type scan. We shall call steps 3-6 of Method I one 'cycle' of the algorithm.

Let p be the set of all subsets of P containing $(k+2)$ objects. Partition p into p'_i and p''_i , where p''_i contains all the subsets of p which, at the end of the i th cycle, are currently maximal complete subgraphs in U (that is, subsets S for which all $\frac{1}{2}(k+2)(k+1)$ similarities η have been filled in U). Let p'_i be the complement of p''_i in P ; that is

$$p = p'_i + p''_i \text{ and } p'_i \cdot p''_i = \emptyset$$

We observe that for each subset $S \in p'_i$, $|\eta'| \geq 2$, since step 4 of the algorithm always completes the subgraphs for those subsets having $|\eta'| = 1$. Let

$$E_i = p_i'' - p''_{i-1}$$

be the set of all maximal complete subgraphs of $(k+2)$ objects which are completed during the i th cycle; then for each $S \in E_i$ we have $|\eta'| \geq 2$ at the start, and $|\eta'| = 0$ at the end of the i th cycle. Hence at least 2 of the similarities η_i for S are filled during the i th cycle, and therefore, by virtue of the ordering on Q , R is the non-unique minimum similarity in η_i for all $S \in E_i$. It follows that the similarities η_i cannot be further modified by a Method I-type reduction on any $S \in E_i$. But

$$p_i'' = E_i + p_{i-1}'' = E_i + E_{i-1} + p_{i-2}'' = \dots = E_i + E_{i-1} + \dots + E_1$$

Hence p_i'' contains no subset S for which η_i can be modified by a Method I-type reduction. Now suppose at the t th cycle of Method II all elements of U have been filled and the algorithm terminates, then

$$p_t' = \emptyset, \text{ and } p_t'' = p.$$

Hence for all $S \in p$, the similarities in η cannot be further modified by a Method I-type ultrametric reduction. It follows that the matrix U obtained by Method II cannot be further modified by Method I.

Stage 4 of Method II requires a search through all $(k+2)$ -

subsets of P which contain the pair of objects (g,h). Since there are $\binom{N-2}{k}$ such subsets for each of the $\frac{1}{2}N(N-1)$ entries to be filled into U, the ratio of the work required for Methods I and II is

$$\frac{W_{II}}{W_I} = \frac{\frac{1}{2}N(N-1)}{P} \times \frac{\binom{N-2}{k}^C}{N^C(k+2)} = \frac{1}{2p} (k+2)^{(k+1)}$$

where p is the number of scans of U which are required for the completion of Method I. Since p is usually in the range 3 to 5, it is clear that Method II requires considerably more work than Method I except for the marginal case when $k = 1$. However, the approach of Method II now suggests the following three ways of reducing the amount of work W_{II} :

(i) After any one entry in U, the number of subsets that must be examined at stage 4 can be minimised by excluding certain objects from the $\binom{N-2}{k}$ possibilities.

(ii) During stage 4, the size of these subsets can be reduced under certain conditions when a local value of k, which is smaller than the general value of k, is adopted. In general, this also reduces the number of subsets that must be examined.

(iii) We also consider situations in which the general value of k can be reduced, and it is shown, in the paragraph on ending conditions, that the algorithm can be terminated as soon as k rows of U have been completely filled. Hence the factor $\frac{1}{2}N(N-1)$ for W_{II} can be improved.

For the ensuing development we shall adopt certain new terms which are defined as follows:

1. If, at some stage of Method II, the similarity U_{ij} is filled in U then we say that objects i and j are 'connected'; similarly, if U_{ij} is not filled then i and j are 'not connected' or 'disconnected'.

2. Any subset of q objects for which all pairs of objects are connected is called a 'complete q -subset'.

3. Any subset of q objects for which all but one of the possible pairs of objects are connected is called an 'almost complete q -subset', which we abbreviate to an 'a.c. q -subset'.

4. In any subset of objects S , s_i is the number of objects in S which are connected to the i th member of S . This convention is applied to the base subset B (defined below), and the object universe P , where b_i and p_i are the respective connection counts for the i th object.

Step 4 of Method II can now be described as a search for all a.c. $(k+2)$ -subsets of P which contain objects g and h , and in the next three sections we discuss means of improving the efficiency of this search.

Reducing the number of subsets

Let $S = (\alpha_1, \alpha_2, \dots, \alpha_k, g, h)$ be an a.c. $(k+2)$ -subset, then by definition exactly one of the connections within S is missing. Hence k of the members of S have $(k+1)$ connections within S , while

the remaining two members are disconnected and have exactly k other connections. Since the connection U_{gh} is completed during step 3 of Method II, g and h cannot be the disconnected pair of objects, and therefore

$$\begin{aligned} \text{either } s_g \geq k \text{ and } s_h = k+1 \\ \text{or } s_g = k+1 \text{ and } s_h \geq k \end{aligned} \tag{9.2.1}$$

Also for every object $i \in S$, we have

$$s_i \geq k \tag{9.2.2}$$

and it follows that each $i \in S$, $i \neq g, h$ is at least connected to g or h . We can, therefore, restrict the objects α_i , which are placed in S during the search for a.c.($k+2$)-subsets, to those objects which are at least connected to g or h , and we define the base subset B as the set of all such objects together with objects g and h . Associated with each object $i \in B$ we have the number of connections b_i between i and the other objects in B .

Note that for each object $i \in S$, where S is any subset of ($k+2$) objects taken from B , $s_i \leq b_i$; it follows, from 9.2.1 above, that no such a.c.($k+2$)-subset S exists unless

$$\begin{aligned} \text{either } b_g \geq k \text{ and } b_h \geq k+1 \\ \text{or } b_g \geq k+1 \text{ and } b_h \geq k \end{aligned} \tag{9.2.3}$$

Similarly, suppose that for some object $s_i \in B$, $i \neq g, h$, we have

$$b_i < k.$$

then for any subset S containing i , $s_i \leq b_i < k$, and hence S is not an a.c.($k+2$)-subset, by (9.2.2) above. Therefore, no a.c.($k+2$) subset exists which includes an object i for which $b_i < k$, and hence such objects can be removed from B . If there are any removals, the b_i 's for the remaining objects are recomputed and the procedure iterates until there are no further removals. The search for a.c.($k+2$)-subsets can be concluded if, at any stage of this procedure,

- either (i) B is now empty,
- or (ii) B is not empty, but either g or h have been removed,
- or (iii) B is not empty, but condition (9.2.3) is no longer satisfied,

in which case we return to stage 3 of the algorithm. Otherwise, B contains a list of objects, including g and h , which may form a.c.($k+2$)-subsets and we proceed to consider methods for reducing the value of k and hence the size of the subsets that must be examined in the subsequent search.

Reducing the size of the subsets

We denote the number of objects in the base subset B by $|B|$, and hence

$$b_i \leq |B| - 1$$

Suppose that for some object $i \in B$, $b_i = |B| - 1$; that is, i is connected to every other object in B . Then if i is removed from B we can reduce the value of k by 1 in the search for a.c.($k+2$)-subsets of B . Let B' be the subset of B which excludes object i , and let

$S' = (\alpha_1, \alpha_2, \dots, \alpha_{k+1})$ be any a.c.(k+1)-subset of objects taken from B' . Then the subset $S = (\alpha_1, \alpha_2, \dots, \alpha_{k+1}, i)$ is an a.c.(k+2)-subset since $b_i = |B| - 1 \rightarrow i$ is connected to every $\alpha_j \in S'$. Notice that the removed object i could be either object g , object h or some other member of B , but we always require that S contains both objects g and h (the only a.c.(k+2)-subsets which can be found contain both objects g and h by virtue of the insertion of U_{gh}). This result can be generalised to the extent that if t members $i \in B$ satisfy $b_i = |B| - 1$, and these t objects are removed to form the residual base subset B' , then an a.c.(k+2-t)-subset S' of B' becomes an a.c.(k+2)-subset S of B with the addition of the t completely connected objects previously removed from B . Furthermore, the single disconnected pair of objects in S' will be the same disconnected pair in S , and in the limiting case when $t \geq k$ but $|B'| > 0$, any disconnected pair of objects in B' is associated with an a.c.(k+2)-subset of B .

To implement these results we use a local value k_L of k which applies throughout the search for a.c.(k+2)-subsets of B . Initially $k_L = k$, and each object $i \in B$ for which $b_i = |B| - 1$ is removed from B and k_L is reduced by 1. When all such removals are complete, and provided that $k_L > 0$, B is searched for a.c.(k_L+2)-subsets subject to the inclusion of either or both objects g and h provided that either or both objects g and h are retained in B ; when $k_L \leq 0$ and $|B| > 0$, B is searched for disconnected pairs of objects. When such

a subset or disconnected pair is found, its single residual disconnection is completed in U with the current similarity R , and the associated object pair is added to L (as described at Stage 4). If after all removals B is empty, no a.c. $(k+2)$ -subsets are to be found and the search is concluded. In the computer program KDEND, the search for (k_L+2) -subsets is further reduced by first considering triples of objects from B . When a complete or a.c. triple is found in B , other objects are tested for addition until a complete or a.c. 4-subset is found, and so on, until an a.c. (k_L+2) -subset has been obtained. The choice of objects for addition is, at each stage, restricted to those which have neither been considered in a previous base triple, nor have been considered at a previous stage in the generation of the current subset. Furthermore, when objects are being tested for addition to a current subset of size r , it is only necessary to examine r connections to determine if the addition yields a complete or a.c. $(r+1)$ -subset. This procedure, which is used only when $k_L > 1$, further reduces the amount of work required for the consideration of all subsets of size (k_L+2) taken from B .

Ending conditions

In the previous section it was shown that we can remove from B any object i for which $b_i = |B| - 1$, and reduce the local k_L value by 1. If we consider the universe of objects P as base subset, where p_i is the number of overall connections for object i ,

then the result can be generalised as follows: if there exists an object $i \in P$ such that $p_i = N - 1$, then we can remove object i from further consideration and reduce the general value of k by 1. Furthermore, when $k = 0$ after such a reduction, all empty elements of U can be filled with the current similarity \bar{a} , that is, the matrix U can be completed as soon as k rows or columns are filled. In the computer program, the overall connections counts p_i are stored and updated at each insertion into U . When one such count reaches $N - 1$, the associated object i is deleted from further consideration and the value of k is reduced by 1; however, if such an insertion occurs during a scan of the base subset B and $k \neq 0$, it is important to retain object i with the current local value of k_L unmodified until the scan of B has been completed. When an insertion reduces k to zero, we exit from further scanning of B and complete all empty elements of U with the current similarity R . We are now finished, and U contains the final reduced matrix.

9.3 CLUSTER RECOGNITION ALGORITHM

Jardine and Sibson (1968) appear to have avoided the problem of automatically isolating clusters from the reduced matrix U , and simply advocate their construction by hand for any chosen similarity threshold h . This process, although not difficult, can be tedious and, therefore, we propose an algorithmic solution to the problem.

A binary linkage matrix L is defined as follows: for any chosen similarity threshold h , we set

$$\begin{aligned} L_{ij} &= 1 \text{ if } U_{ij} \geq h \\ \text{or } L_{ij} &= 0 \text{ if } U_{ij} < h \\ \text{and } L_{ii} &= 1 \text{ for all } i \end{aligned}$$

Figure 9.3.1 shows the linkage matrix L obtained when $h = 5.50$, from the reduced matrix U derived in the example, used by Jardine and Sibson, when $k = 3$. Also shown is a linkage diagram for this set of 9 objects, on which connected objects are joined and clusters are indicated by dotted lines. We notice that there are two types of objects: 'explicit' objects belong to one cluster only, and 'overlap' objects belong to two or more clusters. Also, for each cluster of objects we know that every member is connected to every other member. The problem of cluster recognition is to isolate the maximal complete subgraphs expressed in L , and remove them one at a time until all such clusters have been found. Furthermore, in removing connections from L , it is important to retain those connections to overlap objects which are used to describe other clusters. For example, suppose we remove cluster (159) (shown in figure 9.3.1) and in doing so delete the (5,9) connection, then cluster (569) will not be recognised later. Our approach is to search for explicit (see below) objects, remove the clusters containing them, and then delete only those connections to each explicit object. Hence, in our example of cluster (159) we discover that object 1 is explicit and so we reset the first row and column of L to zero.

An 'overlap' object is defined as an object i , connected to

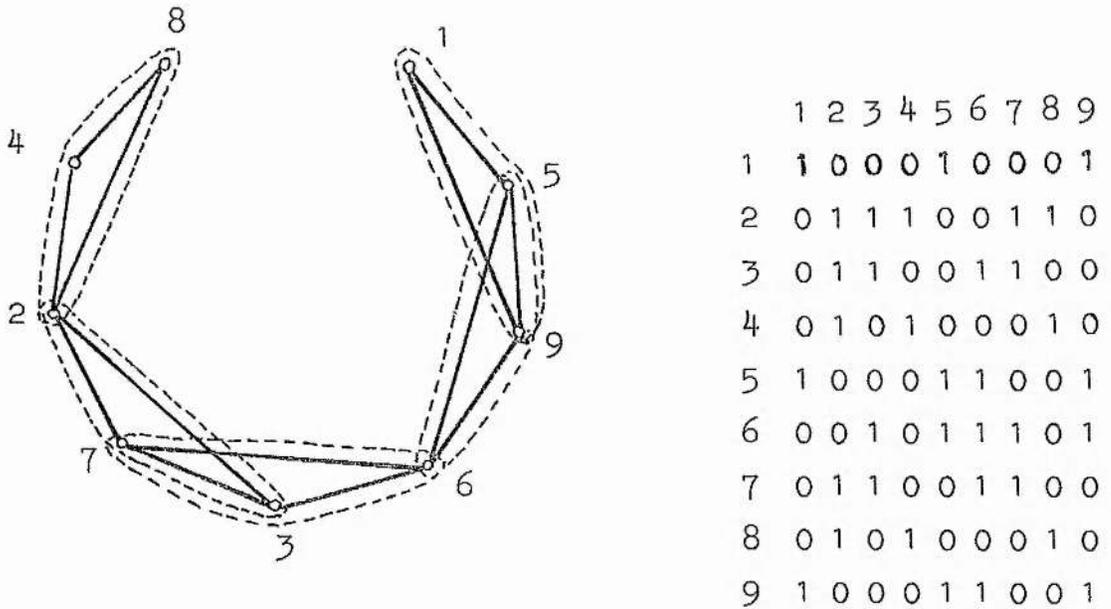


Figure 9.3.1. An example of overlapping clusters for a 9-object population, and the associated linkage matrix derived from the ultrametric. Clusters are indicated by dotted lines, and solid lines join objects whose similarity exceeds the threshold. In the linkage matrix significant similarities are coded 1. Observe that only objects 1, 4 and 8 are explicit.

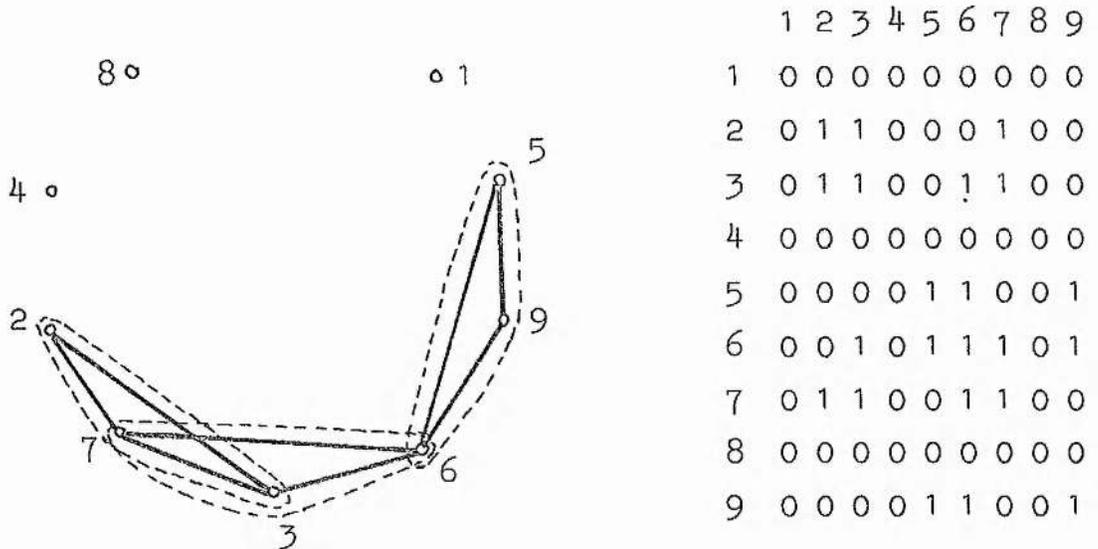


Figure 9.3.2. Residual clusters and linkage matrix for the 9-object population in Figure 9.3.1 after 2 cycles of the cluster recognition algorithm. Clusters (248) and (159) have been recognised and removed from the linkage matrix with the result that objects 2, 5 and 9 are now explicit. The next cluster that will be recognised is (237).

objects $(\alpha_1, \alpha_2, \dots)$ for which at least one of the pairs (α_j, α_k) is not connected. Any other object is termed 'explicit'. It follows that only overlap objects belong to more than one cluster. In our example, object 1 is explicit because it is connected to 5 and 9, and the pair (5,9) is also connected. By contrast, object 5 has connections to objects 1, 6 and 9, and since the pair (1,6) is not connected, 5 is an overlap object. We observe that objects 1, 4 and 8 are the only explicit objects in figure 9.3.1.

The first step in the cluster recognition algorithm is to search for isolated objects. Each such object is characterised by a 1 in the diagonal and zeros elsewhere of its associated row and column in L. The diagonal 1 element is replaced by 0 for each isolated object, and a 'single object cluster' message is printed. Next we look for explicit objects and their parent clusters by the following terminating procedure:

- (1) We scan L for objects that have 1's in their associated row: suppose we find that the i th object has connections in L, then vectors I and I' are set equal to the i th row of L.
- (2) For each '1' in I , corresponding to a connection (i,j) , we replace I' by the product of vector I' and the j th row in L.
- (3) When all 1's in I have been treated in this way, the current vector I' is compared with the original vector I : if they are the same, then object i is explicit; if they differ, then i is an overlap object. We ignore any overlap objects and continue the scan of

L until an explicit object is found, when we proceed to 4.

(4) Suppose there are c 1's in I , then each 1 corresponds to a member of the current cluster C , whose size is c . Since the row and column of L corresponding to each explicit object in C will be identical to I , the number of connections in L for each such object will be c . We now form a list $E = (\alpha_1, \alpha_2, \dots)$ of objects $\alpha_i \in C$ such that the row in L corresponding to α_i contains exactly c 1's; E is therefore a list of all the explicit objects in C .

(5) The rows and columns of L corresponding to each $\alpha_i \in E$ are now reset to zero, and we return to (1) and look for new explicit objects and their parent clusters. When every element of L is zero, the procedure is terminated and all clusters have been found.

This process, when applied to the linkage diagram of figure 9.3.1, will first recognise cluster (159), find the explicit object 1 and then reset row 1 and column 1 of L to zero. Next, explicit object 4 will be discovered and the parent cluster (248) recognised. In this case, E will contain explicit objects 4 and 8, and so the corresponding rows and columns of L will be reset to zero. Figure 9.3.2 shows the stage that has now been reached. Two clusters have been peeled off the chain of overlapping clusters with the result that objects 2, 5 and 9, which were previously overlap objects, are now explicit. When the algorithm is reapplied, the clusters (237), (367) and (569) are recognised, at which point

L contains zeros everywhere and the scan is concluded.

So far we have omitted one special case: the situation when every object is in an overlap position. This may occur with the entire population or with a subset of objects, but in any event, it is detected when steps 1-3 of the algorithm fail to find an explicit object while there are still 1's in L. Figure 9.3.3 shows the connections between six overlap objects, for which the clusters are (1245), (1246), (1345) and (1346). We observe that if any overlap object i has p_i residual connections in L (the connection L_{ii} is included), then the maximum size of cluster to which i can belong cannot exceed $p_i - 2$. Let p_{\max} be the maximum number of residual connections in L for any one object, then a search through all subsets comprising from 2 to $p_{\max} - 2$ overlap objects will reveal all overlapping clusters. Our approach is to first consider all subsets of size $p_{\max} - 2$ objects, and form a list of those subsets which are maximal complete subgraphs in L. Next, all subsets of size $p_{\max} - 3$ are considered and any maximal complete subgraph is added to the list provided that it does not form a subset of a cluster previously found. As each new cluster is discovered, a check is made to test whether the present list of clusters accounts for all the residual connections in L. When this happens the process is terminated. Although this procedure appears to be lengthy, in practice the total overlap condition occurs for small spherical groupings of overlapping clusters (as in figure 9.3.3) or for connected circuits of overlapping clusters, because of the severe

requirement that every object must belong to more than one cluster. Consequently p_{\max} is usually small, and the exhaustive search for clusters is often terminated at or near the subset size $p_{\max} - 2$ level. In the example shown in figure 9.3.3, $p_{\max} = p_1 = 6$ so that only the 15 subsets of size 4 need be considered.

Finally, one modification to the cluster recognition algorithm has to be made. A cluster can contain subsets of explicit and overlap objects, respectively. When a cluster is recognised, those connections to the explicit objects are removed from L while the connections to the overlap objects are retained in L. If the overlap objects happen to belong to more than one other cluster, then this subset of overlap objects will be recognised later as a separate cluster. This means that not only clusters, but also subsets of clusters will be recognised by the algorithm. To correct this case, the computer program compiles a list of the clusters as they are found. When the cluster recognition phase is terminated, this list is searched for duplicated cluster subsets which are removed before the final classifications are printed.

Obtaining a hierarchy of clusterings

The cluster recognition algorithm is defined for one chosen similarity threshold h . However, if we apply the algorithm using each unique entry contained in U as threshold, we obtain the hierarchy of all the clusterings that can possibly be generated for any chosen value of h . In their introduction of Method I, Jardine and Sibson distinguish between the variants of the sequence which yield

non-overlapping ($k = 1$) and overlapping ($k > 1$) clusters by the terms 'hierarchical' and 'non-hierarchical' respectively. This use of the term 'hierarchy' to describe a sequence of nested partitions which give rise to strictly disjoint subsets, differs from the conventional meaning where 'hierarchical' refers to those methods that produce ordered clusterings from a monotonic decreasing similarity threshold. For this reason, we prefer to use the terms 'non-overlapping' and 'overlapping' to describe the type of clusters which are obtained. In our terminology therefore, we obtain a hierarchy of clusterings for all values of k by applying the ultrametric similarities contained in U to the cluster recognition algorithm in order of decreasing similarity.

In the computer program KDEND, the ordering of ultrametric similarities is stored in Q at step 2 of Method II. Any value Q_p , corresponding to an entry U_{ij} , which is not filled, is retained in Q for subsequent use as a threshold with the cluster recognition algorithm.

Large Populations

In a situation where we wish to describe a large population in terms of a few 'types', we tend to look for large clusters that signal a concurrence of pattern. These large groups need not be the most interesting - the peripheral and intermediate objects may attract our attention through being unusual - nevertheless, we must first detect and isolate the types before the classification

can be examined in detail. Earlier we introduced the concept of a chain cluster for the k -partition, which consists of a straggle of maximal complete subgraphs that intersect in at least k vertices. We shall restrict our interest in the clusters of a large population to the chain clusters, or those distinct maximal complete subgraphs that at least have the potential to chain if there were any local connections. We observe that each individual maximal complete subgraph (which we shall call a 'unit') must possess at least $k+1$ mutually connected objects. Furthermore, since each unit is maximally connected at threshold r its diameter cannot exceed r . Hence the units can be considered as spherical dense neighbourhoods, containing at least $k+1$ objects, and clusters are formed by connecting intersecting units (provided that they intersect in at least k vertices). This is essentially the concept on which the probabilistic method is based, and it is interesting to note that both methods degenerate to single linkage when $k = 1$. Of course, there are differences: two intersecting chain clusters are separated by the k -partition if they fail to intersect in more than $k-1$ objects, and the k -partition also recognises maximal complete subgraphs which contain fewer than $k-1$ objects. Nevertheless, with a large population which exhibits distinct data swarms we can expect the two methods to yield very similar major clusters (that is, if it were possible to use the k -partition method with a large population).

These considerations now suggest an algorithm for the specific solution of large clusters by the k-partition when a single initial threshold value is given. We observe that every member of a unit is connected to at least k other objects, and conversely any object which has less than k connections at threshold h cannot belong to a large cluster. We can therefore eliminate some of the objects which are not members of units at threshold h by removing those objects having less than k connections. Furthermore, since the connection counts within the residual population may be modified by these removals, we now recompute the connection counts, reapply the test for k connections and repeat this procedure until there are no further removals. The residual population now contains all the members of units, and possibly some others. We now perform the Method II ultrametric reduction on the similarity submatrix for the residual population, and use the cluster recognition algorithm at threshold h to identify the large cluster members. This algorithm will evidently work well for any size of population, provided that the similarity threshold is sufficiently large to yield a residual population of order less than 60. However, since this restriction reduces the effective generality of the k-partition method, we have not programmed the algorithm.

9.4 DISCUSSION

Figure 9.4.1 shows the times required for Methods I and II

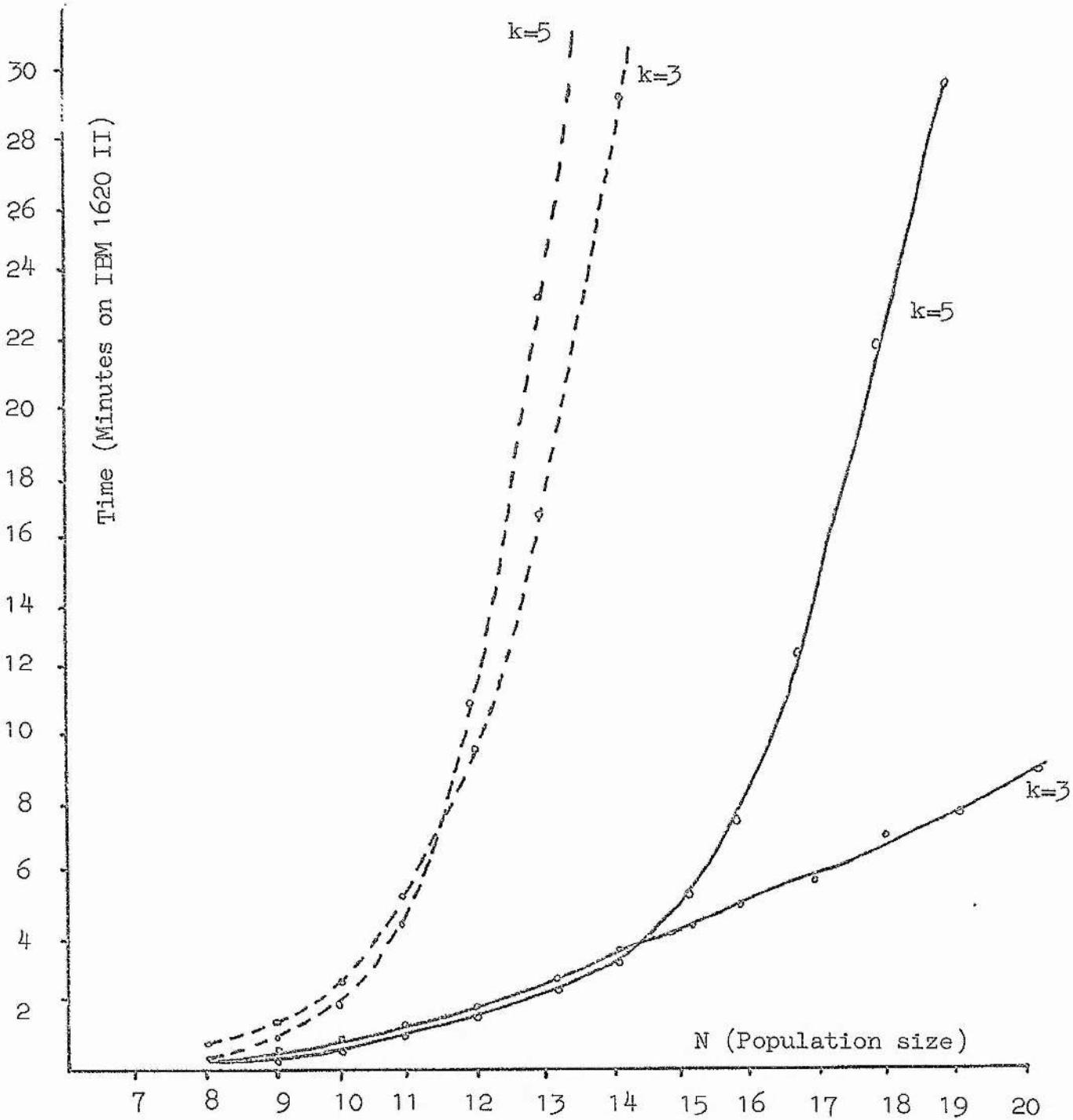


Figure 9.4.1. An indication of the times required for the two algorithms. Dotted lines correspond to Method I; solid lines are Method II.

using different population sizes. Each timing represents the average of two trials using different 2-cluster populations generated by a pseudo-normal number routine. The similarity matrix was computed from Euclidean distances, and the tests were completed by Fortran II programs on the IBM 1620 II, and Fortran IV programs on the IBM 360/Model 44. It became clear that the sequence $k = 1$ to 5 could only be reasonably computed with the IBM 360/Model 44 for up to 20 individuals by Method I and 60 individuals for Method II (these limits can probably be extended by 5 and 20 objects respectively on a faster computer). The largest population size that we tried was 35, for which Method II required 18 minutes for the completion of the sequence $k = 1$ to 5 on the IBM 360/Model 44. In view of the fact that the similarity matrix must be held in core for these algorithms, the method is necessarily restricted to small data problems.

An important practical feature of the k -partition is the large number of clusterings that are obtained. In our trials with the 9-object population cited by Jardine and Sibson, the cluster recognition algorithm produced 70 separate classifications for the sequence $l = 1$ to 5, and although not all of these groupings were unique, a user must be severely selective when presenting his results. This is in complete contrast to hierarchic mode analysis which incorporates a selective mechanism that seldom presents the user with more than about 12 groupings for any population size.

On the whole, we feel that the detailed analysis of data structure which is offered by the k -partition is desirable for small populations; large data applications which suggest a natural class approach should be referred to the alternative probabilistic model.

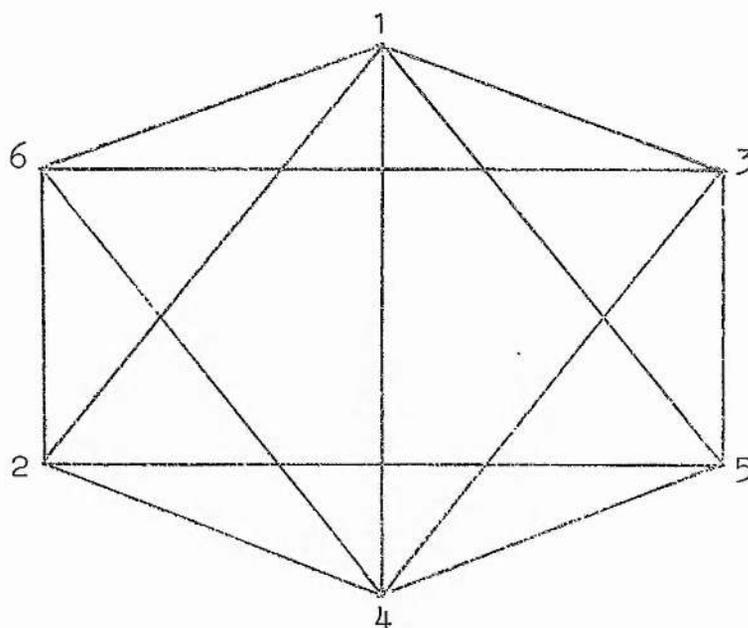


Figure 9.3.3. An example of the complete overlap situation for 6 objects when $k = 4$. The clusters are (1245) , (1246) , (1345) and (1346) .

CHAPTER 10: COMPUTER TECHNIQUES

10.1 TOWARDS GENERALISED STATISTICAL SYSTEMS

There appear to be the following five general types of statistical systems: isolated special-purpose programs, program packages, subroutine systems, conversational mode packages and statistical languages. Chambers (1967) has reviewed some of the advantages and disadvantages of these systems taken in isolation, but it appears that the designs involved can be generalised, if only in a limited way, to that of a system of subroutines. First, we shall review the pro's and con's of each system, taken in isolation.

Special-purpose packages are often inflexible due to rigid input/output conventions and the not infrequent specialisation of the algorithms. For example, many of the single linkage programs which are passed around are restricted to either continuous data or binary data, according to the particular requirements of the authors; seldom do there exist facilities for handling different similarity criteria, allowing for missing observations or drawing dendrograms. In short, such programs are suitable for casual users who wish to repeat, in every respect, the analyses of someone else. The single advantage of such programs is that they are usually trivial to run, often because they require no job control specifications for data sets.

Subroutine systems are most useful at large installations staffed by skilful statistician-programmers. The designer attempts to reduce all the complex statistical procedures that he can think of to a series of simple operations which can then be efficiently programmed as general-purpose subroutines. To rebuild a complex procedure, the programmer then simply links together the required subroutines and appropriate input/output instructions within a single main program. The advantage of this approach is that the user is free to programme a unique task without concerning himself with all the trivial repetitive computations packaged in the subroutines. The disadvantage is that programming is required, and this does not suit the biologist who has little or no experience with computers; furthermore, trivial syntactic mistakes in the short main program can easily be made with the corresponding delay in obtaining results.

Some of the better program packages represent a compromise between the single programs and subroutine systems. Having developed a subroutine system, various main programs can be written which link the subroutines in different combinations in order to perform different statistical tasks, as selected by data card parameters. This enables the non-programmer to 'programme' a complex task using a few simple data card values, while the programmer can still access individual subroutines as before. The package becomes unwieldy when the number of options is large;

rigidly specified data card parameters are often error-prone and tend to reduce the level of programming to that of machine-code (integer) instructions.

Statisticians were quick to realise that these systems could be elegantly brought together within a statistical language which uses 'free-format' commands taken from the English language. For example, the following typical package-type parameter card:

1 -2 -3 -4

would be replaced in the language ASCOP (Cooper, 1967) by the statement:

REGRESSION OF A ON ALL VARIABLES EXCEPT E AND F

which is not only easier to read and write down from memory, but also less error-prone. On the debit side, complex rules of syntax evolve as soon as low-level programming facilities are introduced to the language. As Colin has remarked (personal communication) the designer usually starts by writing a Fortran or Algol compiler and then extends the facilities. This means that such systems are often machine dependent (being written in assembler code), difficult to use by the non-programmer, and represent an enormously expensive investment of development effort. Furthermore, it is usually difficult to modify or extend the standard language features.

Conversational mode packages could be classified as special

types of statistical languages, the difference being that with the languages the user is required to specify his instructions in advance of processing, while the conversational mode program interrogates the user through an on-line terminal in order to set up the required program (Colin, 1967). The advantage of this system to the non-programmer is undeniable, but there comes a time when even the most amateur computer user graduates from the question-and-answer routine to the stage when he can specify in advance (programme) what his requirements are to be. This system is ideal for the very casual user, and as a teaching instrument for beginners. The essential design of the package is often the same as the language, and it is sometimes the case that both batch-processing language and conversation mode terminal facilities are provided with the same system (Cole and Campbell, 1969).

It is now apparent that all of the above systems can be obtained from the subroutine library system, provided that we can define a suitable method for communicating data from one phase to another. This is necessary because we must be able to define a procedure as a series of independent programs which can be both interrupted and restarted. We observe, firstly, that some data items are too big to be stored in core - for example, the triangular matrix of $\frac{1}{2}N(N-1)$ similarity coefficients (Chapter 1) becomes unwieldy when N exceeds about 80. Since magnetic tape is

restricted to sequential operations, the use of a direct-access device such as magnetic disk is indicated. However, it must be remembered that although this permits the extension of problem size beyond the capacity of core storage, we should not equally suffer the penalty of using slow memory with small problems which could otherwise be solved in core alone. Ideally we should be able to utilise all of the available core for primary storage, switching to secondary disk storage when demanded by the problem size (Sect. 10.2).

It is now suggested that most statistical requirements can be met with the FORTRAN subroutine system, building this into the various other systems as follows:

Special-purpose programs can be written as a sequence of subroutine calls using additional programming where necessary.

Program packages can be set up using comprehensive driving programs to link the basic subroutines. A complex analysis may then be assembled as a sequence of independent executions, where data are stored on disk by one program and read from disk by another.

A statistical language can be defined using syntax which is translated into either a sequence of FORTRAN subroutine calls, or a sequence of job control statements which link independent programs from the package. Communication of data would again be via the disk-based file, and the language compiler would therefore take the form of a driver program which generates a single Fortran

program, or a sequence of job control statements which invoke serially the package programs together with appropriate parameter and data cards. This latter design has the advantage that an entire Fortran program can be easily introduced to a statistical procedure; all that the driver program need do is set up the job control statements which are required to enter the Fortran compiler.

A conversational mode package could now be written as a driver program which translates user responses into either the syntax of the statistical language, or a sequence of subroutine calls, or a sequence of job control statements.

These suggestions are neither new nor relevant only to statistical systems. The idea of designing a compiler as a series of independent 'modules' which are phased in and out of core as required is already well-formulated, and the translation of high-level commands into job control statements has a parallel in catalogued procedures. Indeed, since there exist catalogued procedures for cataloguing procedures, so too could there exist a language command for defining new elements of the language. However, these are only ideas and they would require some exploratory work in order to test whether implementation is both possible and reasonable.

Basic system requirements

We shall now list several requirements of the system which

are deemed a priori to be necessary:

- (a) the system should be written using a high-level computer language in order to have maximum potential distribution market. However, allowance should be made within the design to replace deeply nested routines with fast assembler code programs. At the present time, the language Fortran IV is indicated.
- (b) The design of subroutines and programs should anticipate any modifications which may be required to 'tailor' the system for a particular machine (i.e. for varying core allotments, peripheral and other installation dependent characteristics).
- (c) Comprehensive error detection routines should be provided to trap data and parameter errors; all messages should be clearly stated (in English), and ideally, it should be possible to read and understand system output without referring to a manual.
- (d) Adequate documentation should be provided for both users and programmers. In particular, enough information should be given to enable qualified programmers to write additional material for the system, and in view of this prospect, all programs should be designed for maximum machine independence.

From these requirements has emerged a suite of programs called CLUSTAN (CLUSTER ANALYSIS), which was completed in two stages: CLUSTAN I (September, 1968) and CLUSTAN IA (November, 1969). The programs were initially written in Fortran II for the IBM 1620 (Wishart, 1969d) and then alternative Fortran IV routines were

developed for the IBM 360/Model 44. The initial development of Fortran II routines proved to be extremely advantageous, because once the input/output statements had been translated into Fortran IV read/write commands it was found that basic Fortran II satisfies the Fortran IV compilers of all machines on which CLUSTAN has been implemented.

Other characteristics were determined by the subject, and by user requirements. Firstly, the dimension which seems to be of most interest is the maximum population size (N) which can be accommodated. It seems that no matter how large a maximum population is achieved there will always be people who have larger data sets, and successive versions of CLUSTAN had maximum N of 250, 400 and 999. In the design of program specifications it was often necessary to balance the multiplicity of facilities which are to be provided against the simplicity of input descriptions. Every effort was made to reduce the number of input parameters for each program while retaining reasonable generality. Where possible, default parameter options are employed; that is, many parameters, if blank, are allocated standards by the program. Finally, the most variable machine-dependent characteristic is core size, and in order to make optimum use of the package on different machines it was important that the programs were designed for easy modification and efficient use of core. Careful attention was therefore paid to block design of programs so that 'OVERLAY' or 'LOCAL'

structures can be employed. Also, each program uses the same variable and array names throughout its 'personal' subroutines (all excepting general-purpose subroutines), and in almost all cases the DIMENSION and COMMON statements (within the subroutines of one program) are identical. This means that to reduce core storage requirements, a programmer need only prepare one new DIMENSION and COMMON block for each program, duplicate these cards and insert a copy into all of the 'personal' subroutines.

10.2 DATA STORAGE AND RETRIEVAL TECHNIQUES

Following requirements (a) and (b) of the system, it is natural that all direct-access input/output operations should be confined to one subroutine (called DISKIO). This subroutine operates on a 'data set' which is stored on the direct-access device, and can best be described as a 2-dimensional array such that each row constitutes one 'record' and is addressed by a record number. Most advanced computer operating system permit programmers to choose the record length (the number of words in one row of the data set), but of course there is usually an optimum length which makes the most efficient use of storage and software. With the IBM 360/Model 44 Programming System, this optimum is 90 words if each word occupies the standard 4 bytes, because this particular operating system uses a 360 byte buffer for I/O transfers. Some systems do not permit different record

lengths for data sets: such is the case with the KDF9 and IBM 1620 computers, for which the only permitted record lengths are 640 and 10 words respectively. In view of such limitations, it is necessary to describe the data set used within a general-purpose system such as CLUSTAN in two different ways, as follows:

The INTERNAL FILE is the CLUSTAN data set having dimensions (LMAX,10) - that is, containing LMAX records capable of storing up to 10 real or integer words. All CLUSTAN programs reference this fixed-length data set via subroutine DISKIO.

The OBJECT FILE is a machine-dependent data set, having dimensions (NREC,LREC*10), where LREC*10 is the most efficient record length for a particular machine and operating system (LREC is an integer such that $NREC * LREC \geq IMAX$). Therefore, one record of the object file contains LREC records of the CLUSTAN internal file.

Subroutine DISKIO must therefore link these two data sets, and it is the task of each installation programmer to modify DISKIO to suit his particular optimum direct-access specifications. One example of DISKIO (for the IBM 360/44) is given in Appendix II, where the programmer must change three cards in order to alter the record length of the object file from 90.

During the period September 1968 - November 1969, the following eight versions of DISKIO were developed:

1. Fortran II Fetch/Record I/O's for the 1620..
2. Fast SPS (assembler code) version for the 1620.

3. KDF9 version (using the KDF9 software GET ARRAY/PUT ARRAY).
4. Fortran IV paging version.
- *5. ICL 1909 version.
- *6. English Electric System 4/50 assembler code.
7. Fortran IV paging version with partial incore file-simulation (see Appendix II).
8. Fortran IV paging version using dynamic file-simulation in core.

Versions 1 and 2 are written specifically for the IBM 1620 (both are published in the CLUSTAN I manual; Wishart 1969d). In this instance, the 1620 object file exactly matches the CLUSTAN internal file (not accidentally), both having 10 words per record. Consequently, both of these DISKIO versions process the same record addresses as the internal file, and no address modification is required within the subroutine. It is interesting to note that the SPS version reduces execution times to roughly 1/3 of those recorded with version 1. This is because the IBM-supplied direct-access software for the 1620 contains transient routines which are phased in and out of core, except in very special circumstances. Since it is impossible to arrange that the object file and these transient routines occupy different disks, every I/O operation requires two disk seeks and three disk reads. Consequently, about two thirds of

*The ICL 1909 version was written by Atma Trasi of the Bradford University Computing Laboratory, and the 4/50 assembler code version was written by Nick Tyrer, Greater London Council.

the execution time using DISKIO version 1 can be attributed solely to disk head-contention.

Paging version

Versions 3-6 use a paging system, whereby DISKIO holds one page (one record of the object file, containing LREC records of the internal file) in core at any one time. If a transfer call of DISKIO references records solely located on this incore page, then the transfer is made directly to core (no I/O occurs). Any subroutine call which references a page other than the incore page initiates a page swop operation within DISKIO: the current incore page is transferred to disk only if it has been changed while in core, and then the new page is read into core. For the purpose of testing whether the incore page has been changed while in core, a status integer IWR is updated from 1 (read only) to 2 when any write call of DISKIO is encountered. Hence a segment of program which is reading sequentially from the internal file will require only sequential read transfers from the object file.

The first attempt to improve the paging system was to increase LREC as much as possible in order to reduce the number of direct-access I/O operations. This had disastrous effects¹ during such operations as the computation of principal component scores for certain sizes of continuous data files. The

¹ Similar problems have been encountered with virtual memory machines, and with paging schemes in general.

operation requires that one vector of m values be read from the raw data file (LNDATA), transformed to m' factor scores and then written to the factor scores file (LSCORS), this sequence being repeated N times. If the internal file records associated with the raw data and the factor scores vectors are located on different object file pages then the program performs $2N$ read and N write page transfers for the full calculation. Naturally, when a page size of 5400 words is used the direct-access I/O time becomes prohibitive, particularly with an operating system such as 44PS which subdivides each page transfer into separate I/O's of 90 words (hence a single 5400 word page transfer uses 60 I/O's). This large-page scheme was therefore abandoned in favour of the following modified small-page version.

Incore file simulation

An array A is chosen so that it fills all of the available core when DISKIO is linkage edited with the largest program. Array A is logically subdivided into NBLOCK object file pages (each comprises LREC*10 words). These pages are read into core from disk when they are first required, and reside in core throughout the duration of a CLUSTAN program execution. Associated with each incore page I is a read/write status key IWRD(I) which takes the values 0, 1 or 2, according to whether page I has -

- (0) not been read into A,
- (1) been read into A and received only read transfers,
- (2) been read into A and received one or more write transfers.

Those programs which write to disk all finish with an ending call of DISKIO which causes those pages for which $IWRD(I) = 2$ to be written out to the object file. This effects all the internal file updates which have occurred during the program execution.

If the internal file extends beyond the limits of the part of the object file which is simulated in array A, then pages numbered higher than NBLOCK are swopped in and out of a single buffer page (called DISK) as for versions 3-6 of DISKIO. The extent of array A, which determines the amount of incore simulation of the internal file that can be achieved within DISKIO, can be varied to suit different core partitions by changing two statements within the subroutine (see Appendix II).

In order that this rather sophisticated technique for data storage and retrieval may be used effectively, it is necessary to organise the data so that the part of the internal file which is simulated in array A contains those items which are most required by the commonly used programs. Table 10.2.1 shows the layout of the present CLUSTAN internal file; it should be noted that the ordering is purely accidental, having been determined by the sequence of calculations as they are programmed within FILE and CORREL. Fortunately, the raw data are stored at the top of the

| <u>RECORD ADDRESSES</u> | <u>CONTENTS</u> |
|-------------------------|--|
| 1-2 | 11 file parameters N, MB, MN, etc. |
| 3-5 | 15 file addresses LNDATA, LBDATA, etc. |
| 6-7 | Data identification title (array TEXT). |
| *LNDATA | Start of continuous data file. |
| *LBDATA | Start of binary data file: |
| I MEANS | Continuous variable means. |
| I VARS | Continuous variable variances. |
| I CORS | Matrix of continuous variable product-moment correlation coefficients. |
| I EIGS | Eigenvalues. |
| I EIGVS | Eigenvectors. |
| I SCORS | Start of principal components scores file. |
| I LENG | Binary sample list lengths. |
| I FREQS | Binary attribute frequencies. |
| I NMASK | Continuous variable masking array. |
| I BMASK | Binary attribute masking array. |
| *I MAT | Start of similarity matrix. |
| I KLIST | K-linkage lists (nearest neighbours). |
| I NEXT | End of file pointer. |

Table 10.2.1. CLUSTAN internal file format. Zero record addresses indicate that the associated data are not stored. *Denotes data which are scanned within deeply nested program segments.

internal file, so that if the distributed version of DISKIO (Appendix II) is adhered to (array A contains 5400 words), then the data matrix will usually be stored completely in core. This feature turns out to be extremely advantageous, particularly with programs such as DIVIDE and RELOC which do not use a similarity matrix. It

means that the data are read into core by subroutine DISKIO only once (when first required); thereafter, iterative scanning of the population to divide a cluster or improve a classification will require no additional direct-access I/O operations.

DISKIO VERSION 4

| | <u>N=30</u> | <u>N=60</u> | <u>N=90</u> | <u>N=120</u> | <u>TOTALS</u> |
|---------------|-------------|-------------|-------------|--------------|---------------|
| FILE | .18 | .28 | .36 | .44 | 2.06 |
| CORREL | .21 | .39 | 1.00 | 1.34 | 3.34 |
| MODE | .23 | .49 | 1.42 | 3.07 | 6.01 |
| HIERAR | .34 | 1.28 | 3.00 | 5.19 | 10.21 |
| RESULT | .39 | .58 | 1.18 | 1.38 | 4.33 |
| <u>TOTALS</u> | 2.15 | 4.22 | 7.36 | 12.22 | 29.59 |

DISKIO VERSION 7

| | | | | | |
|---------------|------|------|------|------|-------|
| FILE | .16 | .23 | .29 | .33 | 1.41 |
| CORREL | .20 | .32 | .46 | 1.05 | 2.42 |
| MODE | .21 | .30 | 1.04 | 2.39 | 4.34 |
| HIERAR | .25 | .45 | 1.27 | 3.41 | 6.28 |
| RESULT | .40 | .59 | 1.19 | 1.39 | 4.37 |
| <u>TOTALS</u> | 2.02 | 3.09 | 5.05 | 9.36 | 23.07 |

Table 10.2.2. Execution times for five of the CLUSTAN programs using versions 4 and 7 of subroutine DISKIO on the IBM 360/Model 44. In each case, populations of size 30, 60, 90 and 120 were tested, and the figures represent an overall reduction of 23% on the execution times of the paging version 4 of subroutine DISKIO.

Table 10.2.2 contains 360/44 execution times (minutes. seconds) recorded using versions 4 and 7 of subroutine DISKIO with

data sets of 10 continuous variables for populations of size 30, 60, 90 and 120. The data files for populations N=90 and 120 exceeded the simulated part of the internal file (540 records) to the extent of 198 and 615 records respectively. Nevertheless, in all cases except RESULT the partial incore file-simulating version of DISKIO proved superior, and the figures represent an average reduction of 23% on the paging version execution times.

| | <u>RELOC</u> | <u>DIVIDE</u> |
|--|--------------|---------------|
| (4) <u>Paging version of DISKIO</u> | 7.23 | 4.26 |
| (7) <u>Partial file-simulating version of DISKIO</u> | 6.23 | 2.52 |
| <u>Time reduction factor</u> | 14% | 35% |

Table 10.2.3. Comparison of programs RELOC and DIVIDE using a population of size 450 with versions 4 and 7 of subroutine DISKIO. The data are binary (Crawford et al, 1970) and in both cases the error sum of squares was optimised.

Table 10.2.3 shows a more dramatic comparison between the two versions. Execution times for programs RELOC and DIVIDE are given for an analysis of the 450 quadrat x 37 species Andean survey data (Crawford et al, 1970 - see Appendix Ia). In this instance, the internal file occupies 458 records, and is therefore fully simulated in core by version 7 of DISKIO. Consequently, by carefully choosing the repetitive iterative relocation optimisation (from a random start), and a monothetic divisive analysis (error sum of squares), the amount of time spent waiting for the direct-access

device with DISKIO version 4 is demonstrated (particularly in the case of DIVIDE).

Dynamic file-simulation

The last stage in the improvement of DISKIO (version 8) was to make use of all available core for file simulation. This is achieved on the IBM 360/Model 44 at St. Andrews using two assembler code functions which pass to DISKIO the core addresses associated with the end of the problem program and the end of core. This free space is then utilised for file-simulation within array A, and it is evident that small programs will therefore simulate bigger areas of the file than the large programs. In particular, if this scheme could be adapted to a multi-programming system, the user would be able to select a bigger area of core for file-simulation by specifying a larger partition; the result would be that the selection of a large partition (slow turnaround) would reduce the program wait time (time spent in core waiting for input/output transfers to be completed), while a smaller partition would improve the turnaround at the expense of wait time. The user can therefore balance efficiency of computation against speed of turnaround.

Summary

The greatest advantage provided by DISKIO is that any address modification required to associate data on the internal file with the same data on a particular object file is carried out within the subroutine. Therefore, once a modified version of DISKIO has been

designed for a particular machine, and the job control statements required for the object file have been defined, then all CLUSTAN programs immediately fit the system. All programs are oriented to the internal file, and are therefore machine independent. Consequently, any new program which is distributed will immediately fit at all installations (provided that no dimension modification is required to squeeze it into a smaller memory).

Of secondary importance is that DISKIO frees the user from the limitations of core storage. The paging versions (3-6) could be criticised for using direct-access operations with every problem: all populations whether large or small use the direct-access device for work space, and therefore the programs are unnecessarily disk-bounded for small problems. Versions 7 and 8, on the other hand, can be tailored to make full use of the available core to simulate all or the most-used part of the object file in core. Hence, small problems will be solved in core, while the number of direct-access I/O's for large problems is reduced.

10.3 FEATURES OF CLUSTAN

The suite of computer programs described here was developed to enable different methods of cluster analysis to be compared for large data matrices. Magnetic disk was used as secondary storage right from the start, and a significant proportion of the work was concerned with improving direct-access data retrieval techniques, this culminating in subroutine DISKIO (Sect. 10.2).

Another design feature was the organisation of independent programs to perform common tasks such as data input, computation of similarity matrices, and so on. It was considered important that new techniques should be easily introduced to the package with the minimum of repetitive programming. This led naturally to the subroutine system (Sect. 10.1), although a driver program which links subroutines according to crude language syntax has not been written for general distribution¹. The restrictions of the package are as follows:

Maximum number of individuals = 999

Maximum number of binary attributes = 400

Maximum number of continuous variables = 200

No missing data permitted.

Essential peripherals: Card reader, line printer, magnetic disk.

Desirable peripherals: Magnetic tape, card punch, graph plotter.

All of the binary and intercluster similarity criteria discussed in Sections 1.4 and 5.3 are incorporated within the package, together with the USER facility mentioned in Sect. 5.3. Each of the programs contained in the distributed version of CLUSTAN (release IA) is now briefly described.

¹At the time of writing, such a program has been conceived, if not implemented, for the Atlas computer.

FILE is used to create the data file for either binary or continuous data, or both. Simple transformations such as standardisation and rotation to principal coordinates are built in, as are optional product-moment correlation coefficients. Initial input transformations can be completed by the user within a subroutine called READ, so that nonstandard normalising or data generating methods may be used. This latter facility is particularly useful for treating existing data decks, and for generating data such as the Gaussian distributions of Chapter 7.

CORREL computes a similarity matrix and the ordered lists of the k 'nearest neighbours' for each individual. The data medium can be one of the following:

Raw continuous data

Standard scores

Principal coordinates

Binary data

All of the similarity coefficients discussed in Section 1.4 are available, and facilities exist which enable selected variables to be masked if desired. The USER function (Sect. 5.3) may be reprogrammed to evaluate a new similarity measure.

MODE clusters by the probabilistic mode-seeking algorithm (Chapter 6) using both the k th greatest similarity and average of the $2k$

greatest similarities as density estimates (the latter is the default option).

HIERAR uses the combinatorial algorithm of Chapter 8 to evaluate all of these hierarchical procedures. Any similarity matrix may be given any combinatorial transformation, but if the two are incompatible (for example, the transformation for Ward's method used with a nondistance measure) then the program prints a warning message and continues processing, thereby enabling experienced users to test new solutions.

CENTRO evaluates 'centroid sorting' by the usual centroid-forming technique, and is therefore capable of treating those similarity criteria which cannot be given a combinatorial solution with HIERAR. The USER function may also be linked with CENTRO for hierarchical centroid sorting to optimise a new similarity measure.

DIVIDE performs monothetic division, having as options all variants of Association analysis, group analysis and the generalised divisive technique which optimises any similarity measure (Chapter 3). Either nested or hierarchic subdivision may be adopted, and the program contains an optional trace facility to provide the user with the alternatives of minimal or complete information. Function USER may be linked with DIVIDE for the

evaluation of a new similarity measure under monothetic division.

RELOC uses the generalised iterative relocation procedure (Chapter 4) to optimise any of the standard similarity measures, or a new measure programmed in function USER. Facilities exist for the selection of a similarity threshold value to create a residue if desired, to eliminate clusters whose size reduces below a specified size limit, and to select the non-removal test (4.1.2) if preferred (the default test option is 4.1.1).

KDEND evaluates the Jardine-Sibson k-partition method for finding overlapping clusters by the faster Cole-Wishart algorithm (Chapter 9). The cluster recognition algorithm is normally entered, but this can be optionally suppressed.

DNDRIT uses the Calinski-Harabasz method (Sect. 3.2) to optimise the error sum of squares with a partition of the minimum spanning tree. Alternatively, the within-group sum of similarities may be chosen for maximisation by the same technique.

RESULT is used to print any of the data file values (Sect. 10.2) stored by FILE and CORREL, and also to compute some simple cluster diagnostic statistics using as input any selected classifications obtained from the clustering programs.

SCAT plots scatter and cluster diagrams on the graph plotter, taking as coordinates any mixture of raw or standardised variables, or principal components. Typical examples of SCAT output are to be found in Chapter 7 and Appendix Ie.

PLINK plots full or partial dendrograms on the graph plotter, for any result of programs HIERAR, CENTRO or DIVIDE.

Programs STORE and RESTART are supplied for the tasks of copying the data file from disk on to magnetic tape for permanent storage, or vice-versa.

Discussion

A typical CLUSTAN job often consists of several different methods of cluster analysis which are 'programmed' by linking together appropriate programs. For example, the following sequence would evaluate Association analysis, group analysis, Ward's method and Information-analysis, calling RESULT to compute cluster diagnostics for specified classifications.

FILE

| | | |
|---------------|---|------------------------------|
| <u>DIVIDE</u> | - | Association analysis |
| | - | Group analysis |
| <u>CORREL</u> | - | Distance matrix |
| <u>HIERAR</u> | - | Ward's method |
| <u>CORREL</u> | - | Information statistic matrix |
| <u>CENTRO</u> | - | Information analysis |

STORE

RESULT - Cluster diagnostics

The order is not strictly important, provided that CORREL logically precedes HIERAR and CENTRO for the computation of an appropriate similarity matrix, and that RESULT succeeds those programs for which classifications are to be investigated. An alternative sequence to achieve the same analyses might be: FILE, CORREL, CENTRO, CORREL, DIVIDE, HIERAR, RESULT, STORE. The inclusion of program STORE ensures that the data file is copied on to a magnetic tape for permanent storage. The continuing analysis of these data might therefore be achieved with the sequence: RESTART, SCAT, RELOC, PLINK, which would plot scatter diagrams and dendrograms, and improve some of the earlier classifications using iterative relocation.

The extension of this package to provide a wider range of multivariate procedures is now indicated, and the development of driver programs to translate simple task specifications into the appropriate control statements for sequences, such as the above examples, is suggested. In particular, the format for the internal data file should be generalised to enable programmers to store extra file parameters and file addresses, and thus store intermediate results. The present list of installations which are using the CLUSTAN package is given in Appendix III, and it is hoped to stimulate the development and exchange of new routines at some of these centres.

CONCLUSIONS

It has been shown, with a fairly comprehensive survey of the literature (Chapters 1-4), that many independently proposed techniques are variants of a tripartite grouping procedure associated with a generalised similarity function $S(p,q)$. Twelve different definitions of $S(p,q)$ are compared with test populations and shown to behave in very different ways (figure 7.1.2), suggesting that clusters are influenced far more by the choice of similarity criterion than by the choice of method. Such methods are referred to as 'minimum variance' techniques because they are based on the principle that each member of a cluster should be very similar to every other member. If we draw a graph such that each vertex represents an individual and the distance between pairs of individuals relates to their similarity, then good minimum variance clusters will appear on this graph as tight spherical arrangements of the vertices. When a criterion such as E or d^2 is used to measure similarity, the graph theoretic model maps directly into euclidean space: that is, tight clusters on the graph correspond to tight clusters in euclidean space. On the other hand, if a measure such as size difference is chosen, tight clusters on the graph are associated with elongated clusters in euclidean space (figure 7.1.2). Minimum variance methods can now be tidily generalised as algorithms which find tight clusters from the graph theoretic model, defined by a similarity criterion approved by the user. However, this implies that the user is well

aware of the properties of the chosen criterion S , and has established that these properties match the desired characteristics of the clusters. It has been shown (Chapter 7) that while the properties of some obvious definitions of S (such as E and d^2) are invariant under changes in origin and scale, real clusters are not invariant under such changes. Alternatively, measures which vary directionally with origin and scale are unlikely to match the internal variation within all clusters (figure 7.1.7), since for a particular S , specific directions of variation are predefined in relation to the coordinate axes and origin. It is therefore suggested that algorithms developed from the topological representation of the population constitute dangerous mathematical interpretations of statistical data which are invalidated by the absence of a sound definition of what is meant by a cluster.

The alternative 'natural class' concept is much more like statistics, although rigorous statistical procedures have yet to be worked out. Given spherical clusters in euclidean space (figure 7.1.1), standard transformations have been shown to stretch and squeeze the swarms so that they are no longer necessarily spherical, but may be elongated (figure 7.1.7); however, the swarms are still disjoint. It seems reasonable, therefore, that a 'cluster' should be regarded as a swarm of points (of any shape) which should be separated from other such swarms. This idea is further extended to the probabilistic model, where clusters are

defined by the underlying probability density function $F(\underline{u})$ as those closed connected volumes at level p given by $F(\underline{u}) \geq p$.

The generalisation of single linkage to hierarchical mode analysis constitutes a major step towards finding such clusters, but it is by no means the final solution. Indeed, the problem of resolving stable natural classes is very likely to be one of the major tasks of numerical taxonomists for several years to come. Some ideas for improvements to mode analysis have been given, of which perhaps the most appealing is the suggestion that a modified contingency table technique could be coupled with a method which uses the sample distribution to accurately estimate theoretical densities by iterative smoothing.

Although natural classes appear to be more difficult to find, and have more general applicability, there will always be occasions when the minimum variance solution is indicated. In particular, the single multivariate normal sample will often have to be split into tight clusters by some arbitrary partition, in order that the population may be adequately described. For this purpose, it is suggested that the optimisation of E or d^2 by iterative relocation, using Ward's method to provide the initial solution and a supplementary hierarchical classification, is appropriate. Some work still has to be done on the selection of a good starting solution, and it would also be nice to modify the

technique to trap 'local' optima, so that by "sliding the final partitions back and forth" (Forgéy, 1965) convergence to a global solution may be more assured.

On the subject of statistical computing, there is clearly a need for more effort to be directed towards the development of a standardised statistical system. It has been suggested (Chapter 10) that a general system of subroutines, if properly designed, might meet all statistical needs from batch processing programs to statistical languages and conversational mode packages. The key to this lies in developing powerful data handling and storage facilities, for which paging systems such as DISKIO (Sect. 10.2) may be the answer. Some statistical procedures must be generalised, and new algorithms formulated, in order that they may be expressed in terms of sequences of simple computational operations. In this respect, the tripartite grouping procedure with generalised $S(p,q)$ represents a major step in the standardisation of the minimum variance methods of cluster analysis, and the CLUSTAN suite of computer programs provides researchers with a useful basic tool for the future evolution of programmed techniques and their collective evaluation.

Bibliography

- Ball, G. H. (1965), 'Data analysis in the social sciences: What about the details?', Proc. Fall Joint Computer Conference, Stanford Research Inst., Menlo Park, California.
- Ball, G. H. (1966), 'A comparison of some cluster seeking techniques', Stanford Res. Inst., California.
- Ball, G. H., and Hall, D. J. (1966), 'ISODATA, an iterative method of multivariate data analysis and pattern classification', Proc. IEEE International Communications Confr., p. 116-117.
- Beale, E. M. L. (1969), 'Euclidean cluster analysis', Proc. Int. Stat. Inst. (London), p. 99-101.
- Berry, B. L. J. (1961), 'A method for deriving multi-factor uniform regions', Przegląd geogr., v. 33, p. 263-279.
- Berry, B. L. J. (1965), 'The mathematics of economic regionalisation', Proc. 4th General Meeting, Commn Methods Econ. Regionalisation, Int. geogr. Un., p. 77-106.
- Bolshev, L. N. (1969), 'Cluster analysis', Proc. Int. Stat. Inst. (London), p. B0. 33 1-15.
- Bonner, R. E. (1964), 'On some clustering techniques', IBM J. Res. Devlpmt., v. 8, p. 22.
- Boyce, A. J. (1969), 'Mapping diversity: A comparative study of some numerical methods', In - Numerical Taxonomy, Academic Press, London, p. 1-31.
- Brillouin, L. (1962), Science and information theory. (2nd ed.), Academic Press, N.Y.
- Calinski, T. (1969), 'On the application of cluster analysis to experimental results', Proc. Int. Stat. Inst. (London), p. 108.
- Calinski, T., and Harabasz, J. (1970), 'A dendrite method for cluster analysis', Biometrics (in press).
- Carmichael, J. W., George, J. A., and Julius, R. S. (1968), 'Finding natural clusters', Syst. Zool., v. 17, p. 144-150.
- Chambers, J. M. (1967), 'Some general aspects of statistical computing', Appl. Statist., v. 16, p. 124-132.

- Cole, A. J. (1966), 'Plane and stereographic projections of convex polyhedra from minimal information', *Comp. J.*, v. 9, p. 27-31.
- Cole, A. J., and Adamson, P. G. (1969), 'A simple method for drawing molecules using a digital plotter', *Acta Cryst.*, v. A25, p. 535-538.
- Cole, A. J., and Campbell, R. M. (1969), 'Yet another conversational mode program', *Appl. Statist.*, v. 18, p. 190-191.
- Cole, A. J., and Wishart, D. (1970), 'An improved algorithm for the Jardine-Sibson method of generating overlapping clusters', *Comp. J.* (in press).
- Colin, A. J. T. (1967), 'On-line systems in statistics', *Appl. Statist.* v. 16, p. 111-119.
- Cooley, W. W., and Lohnes, P. R. (1964), Multivariate procedures for the behavioural sciences, Wiley, New York and London.
- Cooper, B. E. (1967), 'ASCOP - A statistical computing procedure', *Appl. Statist.*, v. 16, p. 100-110.
- Crawford, R. M. M. (1969), 'The use of graphical method in classification', In - Numerical Taxonomy, Academic Press, London, p. 32-41
- Crawford, R. M. M., and Wishart, D. (1966), 'A multivariate analysis of the development of dune slack vegetation in relation to coastal accretion at Tentsmuir, Fife', *J. Ecol.*, v. 54, p. 729-743.
- Crawford, R. M. M. and Wishart, D. (1967), 'A rapid multivariate method for the detection and classification of groups of ecologically related species', *J. Ecol.*, v. 55, p. 505-24.
- Crawford, R. M. M. and Wishart, D. (1968), 'A rapid classification and ordination method and its application to vegetation mapping', *J. Ecol.*, v. 56, p. 385-404.
- Crawford, R. M. M., Wishart, D., and Campbell, R. M. (1970), 'A numerical analysis of high altitude scrub vegetation in relation to soil erosion in the eastern cordillera of Peru', *J. Ecol.*, v. 58, p. 173-191.
- Dagnelie, P. (1967), 'Introduction aux problemes et aux methodes de classification numerique', Read at - Societe Adolphe Quetelet (Bruxelles); also in - *Biometrie-Praximetrie*, v. 9 (page number not known).

- Dawson, A. H. (1970), 'The changing distribution of Polish industry, 1949-65; a general picture', *Trans. Inst. Br. Geogr.* (in press).
- Day, N. E. (1970), 'Estimating the components of a mixture of normal distributions', *Biometrika*, v. 56, p. 463-474.
- Edwards, A. W. F., and Cavalli-Sforza, L. L. (1965), 'A method for cluster analysis', *Biometrics*, v. 21, p. 362-375.
- Florek, K., Lukaszewics, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951), 'Sur la liason et la division des points d'un ensemble fini', *Colloquium Math.*, v. 2, p. 282-285.
- Forgey, E. W. (1964), 'Evaluation of several methods for detecting sample mixtures from different N-dimensional populations', *Amer. Psychol. Ass., Los Angeles, California*.
- Forgey, E. W. (1965), 'Cluster analysis of multivariate data: efficiency versus interpretability of classifications', *AAAS - Biometric Soc. (WNAR), Riverside, California*.
- Gengerelli, J. A. (1963), 'A method for detecting subgroups in a population and specifying their membership', *J. Psychol.*, v. 55, p. 457.
- Gower, J. C. (1966), 'Some distance properties of latent root and vector methods used in multivariate analysis', *Biometrika*, v. 53, p. 325-358.
- Gower, J. C. (1967), 'A comparison of some methods of cluster analysis', *Biometrics*, v. 23, p. 623-637.
- Gower, J. C., and Ross, G. J. S. (1969), 'Minimum spanning trees and single linkage cluster analysis', *Appl. Statist.*, v. 18, p. 54-64.
- Hodson, F. R., Sneath, P. H. A., and Doran, J. E. (1966), 'Some experiments in the numerical analysis of archaeological data', *Biometrika*, v. 53, p. 311-324.
- Hyvärinen, L. (1962), 'Classification of qualitative data', *B.I.T.*, v. 2, p. 83.
- Jancey, R. C. (1966), 'Multidimensional group analysis', *Aust. J. Bot.*, v. 14, p. 127.

- Jardine, C. J., Jardine, N., and Sibson, R. (1967), 'The structure and construction of taxonomic hierarchies', *Math. Biosci.*, v. 11, p. 173.
- Jardine, N., and Sibson, R. (1968), 'The construction of hierarchic and non-hierarchic classifications', *Comp. J.*, v. 11, p. 177.
- Johnson, S. C. (1967), 'Hierarchical clustering schemes', *Psychometrika*, v. 32, p. 241-254.
- Jones, K. S., and Jackson, D. (1967), 'Current approaches to classification and clump-finding at the Cambridge Language Research Unit', *Comp. J.*, v. 10, p. 29-37.
- Kaiser, H. F. (1959), 'Comments on communalities and the number of factors', paper read at the symposium 'Applications of computers to psychological problems' at the American Psychological Association meeting, 1959.
- Kelly, F. (1969), 'Classification of urban areas', *GLC Research and Intelligence Unit Quarterly Bulletin*, No. 1, p. 13-19.
- Kendrick, W. B., and Proctor, J. R. (1964), 'Computer taxonomy in the fungi imperfecti', *Can. J. Bot.*, v. 42, p. 65-88.
- Kruskal, J. B. (1964), 'Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis', *Psychometrika*, v. 29, p. 1-27.
- Kullback, S. (1959), Information theory and statistics, John Wiley and Sons, N.Y.
- Lambert, J. M., and Williams, W. T. (1966), 'Multivariate methods in plant ecology. VI. Comparison of Information-analysis and Association-analysis', *J. Ecol.*, v. 54, p. 635-664.
- Lance, G. N. and Williams, W. T. (1965), 'Computer programs for monothetic classification ("Association analysis")', *Comp. J.*, v. 8, p. 246.
- Lance, G. N., and Williams, W. T. (1966a), 'A generalised sorting strategy for computer classifications', *Nature*, v. 212, p. 218.
- Lance, G. N., and Williams, W. T. (1966b), 'Computer programs for hierarchical polythetic classification ("similarity analyses")', *Comp. J.*, v. 9, p. 60.

- Lance, G. N., and Williams, W. T. (1966c), 'Computer programs for classification', Proc. 3rd Aust. Comp. Conf. (Canberra), p. 12/3/1.
- Lance, G. N., and Williams, W. T. (1967a), 'A general theory of classificatory sorting strategies. I. Hierarchical systems', Comp. J., v. 9, p. 373.
- Lance, G. N., and Williams, W. T. (1967b), 'Note on the classification of multi-level data', Comp. J., v. 9, p. 381-382.
- Lance, G. N., and Williams, W. T. (1967c), 'A general theory of classificatory sorting strategies. II. Clustering systems', Comp. J., v. 10, p. 271-276.
- Lance, G. N. and Williams, W. T. (1968), 'Note on a new information statistic classificatory program', Comp. J., v. 11, p. 195.
- MacArthur, R. H., and MacArthur, J. W. (1961), 'On bird species diversity', Ecol., v. 42, p. 594-598.
- MacNaughton-Smith, P. (1965), 'Some statistical and other numerical techniques for classifying individuals', H.M.S.O. Home Office Research report no. 6.
- MacNaughton-Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. (1964), 'Dissimilarity analysis: a new technique of herarchical subdivision', Nature, v. 202, p. 1034.
- MacQueen, J. (1967), 'Some methods for classification and analysis of multivariate observations', Proc. 5th Berkeley Symp. 1965, v. 1, p. 281-297.
- McQuitty, L. L. (1957), 'Elementary linkage analysis for isolating orthogonal and oblique types and typl relevancies', Educ. and Psychol. Msrmnt., v. 17, p. 207-229.
- McQuitty, L. L. (1961), 'Elementary factor analysis', Psychological Reports, v. 9, p. 71-78.
- McQuitty, L. L. (1966a), 'Improved hierarchical syndrome analysis of discrete and continuous data', Educ. and Psychol. Msrmnt., v. 26, p. 577-582.
- McQuitty, L. L. (1966b), 'Similarity analysis by reciprocal pairs for discrete and continuous data', Educ. and Psychol. Msrmnt, v. 26, p. 825-831.

- McQuitty, L. L. (1967a), 'A mutual development of some typological theories and pattern-analytic methods', *Educ. and Psychol. Msrmt.*, v. 27, p. 21-46.
- McQuitty, L. L. (1967b), 'A novel application of the coefficient of correlation in the isolation of both typal and dimensional constructs', *Educ. and Psychol. Msrmt.*, v. 27, p. 591-599.
- Morrison, D. F. (1967), Multivariate statistical methods, McGraw-Hill, New York and London.
- Needham, R. M. (1962), 'A method for using computers in information classification', *Proc. I.F.I.P. Congress 62*, p. 284.
- Needham, R. M. (1965a), 'Automated classification: models and problems', Mathematics and computer science in biology and medicine, Medical Research Council, London.
- Needham, R. M. (1965b), 'Computer methods for classification and grouping', The use of Computers in Anthropology, Mouton and Co., The Hague.
- Orloci, L. (1966), 'Geometric models in ecology. I. The theory and application of some ordination methods', *J. Ecol.*, v. 54, p. 193-215.
- Orloci, L. (1967a), 'Data centering: a review and evaluation with reference to component analysis', *Syst. Zool.*, v. 16, p. 208-212.
- Orloci, L. (1967b), 'An agglomerative method for classification of plant communities', *J. Ecol.*, v. 55, p. 193-205.
- Orloci, L. (1968a), 'Information analysis in phytosociology: partition, classification and prediction', *J. Theoret. Biol.*, v. 20, p. 271-284.
- Orloci, L. (1968b), 'A model for the analysis of structure in taxonomic collections', *Canadian J. of Bot.*, v. 46, p. 1093-1097.
- Orloci, L. (1968c), 'Definitions of structure in multivariate phytosociological samples', *Vegetatio*, v. 15, p. 281-291.
- Orloci, L. (1969a), 'Information analysis of structure in biological collections', *Nature*, v. 223, p. 483-484.
- Orloci, L. (1969b), 'Information theory models for hierarchic and

- non-hierarchical classifications', In - Numerical Taxonomy, Academic Press (London), p. 148-164.
- Parker-Rhodes, A. F., and Jackson, D. M. (1969), 'Automatic classification in the ecology of the higher fungi', In - Numerical Taxonomy, Academic Press. (London), p. 181-215.
- Parks, J. M. (1969a), 'Classification of Mixed Mode Data by R-Mode Factor Analysis and Q-Mode Cluster Analysis on Distance Function', In - Numerical Taxonomy, Academic Press, London.
- Parks, J. M. (1969b), 'Multivariate facies maps', Proc. of Symp. on Computer Applications in Petroleum Exploration, Computer Contr. No. 40, State Geological Survey, Kansas, USA.
- Penrose, L. S. (1954), 'Distance, size and shape', Ann. Eugenics, v. 18, p. 337-343.
- Pocock, D. C. D., and Wishart, D. (1969), 'Methods of deriving multi-factor uniform regions', Trans. Brit. Inst. of Geographers, No. 47, p. 73.
- Proctor, J. R. (1966), 'Some processes of Numerical Taxonomy in terms of distance', Syst. Zool., v. 15, p. 131-140.
- Rao, C. R. (1952), Advanced Statistical Methods in Biometric Research, John Wiley, New York.
- Ray, D. M., and Berry, B. L. J. (1965), 'Multivariate socio-economic regionalisation: a pilot study in central Canada', Regional Statistical Studies, p. 1-48.
- Rogers, D. J., and Fleming, H. (1964), 'A computer program for classifying plants. II. A numerical handling of non-numerical data', Bioscience, v. 14, p. 15.

- Rogers, D. J., and Tanimoto, T. T. (1960), 'A computer program for classifying plants', *Science*, v. 132, p. 1115.
- Roux, M. (1969), 'An algorithm to construct a particular kind of hierarchy', In - Numerical Taxonomy, Academic Press, London, p. 234-240.
- Sebestyen, G. S. (1962), 'Pattern recognition by an adaptive process of sample set construction', *IRE Trans. on Info. Theory*, v. IT-8.
- Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Techn. J.*, v. 27, p. 379-423, 623-656.
- Shepherd, M. J. (1966), 'The methods of numerical taxonomy. A program for treating 'chained' data', Unpubl. M.Sc. thesis, Univ. of Manchester Inst. of Science and Technology.
- Shepherd, M. J., and Willmott, A. J. (1968), 'Cluster analysis on the Atlas computer', *Comp. J.*, v. 11, p. 57-62.
- Sneath, P. H. A. (1957), 'The application of computers to taxonomy', *J. Gen. Microbiol.*, v. 17, p. 201.
- Sneath, P. H. A. (1966a), 'A comparison of different clustering methods as applied to randomly-spaced points', *Classification Soc. Bull.*, v. 1, p. 2-18.
- Sneath, P. H. A. (1966b), 'A method for curve seeking from scattered points', *Comp. J.*, v. 8, p. 383.
- Sneath, P. H. A. (1968), 'Vigour and pattern in taxonomy', *J. Gen. Microbiol.*, v. 54, p. 1-11.
- Sneath, P. H. A. (1969), 'Evaluation of clustering methods', In - Numerical Taxonomy, Academic Press, London, p. 257-271.
- Sokal, R. R. and Michener, C. D. (1958), 'A statistical method for evaluating systematic relationships', *Kans. Univ. Sci. Bull.*, v. 38, p. 1409.
- Sokal, R. R. and Sneath, P. H. A. (1963), Principles of Numerical Taxonomy, Freeman, London.
- Sørensen, T. (1948), 'A method of establishing groups of equal amplitude in plant sociology based on similarity of species content', *Biol. Skrifter*, v. 5, paper 4.

- Struve, O., and Zebergs, V. (1962), Astronomy of the 20th century, Macmillan, New York, p. 259.
- Thorndike, R. L. (1953), 'Who belongs in the family?', *Psychometrika*, v. 18, p. 267-276.
- Tolman, R. C. (1938), Principles of Statistical Mechanics, Oxford, Clarendon.
- Ward, J. H. (1963), 'Hierarchical grouping to optimize an objective function', *J. Amer. Stat. Ass.*, v. 58, p. 236.
- Williams, W. T. and Lambert, J. M. (1959), 'Multivariate methods in plant ecology. I. Association analysis in plant communities', *J. Ecol.*, v. 47, p. 83-101.
- Williams, W. T., and Lambert, J. M. (1960), 'Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis', *J. Ecol.*, v. 48, p. 689-710.
- Williams, W. T., and Lambert, J. M. (1961), 'Multivariate methods in plant ecology. III. Inverse association-analysis', *J. Ecol.*, v. 49, p. 717-729.
- Williams, W. T., Lambert, J. M., and Lance, G. N. (1966), 'Multivariate methods in plant ecology. V. Similarity analyses and information-analyses', *J. Ecol.*, v. 54, p. 427.
- Wishart, D. (1968), 'A Fortran II programme for numerical classification', St. Andrews, Scotland.
- Wishart, D. (1969a), 'A numerical classification method for deriving natural classes', *Nature*, v. 221, p. 97.
- Wishart, D. (1969b), 'The use of cluster analysis in the classification of diseases', Read at - Proc. Scot. Soc. for Experimental Medicine (Glasgow); abstract in - *Scottish Medical J.*, v. 14, p. 96.
- Wishart, D. (1969c), 'An algorithm for hierarchical classifications', *Biometrics*, v. 22, p. 165-170.
- Wishart, D. (1969d), 'Fortran II programs for 8 methods of cluster analysis (CLUSTAN I)', Kansas Comp. Contr. No. 38, State Geological Survey, Kansas, USA.
- Wishart, D. (1969e), 'Mode analysis: a generalisation of nearest neighbour which reduces chaining effects', In - Numerical Taxonomy, Academic Press (London), p. 282-311.

A MULTIVARIATE ANALYSIS OF THE DEVELOPMENT OF DUNE SLACK VEGETATION IN RELATION TO COASTAL ACCRETION AT TENTSMUIR, FIFE

BY R. M. M. CRAWFORD AND D. WISHART

*The Botany Department and the Computing Laboratories,
The University, St Andrews*

INTRODUCTION

The development of vegetation and soil on dune systems has been the subject of many investigations (Salisbury 1952) but the dune slack communities frequently associated with these habitats have received relatively less attention. A floristic account of dune slacks in Scotland has been given by Gimingham (1964). Detailed examinations of particular slacks have been carried out by Gorham (1961) on the chemical composition of dune slack water, and by Birse (1958) on the growth of byrophytes in relation to the hydrology of the slacks. However, apart from the work of Ranwell (1959, 1960) at Newborough Warren, little attention has been paid to the development of dune slack communities. In these investigations Ranwell was principally concerned with the cyclic development of *Salix repens* slacks and their distribution in relation to the height of the water table.

At Tentsmuir, owing largely to the extent and undisturbed nature of the area (a Nature Conservancy reserve since 1954), there is a great variety of dune slack types and an opportunity is presented to study their relationship to each other.

Gimingham (1964) from unpublished data of Spence has discussed some of these different slack types and considered some of the characteristic soil data for each type, but also notes that no development sequence can be detected without further investigation.

Tentsmuir is a low sandy promontory lying to the south side of the mouth of the Firth of Tay. In recent years it has been noticed that the area has been accreting steadily from the sea. Grove (1953) has mapped these changes in the coastline since 1854, using the Ordnance Survey map of 1854, and Admiralty chart 149, which was first published in 1914 (surveyed in 1912) and then re-surveyed in 1919 and 1939. Further, from plotting the line of anti-tank blocks laid down on what is believed to have been the high water mark in 1940, and using these together with his own maps of 1950 and 1953, he has produced a map of the coastal changes which covers exactly 100 years. The area was re-mapped for the present survey (October 1965) and a check was made of the coastal changes from aerial photographs taken in 1940, 1947, 1954 and 1957. The photographs agree very well with Grove's map and show that the area has accreted further since his survey. Over the past 25 years, the northern extremity of the promontory has grown with an average increment, varying with position, of from 7 to 14 m per annum.

This rate of growth makes Tentsmuir one of the most rapidly accreting areas of the British coastline, and provides an excellent opportunity for studying the developmental sequence of slack vegetation, as well as the factors which govern its distribution.

The aim of this present investigation, however, is restricted to the study of the development of the vegetation in relation to coastal accretion. The plant cover of the slacks is

subjected to a multivariate analysis, using the normal association analysis method of Williams & Lambert (1959). Having thus obtained an objective separation of the slack types on their species composition, an attempt is made to arrange them in order of floristic affinity and hence trace the most probable sequence of slack development. The distribution of these slack associations is then mapped and their location on the ground is related to the coastline changes over the past 112 years. As an additional indication of the development of the slacks, soil analyses were carried out for pH, conductivity, sodium, potassium, calcium and chloride as well as for soil moisture, loss on ignition and water table depth. Thus in one area Tentsmuir presents an opportunity to follow simultaneously the temporal, floristic and edaphic development of dune slacks.

METHODS

Sampling

The location of the Nature Conservancy's reserve at the mouth of the Tay is seen in Phot. 1. Reference posts were sited delimiting the portion of the Reserve to be surveyed. This comprised the dune and dune slack area seen in Phot. 1. A map was then prepared from a plane table survey showing the position of all the main landmarks, the reference posts and the position of the high water mark.

As the alder slacks could be identified without further analysis these were also drawn in on the map. The other slack areas were outlined but left undefined floristically. All the slacks were then visited in turn and a number of 1 metre square quadrat samples were taken at random in each. Vascular plants (excluding trees other than alder), mosses and lichens were recorded using the five-point Braun-Blanquet scale for cover-abundance. The position of each quadrat was determined by compass bearings on the reference posts. In a number of the quadrats a soil sample was taken at a depth of 10 cm and in some of these a pit was dug to the water table and a sample of the soil water was collected. In all, 263 samples were taken of the vegetation, ninety samples of the soil, and thirty-five of the soil water. Measurements of the water table depth were made after the water in the soil pits had settled to a constant height. In many soils this would give an erroneous measure due to the existence of artesian well effects (Rutter 1955). However, from digging pits in the vicinity of narrow bore pipes previously inserted to the level of the summer water table, it was found that in the sandy soil of Tentsmuir such artesian well effects were negligible.

Association analysis

The survey when completed was found to include 142 species of vascular plants, mosses and lichens. As it was felt that it would be undesirable to introduce any selection of the species to be used in the association analysis and thus defeat the objectivity of the survey, it was decided to include all the species in the χ^2 matrix.

The analysis was carried out by the procedure for normal association analysis of Williams & Lambert (1959). The calculations were carried out on an IBM 1620 computer. The matrix of 142×142 correlations was calculated in twelve cycles, the data being stored on the random access disk storage. After the computation of the matrix the values for positive and negative associations were summed separately and from these figures the value for total summed χ^2 was obtained. Subdivisions were carried out in a manner similar to that of Williams & Lambert (1960), the quadrats being divided on the presence or absence of the species with the highest $\Sigma\chi^2$ value, provided this species had a minimum

of nine occurrences. From an examination of the $\Sigma\chi^2$ values for positive and negative associations it was seen that for species with one to five occurrences only, the $\Sigma\chi^2$ value was made up almost entirely of the total for negative associations. Further, these values for negative association exceeded by far any values that were obtained for positive associations. Thus if a species is to be used as a basis for subdivision it should have an arbitrary minimum number of occurrences, otherwise a pattern of division will be arrived at which will result in the samples being split into a large number of small groups determined by the most infrequently occurring species. In this survey a species had to have a minimum of nine occurrences before it was used as a basis for subdivision.

After subdivision the matrix was recalculated and the division process repeated as above. This was repeated until no species occurred with a χ^2 value greater than 8.2. This is the same as the arbitrary figure of $(N^2)^{-5}$ (N = number of samples) introduced by Williams & Lambert (1960) for the 'short' termination of subdivision. It is realized that in a survey such as the present one, where the ratio of species recorded to samples taken, is so large, a short division analysis will not produce homogeneous groups of quadrats. However, the method is the most efficient system of division for approaching a maximum degree of homogeneity with minimum number of subdivisions. The large number of species used is essential to preserve the objectiveness of the analysis even although it increases the variability of the terminal groupings.

To facilitate the sorting of the data after the completion of the computer analysis the quadrat data along with the soil and water analyses were duplicated on punched cards suitable for hand sorting. This proved to be the most convenient way for extracting the soil analysis results as well as for finding the position of the quadrats on the maps.

Floristic comparison of quadrat groupings

Having determined the quadrat groupings by association analysis, species lists were extracted for each of these associations. The floristic similarity of these groups was then compared using Jaccard's coefficient of community (Gemeinschaftskoeffizient—see Braun-Blanquet 1964):

$$C = \frac{c}{a+b+c}$$

where C = coefficient of community, a = the number of species exclusive to one community, b = the number of species exclusive to the other community, c = the number of species common to both communities. Because of the large number of species used in the initial analysis, it was decided to limit the calculation of the coefficient of community to those species which occurred with a presence of 25% or more.

Soil and water analysis

Conductivity and pH were measured on a 1 : 2 soil to water extract. Ion analysis was carried out on a 1% citric acid extract obtained by shaking the soil sample in the extractant for 6 h and then centrifuging. Sodium, potassium and calcium were determined by flame photometry. Before assaying for sodium or potassium, calcium was removed by precipitation with saturated ammonium oxalate solution. Calcium was subsequently determined by dissolving this precipitate in 5% perchloric acid. Chloride analyses were carried out using an EEL chloride meter. The same methods were used for water analysis but aliquots of the soil water were taken instead of the citric acid extract.

RESULTS

Floristic analysis

The floristic composition of the slacks as a whole at the time of the survey is recorded in Table 1. It is realized that this list is not complete and illustrates only the floristic composition of the slacks as seen in their autumnal aspect. However, the slacks appear sufficiently rich in species to warrant an analysis of their varying types.

The pattern of division by normal association analysis is shown in Fig. 1. With the

Table 1. *The floristic composition of the dune slacks as seen in their autumnal aspect; the percentage frequency is calculated from an analysis of 263 quadrats and only species occurring with a presence of 5% or more are listed*

| Species | % Presence | Species | % Presence |
|-----------------------------------|------------|-----------------------------------|------------|
| <i>Festuca rubra</i> | 65 | <i>Potentilla erecta</i> | 8 |
| <i>Carex arenaria</i> | 44 | <i>Epilobium hirsutum</i> | 8 |
| <i>Hieracium pilosella</i> | 35 | <i>Hydrocotyle vulgaris</i> | 8 |
| <i>Filipendula ulmaria</i> | 34 | <i>Anthoxanthum odoratum</i> | 7 |
| <i>Salix repens</i> | 33 | <i>Thymus drucei</i> | 7 |
| <i>Erica tetralix</i> | 25 | <i>Alnus glutinosa</i> | 7 |
| <i>Lotus corniculatus</i> | 25 | <i>Cirsium arvense</i> | 7 |
| <i>Cladonia sylvatica</i> | 24 | <i>Pleurozium schreberi</i> | 7 |
| <i>Ammophila arenaria</i> | 22 | <i>Holcus mollis</i> | 6 |
| <i>Galium palustre</i> | 21 | <i>Lophocolea bidentata</i> | 6 |
| <i>Acrocladium cuspidatum</i> | 17 | <i>Juncus gerardii</i> | 6 |
| <i>Rhinanthus minor</i> | 17 | <i>Centaurium erythraea</i> | 6 |
| <i>Holcus lanatus</i> | 17 | <i>Hypochoeris radicata</i> | 6 |
| <i>Hylocomium splendens</i> | 16 | <i>Galium verum</i> | 5 |
| <i>Agrostis stolonifera</i> | 14 | <i>Cladonia impexa</i> | 5 |
| <i>Juncus balticus</i> | 14 | <i>Carex flacca</i> | 5 |
| <i>Dicranum scoparium</i> | 14 | <i>Chamaenerion angustifolium</i> | 5 |
| <i>Parnassia palustris</i> | 13 | <i>Bryum pendulum</i> | 5 |
| <i>Honkenya peploides</i> | 13 | <i>Barbula cylindrica</i> | 5 |
| <i>Peltigera canina</i> | 12 | <i>Erica cinerea</i> | 5 |
| <i>Rhytidiadelphus triquetrus</i> | 12 | <i>Polytrichum formosum</i> | 5 |
| <i>Hypnum cupressiforme</i> | 12 | <i>Cladonia pyxidata</i> | 5 |
| <i>Juncus effusus</i> | 11 | <i>Plantago maritima</i> | 5 |
| <i>Vicia lathyroides</i> | 10 | <i>Sagina maritima</i> | 5 |
| <i>Potentilla anserina</i> | 9 | | |

limit of division set at a maximum χ^2 value of 8.2 the analysis results in the segregation of ten separate types. The floristic composition of these groups is listed in Table 2.

Slack type 1—the first to be segregated, contains those species which would be expected to be found growing in closest proximity to the sea, e.g. *Honkenya peploides*, *Plantago maritima* and *Juncus gerardii*. Type 2 contains those species which seem to be characteristic of the driest slacks, e.g. *Hieracium pilosella*, *Carex arenaria* and *Ammophila arenaria*. The next four groups, 3, 4, 5 and 6, are all slacks containing *Erica tetralix*. The separation of these slacks from each other appears to be linked with the water relations of the slacks, type 4 being the wettest with *E. tetralix* and *Filipendula ulmaria*, ranging to a drier type 5 with *Erica tetralix* and *Cladonia sylvatica*. Type 7, which was segregated on the presence of *Agrostis stolonifera*, had as its most closely associated species *Alnus glutinosa*. From this and the distribution of these slacks on the map, this proved to be the alder type as already drawn in on the plane table survey. The last quadrats to segregate were defined by the presence of *Juncus effusus*. This left a remainder which is classified as *Filipendula ulmaria* slacks as this was the species with the highest percentage presence in the group.

Table 3 records the coefficient of community calculated from those species that occurred in the associations with a presence of 25% or more.

If the type with *Honkenya peploides* and the other halophytes is taken as the starting point it is possible by consulting Table 3 to decide which type has the highest floristic affinity with type 1 (the *H. peploides* association) and would therefore be the second slack type to evolve. This leads to type 8, the *Lotus corniculatus* slacks, being selected as the second slack type. Similarly by examining Table 3 again for the type that has the highest coefficient of community with type 8, the third type of slack to develop would be type 2,

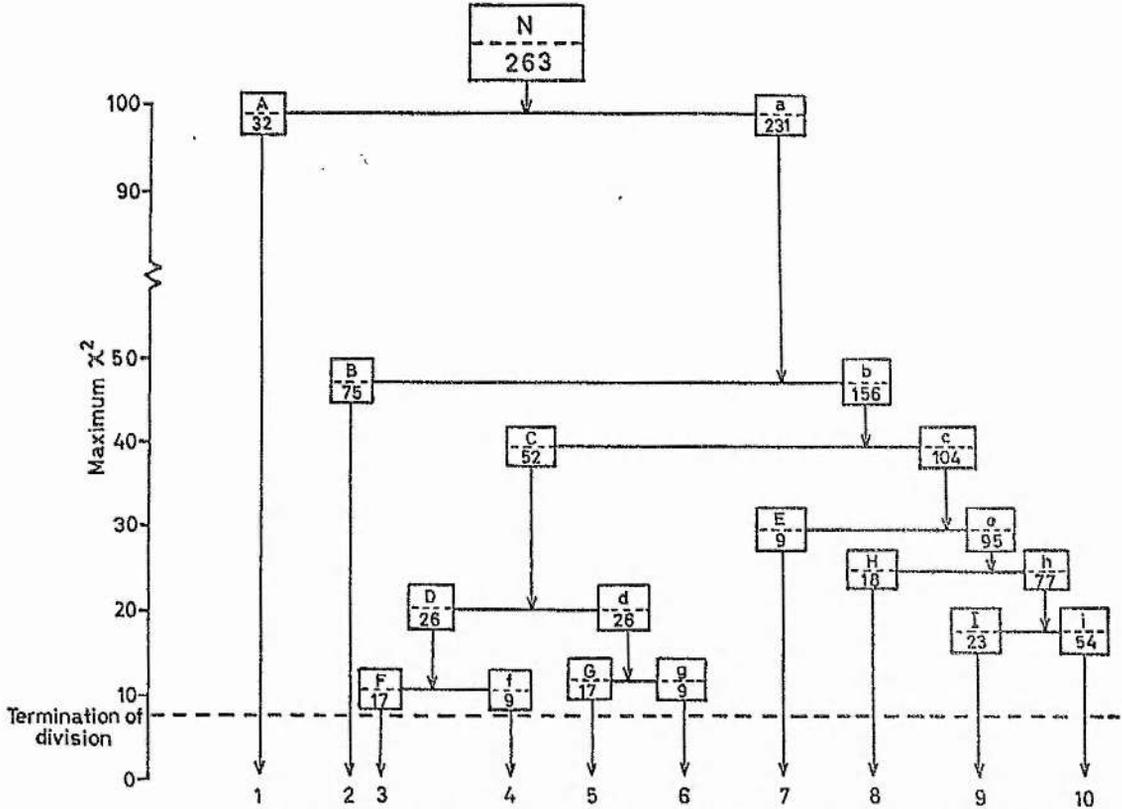


FIG. 1. Normal association analysis of the Tentsmuir slacks based on 263 quadrats and 142 species. The numbers enclosed in boxes represent the number of species involved in each division. The upper case letters represent the presence of a species, the lower case letters its absence. A, *Honkenya peploides*; B, *Hieracium pilosella*; C, *Erica tetralix*; D, *Filipendula ulmaria*; E, *Agrostis stolonifera*; F, *Carex arenaria*; G, *Cladonia sylvatica*; H, *Lotus corniculatus*; I, *Juncus effusus*.

the *Hieracium pilosella* slacks. Fig. 2 shows the arrangement of the associations on this unidirectional basis of floristic affinities.

As Tentsmuir has been steadily accreting and thus developing a gradient of change with increasing distance from the sea, there is some justification for considering the development of the slacks in this way. Nevertheless, this is bound to be a simplification of the situation, and for this reason an ordination of the associations based on the coefficients of community recorded in Table 3 and using the method of Bray & Curtis (1957) is shown in Fig. 3. As the highest community coefficient obtained was 50, the inversion of the data necessary for this method was obtained by subtracting the coefficients of community from 50. The x axis is based on types 1 and 10. In order to illustrate the development of the vegetation from type 1, the y axis also has type 1 as one reference type. Types 1 and 6 were found to give the best separation.

Table 2. The floristic composition of the slack types as segregated by normal association analysis (only species occurring with a presence of 20% or more are recorded)

| | % Presence | | % Presence |
|----------------------------------|------------|----------------------------------|------------|
| SLACK TYPE 1 | | SLACK TYPE 2 | |
| <i>Honkenya peploides</i> | 100 | <i>Hieracium pilosella</i> | 100 |
| <i>Festuca rubra</i> | 97 | <i>Festuca rubra</i> | 71 |
| <i>Rhinanthus minor</i> | 71 | <i>Carex arenaria</i> | 65 |
| <i>Agrostis stolonifera</i> | 51 | <i>Ammophila arenaria</i> | 49 |
| <i>Juncus gerardii</i> | 48 | <i>Centaurium erythraea</i> | 48 |
| <i>Hieracium pilosella</i> | 45 | <i>Salix repens</i> | 47 |
| <i>Parnassia pulustris</i> | 38 | <i>Lotus corniculatus</i> | 43 |
| <i>Plantago maritima</i> | 35 | <i>Peltigera canina</i> | 25 |
| <i>Juncus balticus</i> | 26 | <i>Dicranum scoparium</i> | 25 |
| <i>Lotus corniculatus</i> | 22 | <i>Luzula multiflora</i> | 25 |
| <i>Centaurium erythraea</i> | 22 | <i>Thymus drucei</i> | 21 |
| SLACK TYPE 3 | | SLACK TYPE 4 | |
| <i>Carex arenaria</i> | 100 | <i>Erica tetralix</i> | 100 |
| <i>Erica tetralix</i> | 100 | <i>Filipendula ulmaria</i> | 100 |
| <i>Filipendula ulmaria</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Festuca rubra</i> | 83 | <i>Hylocomium splendens</i> | 78 |
| <i>Salix repens</i> | 59 | <i>Peltigera canina</i> | 33 |
| <i>Hylocomium splendens</i> | 59 | <i>Potentilla anserina</i> | 33 |
| <i>Potentilla erecta</i> | 42 | <i>Salix repens</i> | 33 |
| <i>Vicia lathyroides</i> | 42 | <i>Carex flacca</i> | 22 |
| <i>Rhytidadelphus triquetrus</i> | 35 | <i>Galium palustre</i> | 22 |
| <i>Holcus lanatus</i> | 35 | <i>Holcus lanatus</i> | 22 |
| <i>Cladonia sylvatica</i> | 29 | <i>Potentilla erecta</i> | 22 |
| <i>Pleurozium schreberi</i> | 24 | <i>Rhytidadelphus triquetrus</i> | 22 |
| <i>Potentilla anserina</i> | 24 | <i>Vicia lathyroides</i> | 22 |
| <i>Hypnum cupressiforme</i> | 24 | SLACK TYPE 6 | |
| <i>Galium palustre</i> | 24 | <i>Erica tetralix</i> | 100 |
| <i>G. verum</i> | 24 | <i>Festuca rubra</i> | 67 |
| SLACK TYPE 5 | | <i>Carex arenaria</i> | 55 |
| <i>Erica tetralix</i> | 100 | <i>Hylocomium splendens</i> | 55 |
| <i>Cladonia sylvatica</i> | 100 | <i>Peltigera canina</i> | 55 |
| <i>Festuca rubra</i> | 88 | <i>Ammophila arenaria</i> | 44 |
| <i>Carex arenaria</i> | 53 | <i>Carex flacca</i> | 33 |
| <i>Erica cinerea</i> | 47 | <i>Galium palustre</i> | 33 |
| <i>Hypnum cupressiforme</i> | 47 | <i>Lotus corniculatus</i> | 33 |
| <i>Dicranum scoparium</i> | 41 | <i>Rhinanthus minor</i> | 33 |
| <i>Galium palustre</i> | 30 | <i>Drepanocladus uncinatus</i> | 22 |
| <i>Ammophila arenaria</i> | 30 | <i>Holcus lanatus</i> | 22 |
| <i>Rhytidadelphus triquetrus</i> | 30 | <i>Epilobium hirsutum</i> | 22 |
| <i>Acrocladium cuspidatum</i> | 23 | <i>Hypnum cupressiforme</i> | 22 |
| <i>Hylocomium splendens</i> | 23 | <i>Luzula multiflora</i> | 22 |
| SLACK TYPE 7 | | <i>Parmelia physodes</i> | 22 |
| <i>Agrostis stolonifera</i> | 100 | <i>Potentilla anserina</i> | 22 |
| <i>Alnus glutinosa</i> | 78 | <i>P. erecta</i> | 22 |
| <i>Carex arenaria</i> | 45 | <i>Salix repens</i> | 22 |
| <i>Filipendula ulmaria</i> | 45 | SLACK TYPE 8 | |
| <i>Holcus lanatus</i> | 45 | <i>Lotus corniculatus</i> | 100 |
| <i>Angelica sylvestris</i> | 22 | <i>Salix repens</i> | 83 |
| <i>Festuca rubra</i> | 22 | <i>Festuca rubra</i> | 78 |
| SLACK TYPE 9 | | <i>Carex arenaria</i> | 61 |
| <i>Juncus effusus</i> | 100 | <i>Juncus balticus</i> | 39 |
| <i>Galium palustre</i> | 78 | <i>Acrocladium cuspidatum</i> | 22 |
| <i>Filipendula ulmaria</i> | 65 | <i>Barbula cylindrica</i> | 22 |
| <i>Agrostis stolonifera</i> | 35 | <i>Parnassia palustris</i> | 22 |
| <i>Cirsium arvense</i> | 25 | SLACK TYPE 10 | |
| | | <i>Filipendula ulmaria</i> | 65 |
| | | <i>Festuca rubra</i> | 35 |
| | | <i>Hydrocotyle vulgaris</i> | 27 |
| | | <i>Carex arenaria</i> | 23 |
| | | <i>Holcus lanatus</i> | 22 |
| | | <i>Salix repens</i> | 20 |

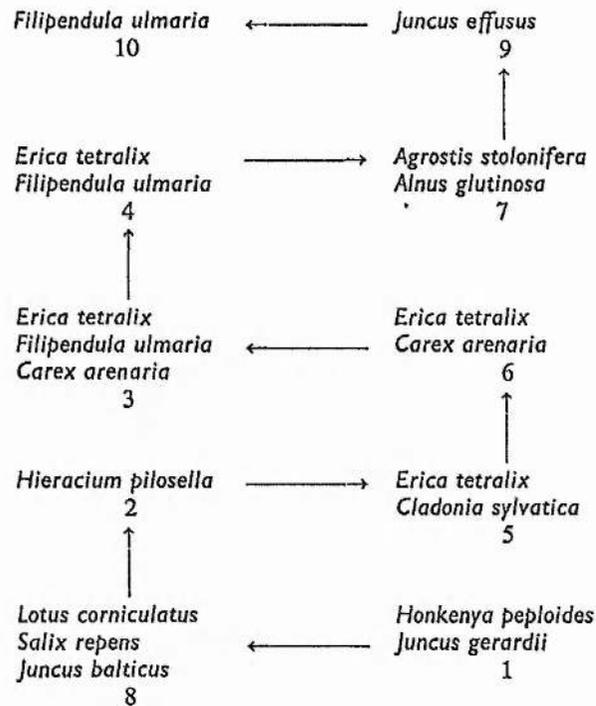


FIG. 2. Probable development of dune slack communities as determined from their floristic affinities. The differing types starting with *Honkenya peploides* are arranged so that pairs with the greatest coefficient of community are proximal to each other (see text).

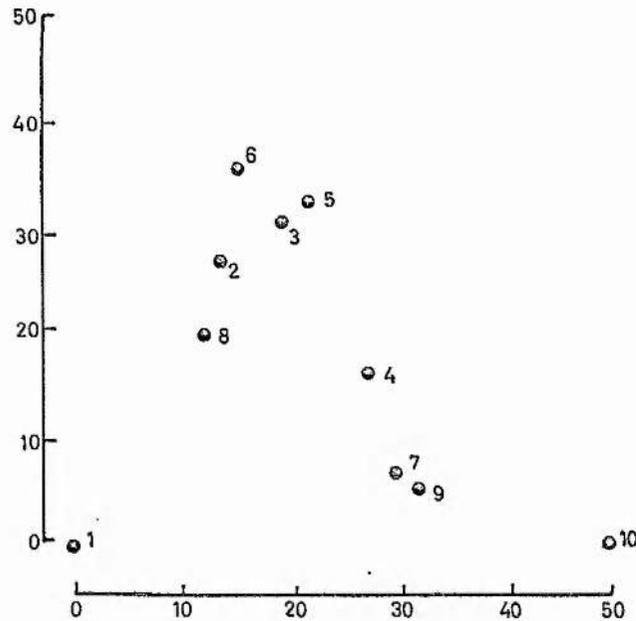


FIG. 3. Ordination of slack types obtained by association analysis; x axis based on types 1 and 10, y axis on 1 and 6.

From the distribution of the types in the ordination diagram the slacks appear to fall into three main groups, viz. young slacks composed of salt tolerant species, older dry slacks and older wet slacks. Again the type most closely related to type 1 is type 8, the *Lotus corniculatus* slacks. After this, slack development appears to follow two possible courses, either a development to the drier types, 2, 6, 3 and 5, with *Hieracium pilosella*,

Carex arenaria and *Cladonia sylvatica* or else to the wetter types, 4, 7, 9 and 10, with *Filipendula ulmaria*, *Alnus glutinosa* and *Juncus effusus*.

Table 3. Coefficient of community between the slack types segregated by association analysis, based on those species which occurred with a presence of 25% or more

| | Group | | | | | | | | | |
|----|-------|------|------|------|------|------|------|---|------|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | | | | | | | | | | |
| 2 | 12.5 | | | | | | | | | |
| 3 | 5.3 | 17.6 | | | | | | | | |
| 4 | 7.7 | 23.0 | 38.0 | | | | | | | |
| 5 | 5.9 | 26.6 | 23.5 | 13.3 | | | | | | |
| 6 | 11.1 | 25.0 | 22.2 | 12.5 | 50 | | | | | |
| 7 | 7.7 | 7.7 | 23.0 | 9.6 | 7.7 | 6.7 | | | | |
| 8 | 15.4 | 36.0 | 21.4 | 18.2 | 14.3 | 21.5 | 10.0 | | | |
| 9 | 7.7 | 0 | 0 | 9.1 | 7.1 | 6.7 | 25.0 | 0 | | |
| 10 | 0 | 0 | 0 | 11.1 | 0 | 0 | 14.3 | 0 | 14.3 | |

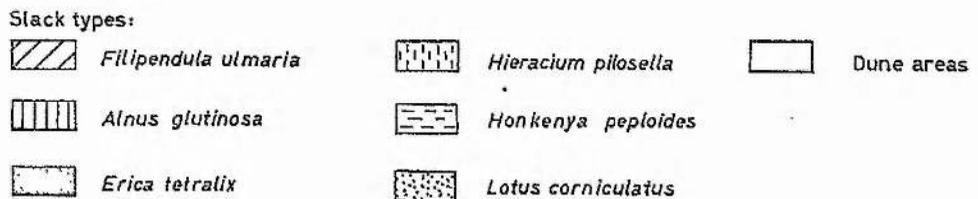
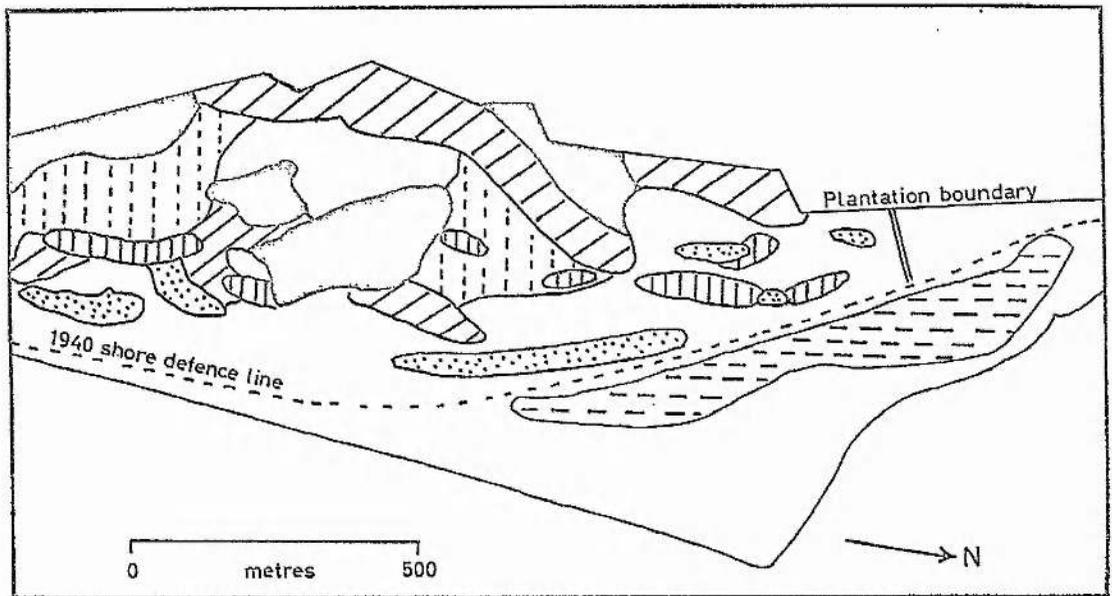


FIG. 4. Distribution of slack types determined by association analysis.

Distribution of slack types

The identification of the slack areas outlined by the plane table survey was made by plotting the positions of the quadrats from each association onto the map. The distribution of the slack associations is shown in Fig. 4. In this way it was possible to distinguish types 1, 2, 7 and 8.

However, there was an overlap on the ground when plotted at this scale between types 9 and 10 and also between types 3, 4, 5 and 6. These last four types all contain *Erica tetralix* and are mapped as *E. tetralix* slacks in Fig. 4. Similarly types 9 and 10 both contain *Filipendula ulmaria* and are mapped as *F. ulmaria* slacks. It can be seen from the map that

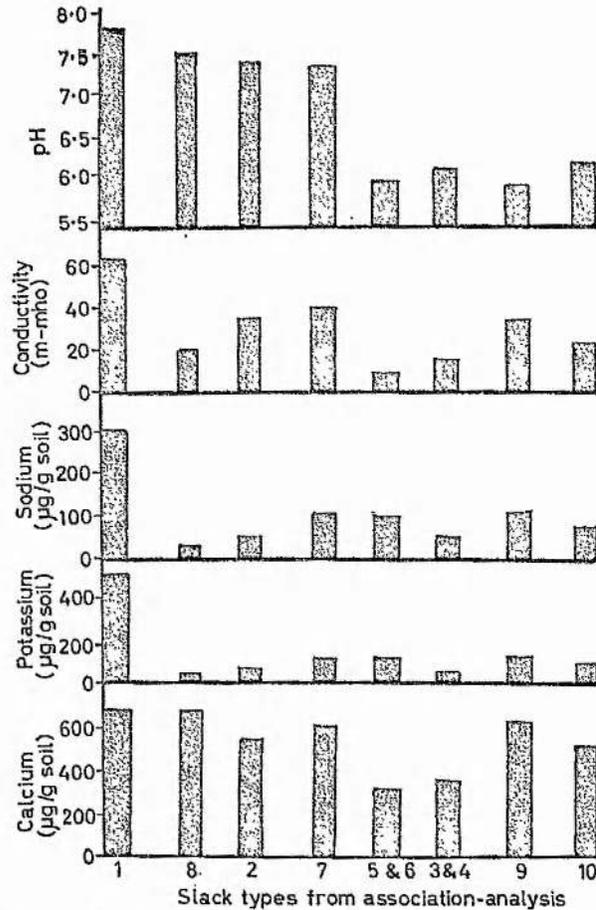


FIG. 5. Soil analysis data for pH, conductivity, sodium, potassium and calcium for each of the slack associations. Each histogram represents the mean of a minimum of ten samples except number 7 which is based on five samples.

the floristic separation of the associations in the ordination diagram is matched closely by their topographical distribution. Type 1 is confined to the areas proximal to the sea with type 8 lying next in order of succession. The dry *Hieracium pilosella* slacks (type 2) have a central position while the rear slacks are dominated by the *Filipendula ulmaria* and *Juncus effusus* (types 9 and 10), although these also have forward extensions. The *Erica tetralix* slacks also match their ordination position with their location on the ground in that they range from the dry central areas to the wetter slacks at the rear. The one exception to the close correlation between floristic and topographical development is the *Agrostis stolonifera*-*Alnus glutinosa* slacks. These have a forward position on the map behind type 8 whereas on the basis of their floristic affinities they appear to have more in common with types 9 and 10. From the position of the alder slacks seen in Phot. 2 they appear to belong to the flood-line alder association described by McVean (1956). This slack type is unique in that the tree canopy must shelter the ground flora from excessive desiccation and this may explain why this forward type of slack has its strongest floristic affinity with the wetter types to the rear.

Soil and water analysis

The results of the soil analysis are shown in Figs. 5 and 6. The slack types are arranged from left to right in order of the floristic development seen in Fig. 2, with the exception of type 7, the alder slacks, which from their position on the ground have been placed next to the *Hieracium pilosella* type. As can be seen clearly in Fig. 5 from the pH values for this type it has more in common with the *H. pilosella* and *Lotus corniculatus* slacks than with the wetter *Juncus effusus* and *Filipendula ulmaria* types. Similarly in Fig. 6 this sequence in slack development presents no marked discontinuities. The wettest slack types are found at the front and rear with the driest types in the centre.

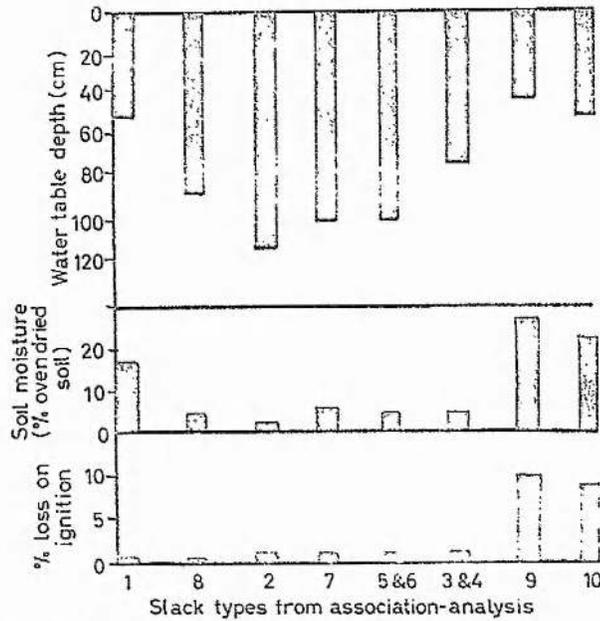


FIG. 6. Soil analysis data for water table depth, soil moisture and loss on ignition for each of the slack associations. Each histogram for water table depth represents the mean of a minimum of four samples, the other measurements are based on a minimum of ten except number 7 which is based on five samples.

It is seen in Fig. 6 and Table 4 that it is only type 1 that is still influenced in its base status by the proximity of the sea. The chloride ions which were present in high concentrations in the water sample from type 1 could not be detected in the citric acid extracts of soil samples from this area.

Table 4. Chloride analysis of soil water from samples taken in each of the floristic groups determined by association-analysis

| | Slack type | | | | | | | |
|--------------------|------------|------|------|-------|---------|---------|------|------|
| | 1 | 8 | 2 | 7 | 5 and 6 | 3 and 4 | 9 | 10 |
| Chlorine (mg/l) | 708 | 21.7 | 21.5 | 94.0 | 21.3 | 21.9 | 13.0 | 21.8 |
| Standard deviation | ±268 | ±6.6 | ±3.6 | ±12.1 | ±6.0 | ±3.6 | ±1.3 | ±2.9 |

Quantitative distribution of *Festuca rubra*

As *F. rubra* was by far the major component of the vegetation at Tentsmuir, occurring with a presence of 65% (see Table 1), an analysis of its distribution was made quantitatively. From the Braun-Blanquet ratings used, it was found that *F. rubra* had been rated from x

(very rare) to 4 (50-75%). The quadrats were sorted into five groups depending on the Braun-Blanquet rating for *F. rubra* and the mean position on the map grid as well as the standard deviation to each axis was calculated. The results are illustrated in Fig. 7. The extent of the cross lines indicates the fiducial limits for the spread of each group at $P = 0.05$. It can be seen that the performance of *F. rubra* decreases in a southerly

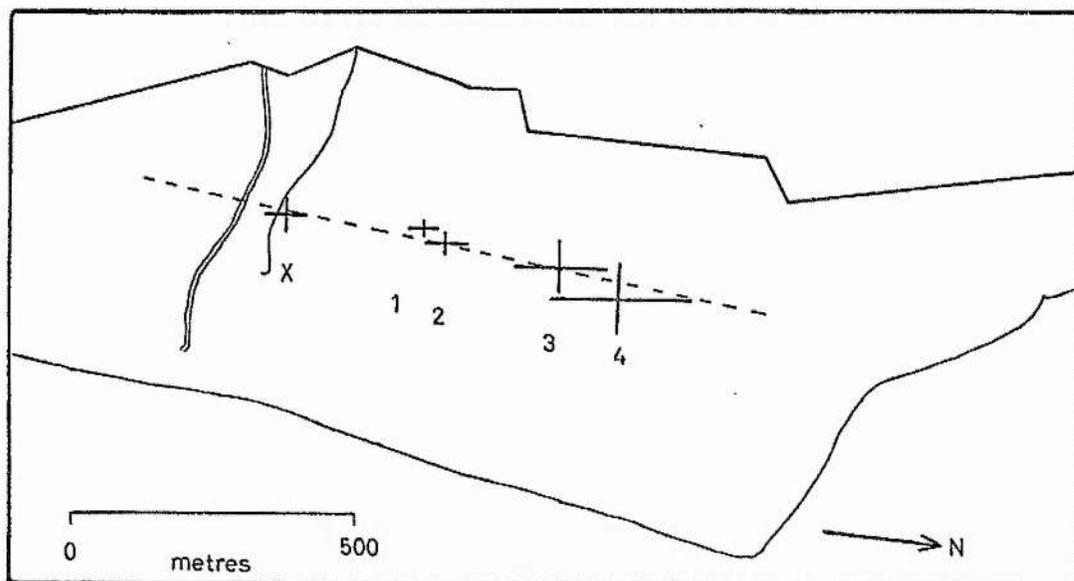


FIG. 7. The quantitative distribution of *Festuca rubra* as determined by the mean position on the map grid for each Braun-Blanquet rating. The cross lines represent the fiducial limits ($P = 0.05$) for the distribution of *F. rubra* at each Braun-Blanquet rating.

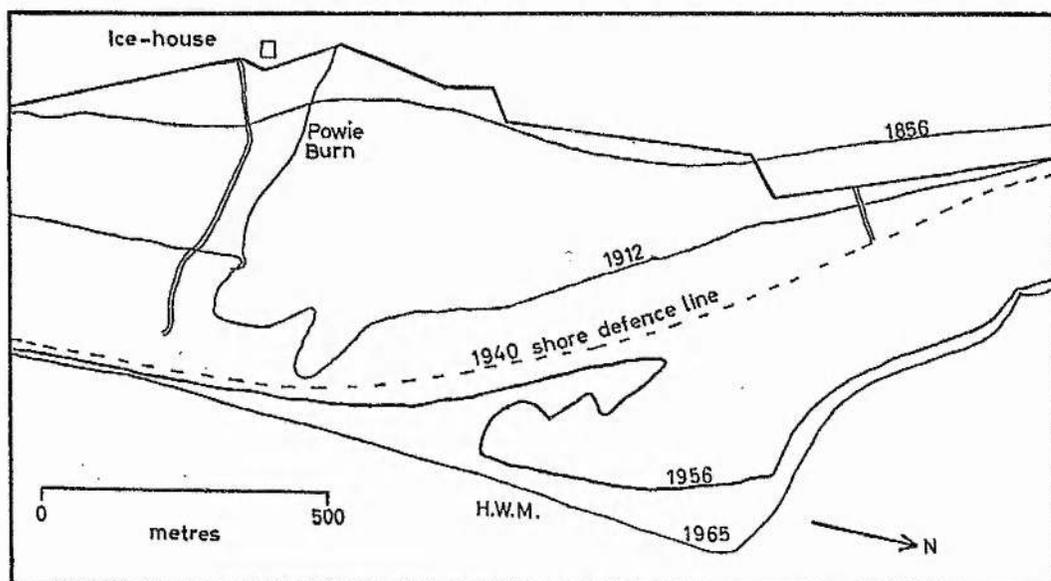


FIG. 8. Coastal changes at Tentsmuir from 1856 to 1965. All the coastlines except the most recent are taken from Grove (1953).

direction. This distribution change for a single species is matched closely by the distribution of the different slack types. The *Honkenya peploides* slacks are found to the north with immediately behind them the *Lotus corniculatus* type and to the south of these the *Erica tetralix* and *Filipendula ulmaria* slacks develop.

Coastline changes

The changes in the position of the high water mark since 1854 plotted in Fig. 8 show a close correlation with the north-south trend in the development of the vegetation. Between 1854 and 1912 the main area of growth lay in a region some 300 m east-north-east of the ice-house. From 1912 to 1953 the region of maximum accretion had moved some 900 m northwards. Recent measurements indicate that this area has now stabilized itself and that the zone of maximum accretion has moved farther north.

DISCUSSION

The slack types described in this investigation range from salt- to freshwater marsh. This is unavoidable, for in a comparative study of vegetation changes, the slacks have been defined not on the basis of their vegetation but on their physical development. In this investigation any area of flat land formed behind the first line of dunes and produced as a consequence of dune erosion has been considered as a dune slack. Due to this erosion the vegetation of these areas comes within the zone of influence of the water table. All the slacks considered in this investigation have been seen to flood with the exception of types 2 and 5, the *Hieracium pilosella* and *Erica tetralix-Cladonia sylvatica* slacks. However, even in these drier slacks the water table has been seen to rise to within the rooting zone.

On this basis undoubtedly the first slack type to appear is the *Honkenya peploides* type. This slack is most clearly seen in Phot. 3 where it has developed seawards of the anti-tank blocks laid down in 1940. The rate of accretion in this area has been particularly rapid as the entire area of the slack has formed during the last 25 years. The advancing line of mobile dunes is no more than 6-8 ft in height. It appears that as they advance seawards, they erode rapidly on their landward side, and due to their lack of height leave a slack that is only 2-3 ft (60-90 cm) above the high water mark. Soil samples taken from this area, although of a high pH and high in sodium and potassium content were totally devoid of chloride ions.

It is probable that the sandy nature of the soil accounts for the poor anion retention, and samples taken in October had been washed free of chloride by the summer rains. However, the soil water contained considerable quantities of chloride. As the slack is only 2-3 ft above the high water mark the drainage seawards will be slow and result in the retention of these chloride ions in the soil water. So the vegetation, although not constantly exposed to high chloride concentrations, will nevertheless be subjected to them periodically, whenever the water table rises.

The next type of slack to develop is the *Lotus corniculatus* type. This is distinguished from the first type by being found on ground that is higher above the high water mark. In a number of slacks that were surveyed, this type was found to lie about 8-10 ft (250-300 cm) above the high tide line. This extra height gives a better drainage gradient towards the sea and where these slacks occur the soil water is free of the large concentrations of chloride found with the *Honkenya peploides* type. These slacks lie to the landward of the 1940 shore defence line and thus appear to be over 25 years old.

To the rear of the *Lotus corniculatus* slacks lies the flood-line alder type. A number of trunk borings were made of the oldest trees in these slacks and the maximum age of tree was found to be 23 years.

This agrees well with their position in relation to the accretion of the Reserve as this is approximately the time that this area would have been free of salt flooding.

The principal development of the *Hieracium pilosella* slacks is to the rear of the alder line. Here the slacks are some 10–14 ft (300–425 cm) above the high water mark. These are the driest slacks with the water table at its maximum depth (see Fig. 6). These slacks are formed from the partially eroded late fixed dunes. In Phot. 2 they can be seen as the dark patches of vegetation extending into the lighter coloured late fixed dunes.

The water table movements at Tentsmuir are very similar to those described by Ranwell (1959) at Newborough Warren except for the absence of any landward drainage. At Tentsmuir the water table descends with an ever-increasing gradient towards the sea. The maintenance of this gradient is probably influenced by a stream, the Powie Burn, whose point of entry into the Reserve is shown in Fig. 8. There is little change in the height of the slacks behind the late fixed dunes but with increasing distance from the sea the water table comes ever closer to the surface. It has also been noted at Tentsmuir that the water table rises much more rapidly in the rear slacks than in those nearer the sea. For every 1 in. of rise in the *Lotus corniculatus* slacks the water table in the *Filipendula ulmaria* slacks rises 3 in. The increased wetness of the rear slacks with concomitant effects of flushing of the soil are seen in Figs. 5 and 6.

Again it is probable that these effects are enhanced by the entry of the stream into this area. These changes in water table level are followed closely by the successional vegetation changes outlined in Fig. 2. As the water table rises *Erica tetralix* is gradually replaced by *Filipendula ulmaria* which finally becomes dominant in the rear slacks.

The first *Erica tetralix* types to appear are found on land that has only accreted from the sea since 1912. This would make these slacks at the most only 53 years old. In this period the pH of the soil has dropped from 7.8 to 6.1. This is much more rapid than the rates of change reported by Salisbury (1952) where the pH of a successive series of dunes at Blakeney Point fell from 7.2 to 6.1 in 235 years and on a similar series at Southport from 8.2 to 6.4 in just over 200 years.

The rear slacks with the *Filipendula ulmaria* and *Juncus effusus* types are most clearly seen in Phot. 4. Those lying closest to the plantation are probably about 100 years old. In these slacks, which are the most prone to flooding, sulphide and ferrous ions were always detectable at the level of the summer water table. This is in marked contrast to the more seaward slacks where sulphides were never found and ferrous iron only occasionally.

The northerly trend in the development of the vegetation already mentioned in connection with Fig. 7 is also clearly seen in Phot. 4, taken looking north over the reserve from the rear slacks. The *Filipendula ulmaria* slacks lie in the foreground with the *Erica tetralix* slacks beyond.

Behind these are the *Hieracium pilosella* slacks bounded on their northern side by the alder line. In the far distance, between the alder slacks and the shore defence line, lie the *Lotus corniculatus* slacks with *Salix repens* and *Juncus balticus*. At the top of the photograph the *Honkenya peploides* slacks are just visible.

Coastal accretion

Although the purpose of this investigation is to describe the development of slack vegetation on a floristic and temporal basis it also provides an opportunity to study some of the factors controlling the accretion of land from the sea. To the north of Tentsmuir point lie the Abertay sands. These are large banks of sand deposited by the river Tay when it reaches salt water. Accretion takes place at Tentsmuir whenever the wind blows to the

Reserve from the banks, that is when the wind is from the north. In one week of northerly winds in November 1965 the fore dunes at Tentsmuir point accreted by up to 1 ft in depth. Grove (1953) suggests that the deposition of such large quantities of sand at the mouth of the Tay may be related to the felling of the Scottish forests and the subsequent land erosion that would increase the sediment load brought down by the river Tay. Lamb (1965) has shown that the percentage of days with westerly winds in Britain over the past 35 years has fallen by 30%. These westerly winds tend to be replaced by northerly ones.

This may well be an additional factor resulting in the very rapid growth of land at Tentsmuir as it is these winds that contribute most to the growth of the foredune systems. The shelter afforded by the Forestry Commission plantation appears also to play a role, for in the dunes to the south of the reserve which have less shelter from the plantation many more 'blow outs', caused by the westerly gales, are found.

ACKNOWLEDGMENTS

We are much indebted to the Nature Conservancy in Edinburgh for access to records, maps and photographs of the reserve. We are also most grateful to Mr A. T. Grove for permission to use his maps of the coastline changes at Tentsmuir. We are further indebted to Mr R. L. Constable for statistical advice, to Mr R. A. L. Oliver for identifying the bryophytes, and to the Honours Botany Class 1965 and Mr R. M. Campbell for assistance in collecting the data.

SUMMARY

The Tentsmuir sands have been noted in recent years for being one of the most rapidly accreting parts of the British coastline. A number of varying dune slack types are represented here, and, with the aid of a computer, an objective classification of these slacks has been made by normal association-analysis based on 142 species, and the resultant types studied in relation to their age and floristic development. A sequence of slack development is suggested which, beginning with a *Honkenya peploides*-*Juncus gerardii* type, evolves into a *Salix repens*-*Juncus balticus* type. Depending on water table depth and flooding frequency the slacks then develop into *Ahus glutinosa*, *Hieracium pilosella* and varying *Erica tetralix* types. Finally the slacks evolve into marsh vegetation with *Filipendula ulmaria* and *Juncus effusus*. This pattern of floristic development is matched closely by the distribution of the slacks in relation to the coastal accretion. The growth of the area over the past 112 years is known from charts and maps and by plotting the results of the association-analysis on a map of coastline changes it was possible to follow the physical and floristic development of the slacks. The first change in slack vegetation from the *Honkenya peploides*-*Juncus gerardii* type to the *Salix repens*-*Juncus balticus* slacks requires a minimum of 25 years. The first *Erica tetralix* slacks appear on land that is, at the most, only 53 years old. The oldest slacks in the area, those of *Filipendula ulmaria* and *Juncus effusus*, are in the region of 100 years old. This floristic and physical development is also matched by changes in soil pH, conductivity, mineral content, moisture and water table depth.

REFERENCES

- Birse, E. M. (1958). Ecological studies on growth form in Bryophytes. III. The relationship between the growth of mosses and ground water supply. *J. Ecol.* 46, 9-27.
Braun-Blanquet, J. (1964). *Pflanzensoziologie*, 3rd edn. Vienna.

- Bray, J. R. & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325-49.
- Gimingham, C. H. (1964). Maritime and sub-maritime communities. *The Vegetation of Scotland* (Ed. by J. H. Burnett), pp. 67-142. Edinburgh.
- Gorham, E. (1961). The chemical composition of some waters from dune slacks at Sandscale, north Lancashire. *J. Ecol.* 49, 79-82.
- Grove, A. T. (1953). *Tentsmuir, Fife; soil blowing and coastal changes*. Unpublished report lodged with the Nature Conservancy, Edinburgh.
- Lamb, H. H. (1965). Britain's changing climate. *Symp. Inst. Biol.* 14, 3-31.
- McVean, D. N. (1956). Ecology of *Ahvis glutinosa* (L.) Gaertn. III. Seedling establishment. *J. Ecol.* 44, 195-218.
- Ranwell, D. S. (1959). Newborough Warren Anglesey. I. The dune system and dune slack habitat. *J. Ecol.* 47, 571-601.
- Ranwell, D. S. (1960). Newborough Warren Anglesey. II. Plant associates and succession cycles of the sand dune and dune slack vegetation. *J. Ecol.* 48, 117-41.
- Rutter, A. J. (1955). The composition of wet heath vegetation in relation to the water table. *J. Ecol.* 43, 507-43.
- Salisbury, E. J. (1952). *Downs and Dunes*. London.
- Williams, W. T. & Lambert, J. M. (1959). Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* 47, 83-101.
- Williams, W. T. & Lambert, J. M. (1960). Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. *J. Ecol.* 48, 689-701.

(Received 11 March 1966)

A RAPID MULTIVARIATE METHOD FOR THE DETECTION AND CLASSIFICATION OF GROUPS OF ECOLOGICALLY RELATED SPECIES

BY R. M. M. CRAWFORD AND D. WISHART

*The Botany Department and Computing Laboratory,
The University, St Andrews*

INTRODUCTION

An expansion in the use of objective methods of vegetation analysis is hindered for many ecologists by the need for access to a computer with a large high speed memory. The original surveys of Williams & Lambert (1960, 1961) contained a maximum of seventy-two species which was just short of the maximum capacity of the computer (seventy-six species). Any increase in the size of surveys that can be investigated by association analysis tends to be limited by the size of the computer available. Because of the larger high speed memory of the Elliot 803 computer, Ivimey-Cook & Proctor (1966) have been able to analyse a survey that contained 132 species and 150 samples. By making use of random access disc storage, association analysis can be carried out on data that exceeds the capacity of the high speed memory. Crawford & Wishart (1966) carried out such an analysis on an IBM 1620 (model II) using 142 species and 263 samples but the time required was too lengthy for the system to be used as a general routine.

It would be an advantage therefore, if a simpler method of analysis could be found that would depend on calculations suited to the capacity of a small machine, rather than trying to follow a method of statistical analysis which becomes unwieldy on a large scale.

The present study was undertaken in order to classify the vegetation of wet land for further work on the metabolism of ecologically related groups of species in relation to flooding tolerance. It was necessary to classify the vegetation into ecological groups or sets of quadrats so that these sets contained the major groups of coincident species. These groups of species and the quadrats in which they were found, could then be examined for any correlations between species adaptations and flooding frequency.

In vegetation classification the concept of the classificatory unit varies with the end in view. For phytosociologists it is the community and this depends on the definition of 'Kennarten' by which the community can be recognized in the field, or in abstracted tables. For Williams & Lambert (1959, 1960) although the process of arriving at the quadrat groupings is similar to the phytosociological method (see Ivimey-Cook & Proctor 1966) in that the presence or absence of species is used to define the quadrat sets, the plant community is considered as a homogeneous set of quadrats or species that is frequently not obvious in the field.

The method described in this paper differs, in that it is an attempt to distinguish the major groups of coincident species and thus searches for gregariousness rather than homogeneity. The method aims, firstly, at being rapid even when the survey is large, and secondly, at obtaining an absolute value for the group significance of any intermediate or final set of quadrats. This is considered important as owing to the continuously variable nature of vegetation, not all classifications can be expected to have

the same significance. Thirdly, the method attempts to measure the significance of each species in forming any final grouping and of each quadrat in belonging to any such group. This retrieval of all the attributes of both samples and species after the process of classification allows the process to be followed by ordination as has been recommended by several authors (Greig-Smith 1961; Gittins 1965).

Using data from two different surveys the results of this method are compared with those obtained using the normal association analysis method of Williams & Lambert (1959). A subsequent communication will describe how these results may be viewed graphically by coupling a digital plotter to the computer and thus enabling the groups and stands to be both mapped and ordinated for immediate inspection.

For the IBM 1620 Model II with 60K core and at least two random access drive units a complete ecological survey would be restricted to 2000 species and 400 000 species records. Hence for an average of 20 species/quadrat the survey would be limited to 20 000 quadrats and with an average of 40 species/quadrat, the quadrat limit would be reduced to 10 000. This method should therefore be adequate for analysing all or any part of the British flora.

DATA

The data used in this paper were taken from two separate surveys, one carried out on wet land vegetation in the north of the Isle of Arran (Buteshire), and the other at the Nature Conservancy's reserve at Tentsmuir (Fife). The north of Arran is considered as that part of the island which lies north of the Highland boundary fault. Wet land is defined physiographically as any area which because of its position is prone to flooding or semi-permanent waterlogging of the soil due to inefficient drainage. Five sites were chosen as representing the major areas of wet land on the island: (1) Glen Sannox (GR NR 995450), the flat floor of a glaciated valley 3 km long by 800 m wide with the entire area lying below the 500 ft (152 m) contour; (2) the fore shore at Corrie (GR NS 020445), an area approximately 1000 × 300 m lying between the foot of cliffs and the high tide mark and kept constantly wet by drainage from above; (3) Glen Diomhan (GR NR 934458), a wet valley lying above the 1000 ft (304 m) contour; (4) Lochranza Bay (GR NR 938505), an alluvial flat at the head of a sea loch (fiord); and (5) a wet upland moor (GR NR 882426), south-east of Pirnmill lying above the 800 ft (243 m) contour. All these areas were sampled at random using 1 m square quadrats. A total of 554 samples were taken and 182 species listed. For the purpose of the analysis only those species that occurred in at least 1% of the samples were included. This reduced the species list to ninety-eight.

The second set of data is taken from a survey of dune slack types in relation to coastal accretion carried out at Tentsmuir (Crawford & Wishart 1966).

METHODS

Group analysis

In this study the field survey data are examined for the occurrence of major groupings of coincident species and therefore it is not only the floristic similarity of the quadrats that is assessed but also their floristic richness. On this basis there are two factors which determine the likelihood of a species being contained in a group; the probability of its

occurrence and the number of species with which it occurs. For a species, X, the first factor, the probability (P) of its occurrence is given by:

$$P_x = \frac{\text{The number of occurrences of species X}}{\text{The number of quadrats in the population sample}} \\ = \frac{f_x}{N} \quad (1)$$

The second factor, the number of species with which it occurs is given by the *mean sample density* (V_x), the average number of species present in those quadrats that contain the species X. This may be calculated as:

$$V_x = \frac{\text{The total number of species occurrences in those quadrats containing X}}{\text{The frequency of X}} \\ = \frac{M_x}{f_x} \quad (2)$$

As we consider it is the species which occur frequently with high mean sample density that determine an ecological group and not those which are frequent but isolated, or infrequent yet occurring in floristically rich areas, we propose to use the product of mean sample density and species probability, symbolically

$$W'_x = P_x \cdot V_x \quad (3)$$

as the measure of the significance of a species contributing to a group. W'_x is termed the *group element potential* (GEP) of a species X.

Having evaluated the W'_x values for the species represented by a set of quadrats it is not only possible to classify each species according to this value but also to constitute a classification of the individual quadrat attributes. If the potential of each species in a quadrat is known for forming a 'general group' then it may be regarded as axiomatic that the greater the significance of those species present for forming a group the more likely it is that the quadrat falls into a 'general group area'. If the species GEP may be used as a measure of species significance then the sum of these values for those species present in a quadrat can be used to describe the group attributes of that quadrat. It is proposed therefore that the sum of the GEP values for those species occurring in a quadrat (represented as S'_j) be taken as a measure of the group attributes of that quadrat (J). It follows that the maximum value for S'_j is obtained when a quadrat contains all the species in the population sample N . In order that the group attributes of the quadrat may be represented as an absolute coefficient for the sample population in question it is proposed to redefine the *set element potential* (SEP) as

$$S_j = \frac{S'_j}{S'_{j_{\max}}} \quad (4)$$

Thus for any quadrat the values for S_j will lie between 0 and 1.

Similarly it is convenient to re-define W'_x so as to obtain an absolute coefficient:

$$W_x = \frac{P_x \cdot V_x}{\bar{V}} \quad (5)$$

where (\bar{V}) is the mean sample density for all the quadrats in the population sample. Hence, it can be shown that W_x will also lie between 0 and 1.

If the SEP value for any quadrat J represents its positive attributes for belonging to a set then the negative attributes, the *non-set element potential* \bar{S}_j is represented by the complement of S_j , with

$$S_j + \bar{S}_j = 1$$

Thus when dividing a set of quadrats on the presence or absence of a species X it is possible to sum the SEP values obtained with species X and those without species X and likewise for the non-set element potential (i.e. S and \bar{S}). On this basis a measure of the interaction between species and group potential can be tested for each species in turn by examining a two-dimensional array as shown below.

| Species X | | |
|-----------|-------|------------------|
| - | + | |
| A | B | $\Sigma \bar{S}$ |
| C | D | ΣS |
| N-f | f_x | N |

Using the statistic

$$\mu'^2 = \Sigma(o_i - e_i)^2,$$

the sum of the squared cell deviations from expectation, as a measure of interaction between species X and the quadrat attributes, it has been found that the species with the maximum interaction (μ'^2) produces the most satisfactory division of the data, segregating all the known ecological types and leaving a minimum number of residual groups. In calculating μ'^2

$$\begin{aligned} \mu'^2 &= \Sigma(o_i - e_i)^2 \\ &= (A - e_A)^2 + (B - e_B)^2 + (C - e_C)^2 + (D - e_D)^2 \end{aligned}$$

the expected values e_i are estimated from the marginal totals, e.g.

$$e_D = \frac{f_x \cdot S}{N}$$

This reduces after manipulation to

$$\begin{aligned} \mu'^2 &= \frac{4}{N^2} (D \cdot N - f_x \Sigma S)^2 \\ &= 4(D - P_x \Sigma S)^2 \end{aligned}$$

or after division by the constant 4,

$$= (D - P_x \Sigma S)^2 \tag{6}$$

The value of the marginal total ΣS , divided by the number of quadrats in the population sample, is taken as a measure of the significance of the sub-set and is termed the group coefficient, $C = \frac{1}{N} \Sigma S$. Division stops when C exceeds an arbitrary limit ϕ .

For the analysis in this present work ϕ was chosen as 0.5. There are, however, reasons for varying this value and these are considered later in the discussion.

Apart from the termination of division as determined above it has been found necessary to set another limit to division in order to avoid the continuous division of quadrat sets that fail to reach the desired level of significance. This again has been found by convenience and is determined by the value of ΣS . When ΣS falls below 10 division is terminated.

Calculation procedure with hypothetical model

Table 1 illustrates the calculation of the above values for a hypothetical model. In the example shown each quadrat is represented by a row and each species present by an 'x' in the appropriate column. Two homogeneous groups A and B are represented with section C containing rare species. The following species distribution types are illustrated.

Type Q is quasi-ubiquitous throughout A.

Type R is ubiquitous throughout A.

Type S is universal.

Type T is quasi-ubiquitous throughout B.

Type U is rare, group C.

Calculation

The species probability of occurrence and species mean sample density are obtained directly by applying formulae (1) and (2). The mean sample density

$$\bar{V} = \frac{(6 \times 7) + (4 \times 4) + (3 \times 2)}{13} = 4.923$$

The GEP values may now be calculated using formula (5)

$$\text{e.g. for species 6 (type Q)} \frac{P.V}{\bar{V}} = \frac{0.384 \times 7.0}{4.923} = 0.55$$

Using the GEP values obtained the programme then refers back to the original quadrat data and the SEP values are calculated for each quadrat:

$$\begin{aligned} \text{e.g. for quadrat (1)} &= \frac{(5 \times 0.55) + (1 \times 0.66) + (1 \times 1.0)}{(5 \times 0.55) + (1 \times 0.66) + (1 \times 1.0) + (4 \times 0.19) + (3 \times 0.3)} \\ &= 0.76 \end{aligned}$$

The sum of these values for all the quadrats, ΣS

$$(6 \times 0.76) + (4 \times 0.27) + (3 \times 0.18) = 6.17$$

The programme may now compute the species quadrat interaction from formula (6), e.g. for species type Q

$$\mu^2 = [(5 \times 0.76) - (6.17 \times 0.384)]^2 = 2.0$$

The group coefficient

$$C = \frac{1}{N} \Sigma S = \frac{1}{13} \times 6.17 = 0.475$$

The summed χ^2 values (with Yates correction applied in all cases) have been calculated for comparison. Since for this example $\Sigma S < 10$ no division would be made, but this is due to the small size of the model. Clearly division would take place on species 7, the most suitable to determine the major group A. (Note: the above calculations obtained by computer are subject to a certain amount of round-off error.)

Comparison of groups

For comparison with the above method the data from both surveys have been analysed by the method of Williams & Lambert (1959). The subsequent quadrat groupings obtained from this analysis as well as those obtained by group analysis are compared for their floristic similarity using Czekanowski's coefficient (see Greig-Smith 1964), calculated by computer;

$$c = \frac{2w}{a + b} \times 100$$

where c is the coefficient of community between the units compared, a and b the species contained in the two areas respectively and w those species contained in common. The mutual floristic affinities of the types segregated by the two forms of analysis are then compared using the ordination procedure of Bray & Curtis (1957). As has been pointed out by Austin & Orloci (1966) this method does not result in a Euclidean representation of interstand distance and exaggerates the appearance of a continuum. However, the ordinations presented here are intended only as a graphic representation of the results of classification, and not as a true geometric representation of species groupings.

Computer programme

The systems used in the analyses described in this paper were written in Fortran II D and are available on application to the authors.

RESULTS

To test the analysis system on a survey which contained no groupings a set of random species occurrences for eighty species in ninety quadrats was produced by means of a pseudo-random number generator. In no case did the group coefficient rise above 0.25 while ΣS was > 10 . The division pattern resulted in the splitting off of groups of two and three quadrats throughout the entire set of ninety. Thus in a homogeneous set of data no groups will be recognized.

North Arran survey

The overall floristic composition of the five sites is shown in Table 2. Only those species that have a presence of more than 5% are listed. When subjected to normal

association analysis, as shown in Fig. 1, fifteen types of wet land vegetation are distinguished. However, the division of the quadrats by group analysis as illustrated in Fig. 2, results in the segregation of thirteen vegetation types although only six of these types rank by definition ($\phi = 0.5$) as significant quadrat sets, i.e. types 1, 2, 3, 6, 9 and 11. The floristic composition of the differing wet land vegetation types as segregated by association and group analysis is recorded in Tables 3 and 4 respectively. Considering

Table 2. *The floristic composition of wet land vegetation in the north of Arran as seen in a random survey of five sites; the percentage frequency is calculated from an analysis of 554 quadrats and only species with a presence of 5% or more are listed*

| Species | % presence | Species | % presence |
|---------------------------------|---------------|----------------------------------|---------------|
| <i>Molinia caerulea</i> | 65 | <i>Ranunculus repens</i> | 9 |
| <i>Potentilla erecta</i> | 58 | <i>Trifolium repens</i> | 9 |
| <i>Sphagnum</i> spp. | 49 | <i>Holcus lanatus</i> | 8 |
| <i>Erica tetralix</i> | 49 | <i>Festuca ovina</i> | 8 |
| <i>Calluna vulgaris</i> | 45 | <i>Rumex acetosella</i> | 8 |
| <i>Trichophorum cespitosum</i> | 38 | <i>Juncus effusus</i> | 8 |
| <i>Narthecium ossifragum</i> | 31 | <i>Pteridium aquilinum</i> | 7 |
| <i>Polygala serpyllifolia</i> | 23 | <i>Poa pratensis</i> | 7 |
| <i>Eriophorum angustifolium</i> | 21 | <i>Iris pseudacorus</i> | 6 |
| <i>Drosera rotundifolia</i> | 21 | <i>Cirsium palustre</i> | 6 |
| <i>Myrica gale</i> | 18 | <i>Carex panicea</i> | 6 |
| <i>Eriophorum vaginatum</i> | 17 | <i>Cladonia arbuscula</i> | 6 |
| <i>Anthoxanthum odoratum</i> | 15 | <i>Luzula multiflora</i> | 6 |
| <i>Festuca rubra</i> | 12 | <i>Hypnum cupressiforme</i> | 5 |
| <i>Carex echinata</i> | 12 | <i>Lycopodium selago</i> | 5 |
| <i>Galium saxatile</i> | 11 | <i>Campylopus atrovirens</i> | 5 |
| <i>Rhytidadelphus loreus</i> | 11 | <i>Glaux maritima</i> | 5 |
| <i>Juncus acutiflorus</i> | 11 | <i>Matricaria matricarioides</i> | 5 |
| <i>Rhacomitrium lanuginosum</i> | 10 | <i>Plantago maritima</i> | 5 |
| <i>Deschampsia flexuosa</i> | 10 | <i>Hydrocotyle vulgaris</i> | 5 |
| <i>Conopodium majus</i> | 9 | | |

only those vegetation types segregated by group analysis that rank as significant sets, these may all be matched with corresponding types (with the exception of type 3) from association analysis. A comparison of these matching sets is shown in Table 5 which sets in juxtaposition those sets from the two systems that have the greatest floristic similarity as measured by Czekanowski's coefficient.

An ordination of the group and association analyses types according to the method of Bray & Curtis (1957) is shown in Figs. 3 and 4. The association analysis groupings display a continuum of floristic change ranging from the salt tolerant *Festuca rubra*-*Glaux maritima* type through the wet fore-shore and alluvial flat types with *Anthoxanthum odoratum* and *Iris pseudacorus* to the base deficient high level bogs with *Molinia caerulea* and *Rhacomitrium lanuginosum*. The pattern obtained in Fig. 4 with the groups segregated by group analysis illustrates the distinctness of the six significant sets while the indeterminate nature of the ecotone types is demonstrated by their position close to each other in the centre of the diagram. In the Arran survey only 67% of the 554 quadrats sampled are classified as belonging to significant sets, the remaining 33% being regarded as transitional or ecotone vegetation.

Tentsmuir dune slacks

Fig. 5 illustrates the division of the quadrat data from the dune slacks by group analysis. From the 263 samples taken, only 22% are classified into significant sets and these fall into three groups. The floristic composition of the vegetation types is shown in Table 6. In an association analysis of the same data (Crawford & Wishart 1966) ten

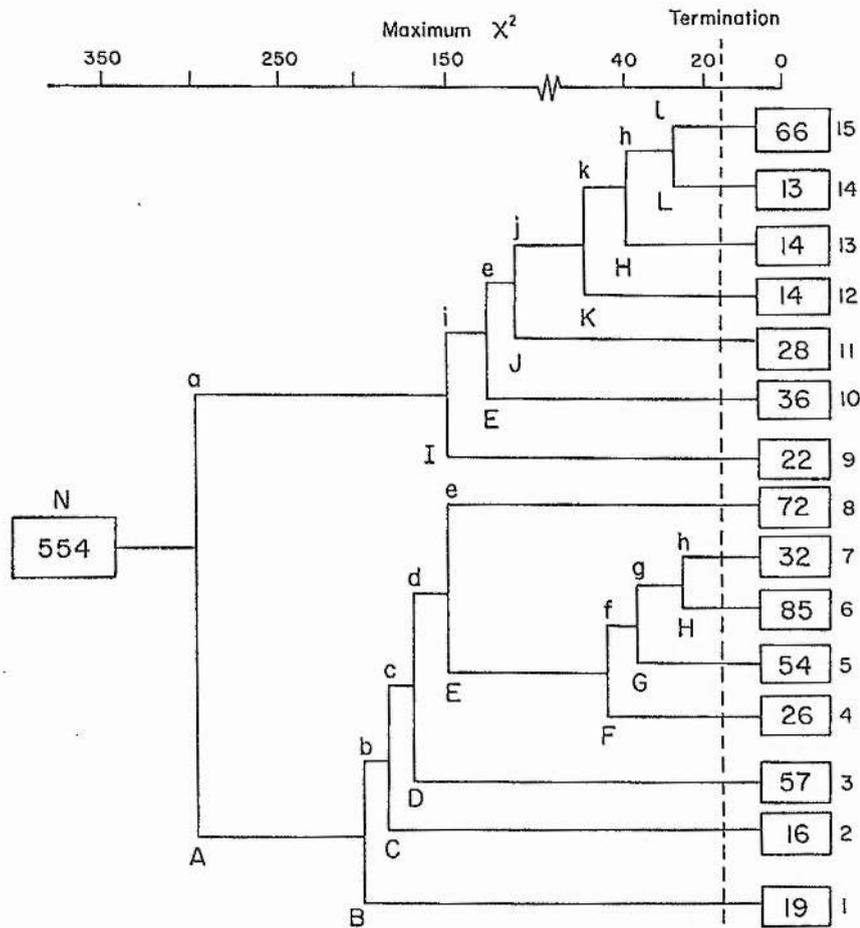


FIG. 1. Normal association analysis of wet land vegetation in the north of Arran based on 554 quadrats and ninety-eight species. The numbers enclosed in the boxes represent the number of quadrats in the terminating classes. The number outside the boxes are the reference numbers of the final sets. The letters at the nodes represent the species involved in each division of the quadrats. Capital letters denoted the presence of a species, lower case letter its absence. A, *Molinia caerulea*; B, *Anthoxanthum odoratum*; C, *Cirsium palustre*; D, *Rhacomitrium lanuginosum*; E, *Sphagnum* spp; F, *Juncus acutiflorus*; G, *Myrica gale*; H, *Calluna vulgaris*; I, *Iris pseudacorus*; J, *Glaux maritima*; K, *Sagina procumbens*; L, *Alnus glutinosa*.

different dune slack types were distinguished. A comparison of the sets obtained by group analysis is made with those obtained by association analysis and the results are recorded in Table 7. As with the Arran survey the salt tolerant type obtained by group analysis matches very closely with that obtained by association analysis. The other two slack types distinguished by group analysis, the mature *Erica tetralix*-*Filipendula ulmaria* slacks and the dry *Hieracium pilosella* slacks, represent only two of the remaining eight types distinguished by association analysis. These remaining types are, however,

variants of the two slack types above and it seems that there is some justification for regarding them as ecotone units both from floristic composition (see Table 6) and by comparing an ordination of the slack types segregated, as shown in Figs. 6 and 7. The ordination of slack types as obtained by association analysis from this area and given by Crawford & Wishart (1966) was based only on those species that had a presence of

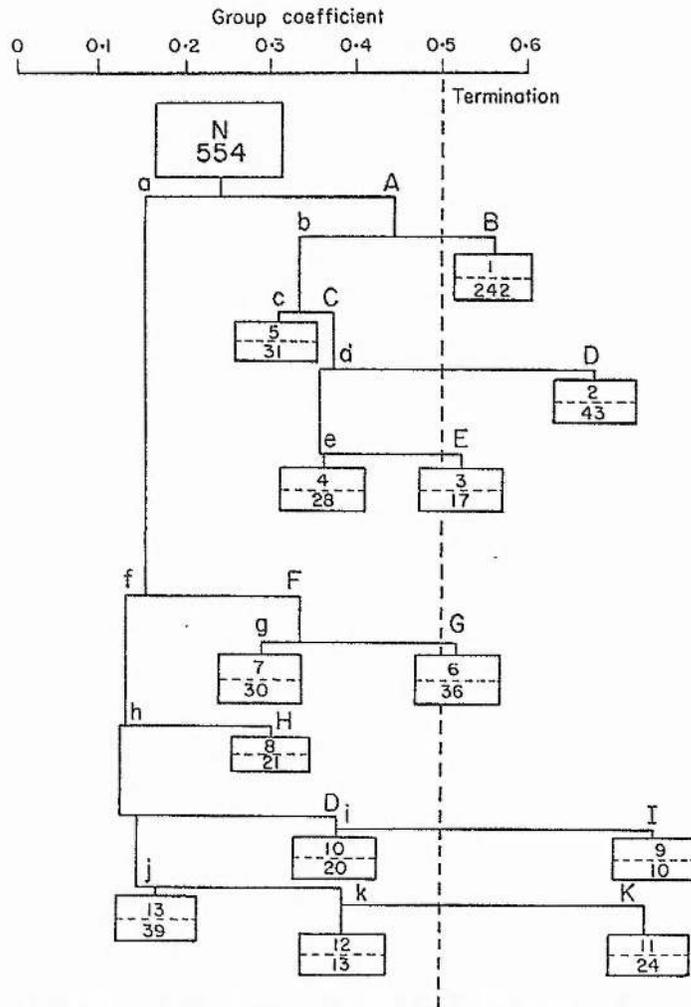


FIG. 2. Group analysis of wet land vegetation in the north of Arran based on 554 quadrats and ninety-eight species. The lower numbers enclosed in boxes represent the number of quadrats in the terminal groups, the upper numbers are the reference numbers of these groups. The letters at the nodes represent the presence of a species, lower case letters its absence. A, *Molinia caerulea*; B, *Erica tetralix*; C, *Potentilla erecta*; D, *Calluna vulgaris*; E, *Juncus acutiflorus*; F, *Anthoxanthum odoratum*; G, *Holcus lanatus*; H, *Rhynchospora alba*; I, *Drosera rotundifolia*; J, *Festuca rubra*; K, *Glaux maritima*.

20% or more. This limitation tends to accentuate the differences of the slacks and is therefore useful in following their sequence of development. However, for the present purpose the ordination diagrams in Figs. 6 and 7 are based on all the species that occurred in the quadrats. There appears from this ordination study further justification for considering the slack vegetation to be composed of three main types, as the relative similarity of some of the association analysis types is clearly illustrated. The ordination of the group analysis sets again illustrates the floristic distinctness of the community sets

Table 3. *The floristic composition of wet land vegetation types in the north of Arran as segrated by normal association analysis (only species with a presence of 33% or more are listed)*

| | % presence | | % presence |
|---------------------------------|------------|---------------------------------|------------|
| TYPE 1 | | TYPE 2 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Anthoxanthum odoratum</i> | 100 | <i>Cirsium palustre</i> | 100 |
| <i>Potentilla erecta</i> | 95 | <i>Potentilla erecta</i> | 75 |
| <i>Festuca ovina</i> | 53 | <i>Iris pseudacorus</i> | 62 |
| <i>Galium saxatile</i> | 53 | <i>Juncus acutiflorus</i> | 43 |
| <i>Pteridium aquilinum</i> | 47 | <i>Sphagnum</i> spp. | 43 |
| <i>Luzula multiflora</i> | 42 | | |
| <i>Carex echinata</i> | 37 | | |
| <i>C. panicea</i> | 37 | | |
| <i>Sphagnum</i> spp. | 37 | | |
| TYPE 3 | | TYPE 4 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Rhacomitrium lanuginosum</i> | 100 | <i>Sphagnum</i> spp. | 100 |
| <i>Calluna vulgaris</i> | 98 | <i>Juncus acutiflorus</i> | 100 |
| <i>Potentilla erecta</i> | 81 | <i>Potentilla erecta</i> | 88 |
| <i>Trichophorum cespitosum</i> | 81 | <i>Erica tetralix</i> | 77 |
| <i>Narthecium ossifragum</i> | 72 | <i>Carex echinata</i> | 58 |
| <i>Cladonia arbuscula</i> | 44 | <i>Eriophorum angustifolium</i> | 54 |
| <i>Erica tetralix</i> | 44 | <i>Narthecium ossifragum</i> | 54 |
| <i>Lycopodium selago</i> | 42 | <i>Trichophorum cespitosum</i> | 54 |
| <i>Pleurozia purpurea</i> | 40 | <i>Polygala serpyllifolia</i> | 50 |
| <i>Sphagnum</i> spp. | 40 | <i>Calluna vulgaris</i> | 38 |
| | | <i>Drosera rotundifolia</i> | 35 |
| TYPE 5 | | TYPE 6 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Sphagnum</i> spp. | 100 | <i>Sphagnum</i> spp. | 100 |
| <i>Myrica gale</i> | 100 | <i>Calluna vulgaris</i> | 100 |
| <i>Erica tetralix</i> | 87 | <i>Erica tetralix</i> | 84 |
| <i>Polygala serpyllifolia</i> | 61 | <i>Trichophorum cespitosum</i> | 78 |
| <i>Potentilla erecta</i> | 52 | <i>Potentilla erecta</i> | 66 |
| <i>Drosera rotundifolia</i> | 48 | <i>Eriophorum vaginatum</i> | 54 |
| <i>Trichophorum cespitosum</i> | 43 | <i>Narthecium ossifragum</i> | 52 |
| <i>Narthecium ossifragum</i> | 41 | <i>Drosera rotundifolia</i> | 47 |
| <i>Eriophorum angustifolium</i> | 39 | <i>Eriophorum angustifolium</i> | 47 |
| <i>Calluna vulgaris</i> | 37 | | |
| TYPE 7 | | TYPE 8 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Sphagnum</i> spp. | 100 | <i>Erica tetralix</i> | 71 |
| <i>Erica tetralix</i> | 75 | <i>Potentilla erecta</i> | 61 |
| <i>Potentilla erecta</i> | 59 | <i>Calluna vulgaris</i> | 57 |
| <i>Eriophorum angustifolium</i> | 56 | <i>Trichophorum cespitosum</i> | 43 |
| <i>E. vaginatum</i> | 56 | <i>Myrica gale</i> | 36 |
| <i>Narthecium ossifragum</i> | 47 | <i>Polygala serpyllifolia</i> | 35 |
| <i>Polygala serpyllifolia</i> | 34 | | |
| TYPE 9 | | TYPE 10 | |
| <i>Iris pseudacorus</i> | 100 | <i>Sphagnum</i> spp. | 100 |
| <i>Epilobium palustre</i> | 64 | <i>Erica tetralix</i> | 53 |
| <i>Mentha aquatica</i> | 64 | <i>Calluna vulgaris</i> | 50 |
| <i>Cirsium palustre</i> | 50 | <i>Trichophorum cespitosum</i> | 50 |
| <i>Juncus acutiflorus</i> | 50 | <i>Drosera rotundifolia</i> | 42 |
| <i>J. effusus</i> | 46 | <i>Narthecium ossifragum</i> | 39 |
| <i>Lotus pedunculatus</i> | 46 | <i>Potentilla erecta</i> | 39 |
| <i>Potentilla erecta</i> | 46 | | |
| <i>Ranunculus repens</i> | 40 | | |
| <i>Filipendula ulmaria</i> | 36 | | |
| <i>Hydrocotyle vulgaris</i> | 36 | | |

The classification of ecological groups

Table 3 (continued)

| | % presence | | % presence |
|---|---------------|-------------------------------|---------------|
| TYPE 11 | | TYPE 12 | |
| <i>Glaux maritima</i> | 100 | <i>Sagina procumbens</i> | 100 |
| <i>Festuca rubra</i> | 89 | <i>Trifolium repens</i> | 86 |
| <i>Plantago maritima</i> | 86 | <i>Festuca rubra</i> | 79 |
| <i>Armeria maritima</i> | 82 | <i>Cynosurus cristatus</i> | 50 |
| <i>Fucus vesiculosus</i> var. <i>muscooides</i> | 68 | <i>Rhytidadelphus loreus</i> | 50 |
| | | <i>Galium saxatile</i> | 43 |
| TYPE 13 | | <i>Potentilla anserina</i> | 43 |
| <i>Calluna vulgaris</i> | 100 | <i>Ranunculus repens</i> | 43 |
| <i>Potentilla erecta</i> | 79 | <i>Carex nigra</i> | 36 |
| <i>Erica tetralix</i> | 71 | <i>Cochlearia officinalis</i> | 36 |
| <i>Festuca ovina</i> | 57 | <i>Deschampsia flexuosa</i> | 36 |
| <i>Previdium aquilinum</i> | 43 | <i>Eleocharis palustris</i> | 36 |
| <i>Galium saxatile</i> | 36 | <i>Hydrocotyle vulgaris</i> | 36 |
| TYPE 15 | | TYPE 14 | |
| <i>Anthoxanthum odoratum</i> | 73 | <i>Alnus glutinosa</i> | 100 |
| <i>Conopodium majus</i> | 59 | <i>Conopodium majus</i> | 77 |
| <i>Holcus lanatus</i> | 55 | <i>Deschampsia cespitosa</i> | 77 |
| <i>Potentilla erecta</i> | 53 | <i>Luzula sylvatica</i> | 77 |
| <i>Rhytidadelphus loreus</i> | 47 | <i>Oxalis acetosella</i> | 77 |
| <i>Rumex acetosella</i> | 44 | <i>Anthoxanthum odoratum</i> | 54 |
| <i>Trifolium repens</i> | 44 | <i>Dactylis glomerata</i> | 46 |
| <i>Deschampsia flexuosa</i> | 42 | <i>Ranunculus repens</i> | 46 |
| <i>Galium saxatile</i> | 39 | <i>Endymion non-scriptus</i> | 39 |
| <i>Matricaria matricarioides</i> | 39 | | |
| <i>Poa pratensis</i> | 38 | | |

Table 4. The floristic composition of wet land vegetation types in the north of Arran as segregated by group analysis (only species with a presence of 33% or more are listed)

| | % presence | | % presence |
|---------------------------------|---------------|---------------------------------|---------------|
| TYPE 1 | | TYPE 2 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Erica tetralix</i> | 100 | <i>Calluna vulgaris</i> | 100 |
| <i>Sphagnum</i> spp. | 73 | <i>Potentilla erecta</i> | 100 |
| <i>Calluna vulgaris</i> | 66 | <i>Trichophorum cespitosum</i> | 79 |
| <i>Potentilla erecta</i> | 65 | <i>Rhacomitrium lanuginosum</i> | 58 |
| <i>Trichophorum cespitosum</i> | 60 | <i>Narthecium ossifragum</i> | 51 |
| <i>Narthecium ossifragum</i> | 52 | <i>Cladonia arbuscula</i> | 47 |
| <i>Polygala serpyllifolia</i> | 44 | <i>Sphagnum</i> spp. | 42 |
| <i>Eriophorum angustifolium</i> | 41 | <i>Lycopodium selago</i> | 34 |
| <i>Drosera rotundifolia</i> | 37 | | |
| TYPE 3 | | TYPE 4 | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 100 |
| <i>Juncus acutiflorus</i> | 100 | <i>Potentilla erecta</i> | 100 |
| <i>Potentilla erecta</i> | 100 | <i>Myrica gale</i> | 36 |
| <i>Sphagnum</i> spp. | 65 | | |
| <i>Carex echinata</i> | 53 | | |
| <i>Cirsium palustre</i> | 53 | | |
| <i>Viola palustre</i> | 47 | | |
| <i>Hylocomium splendens</i> | 41 | | |
| <i>Carex nigra</i> | 35 | | |
| <i>Hydrocotyle vulgaris</i> | 35 | | |

Table 4 (continued)

| | % presence | | % presence |
|---|---------------|----------------------------------|---------------|
| TYPE 5 | | TYPE 6 | |
| <i>Molinia caerulea</i> | 100 | <i>Anthoxanthum odoratum</i> | 100 |
| <i>Sphagnum</i> spp. | 61 | <i>Holcus lanatus</i> | 100 |
| <i>Calluna vulgaris</i> | 45 | <i>Conopodium majus</i> | 75 |
| | | <i>Potentilla erecta</i> | 67 |
| | | <i>Rhytidadelphus loreus</i> | 58 |
| | | <i>Galium saxatile</i> | 53 |
| | | <i>Rumex acetosella</i> | 53 |
| | | <i>Deschampsia flexuosa</i> | 50 |
| | | <i>Matricaria matricarioides</i> | 50 |
| | | <i>Luzula multiflora</i> | 44 |
| | | <i>Ranunculus repens</i> | 42 |
| | | <i>Poa pratensis</i> | 39 |
| | | <i>Rumex acetosa</i> | 39 |
| TYPE 7 | | TYPE 8 | |
| <i>Anthoxanthum odoratum</i> | 100 | <i>Rhytidadelphus loreus</i> | 100 |
| <i>Potentilla erecta</i> | 57 | <i>Festuca rubra</i> | 47 |
| <i>Conopodium majus</i> | 47 | <i>Galium saxatile</i> | 38 |
| | | <i>Potentilla erecta</i> | 38 |
| | | <i>Trifolium repens</i> | 38 |
| | | <i>Deschampsia cespitosa</i> | 33 |
| | | <i>Hydrocotyle vulgaris</i> | 33 |
| | | <i>Iris pseudacorus</i> | 33 |
| | | <i>Myosotis caespitosa</i> | 33 |
| TYPE 9 | | TYPE 10 | |
| <i>Calluna vulgaris</i> | 100 | <i>Calluna vulgaris</i> | 100 |
| <i>Drosera rotundifolia</i> | 100 | <i>Erica tetralix</i> | 65 |
| <i>Sphagnum</i> spp. | 100 | <i>Potentilla erecta</i> | 65 |
| <i>Erica tetralix</i> | 80 | <i>Sphagnum</i> spp. | 40 |
| <i>Trichophorum cespitosum</i> | 70 | <i>Festuca ovina</i> | 35 |
| <i>Narthecium ossifragum</i> | 60 | <i>Trichophorum cespitosum</i> | 35 |
| <i>Eriophorum vaginatum</i> | 40 | | |
| <i>Nardus stricta</i> | 40 | | |
| <i>Potentilla erecta</i> | 40 | | |
| TYPE 11 | | TYPE 12 | |
| <i>Festuca rubra</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Glaux maritima</i> | 100 | <i>Trifolium repens</i> | 65 |
| <i>Plantago maritima</i> | 91 | <i>Deschampsia flexuosa</i> | 54 |
| <i>Armeria maritima</i> | 83 | <i>Potentilla anserina</i> | 46 |
| <i>Fucus vesiculosus</i> var. <i>muscooides</i> | 71 | <i>Sagina procumbens</i> | 38 |
| TYPE 13 | | | |
| <i>Sphagnum</i> spp. | 33 | | |

($\phi = 0.5$) and the indeterminate nature of the ecotone sets. That only 22% of the quadrats are classified into significant sets in the Tentsmuir survey as compared with 67% in the Arran survey is perhaps not surprising as in this area the land has only accreted from the sea in the last 100 years (Crawford & Wishart 1966).

DISCUSSION

Group analysis is intended primarily as a rapid method of sorting ecological data irrespective of the size of the survey. As programmed for the IBM 1620, an analysis of the north Arran survey took by this method only 1% of the time required for an

association analysis of the same data. This is due to the manner in which the species correlations are calculated. In association analysis a survey of N species requires the calculation of approximately N^2 correlations whereas in group analysis N species requires only N 'correlations'. As Williams & Lambert (1960) have pointed out, the time necessary for computation in their studies increased linearly with the number of quadrats examined and in proportion to the square of the number of species. Therefore it is unavoidable

Table 5. *A comparison of the floristic similarities of the group and association analysis sets in the north Arran survey which have the highest coefficient of community (the coefficient of community between matching sets is calculated only on those species that have a presence of 33% or more)*

| Group analysis set | Matching association analysis set | Coefficient of community | Species occurring in both sets with 50% presence or more |
|--------------------|-----------------------------------|--------------------------|---|
| 1 | 7 | 78 | <i>Molinia caerulea</i> <i>Erica tetralix</i> <i>Sphagnum</i> spp. <i>Potentilla erecta</i> <i>Narthecium ossifragum</i> |
| 2 | 3 | 70 | <i>Molinia caerulea</i> <i>Calluna vulgaris</i> <i>Rhacomitrium lanuginosum</i> <i>Potentilla erecta</i> <i>Trichophorum cespitosum</i> <i>Narthecium ossifragum</i> |
| 3 | 4 | 45 | <i>Molinia caerulea</i> <i>Juncus acutiflorus</i> <i>Potentilla erecta</i> <i>Sphagnum</i> spp. <i>Carex echinata</i> |
| 6 | 15 | 88 | <i>Anthoxanthum odoratum</i> <i>Holcus lanatus</i> <i>Conopodium majus</i> <i>Potentilla erecta</i> |
| 9 | 6 | 84 | <i>Sphagnum</i> spp. <i>Erica tetralix</i> <i>Trichophorum cespitosum</i> <i>Narthecium ossifragum</i> |
| 11 | 11 | 100 | <i>Festuca rubra</i> <i>Glaux maritima</i> <i>Plantago maritima</i> <i>Armeria maritima</i> <i>Fucus vesiculosus</i> var. <i>muscooides</i> |

that as the survey becomes larger and species lists extend there comes a time when the method is no longer practicable. In group analysis the data are stored as species records per sample and the number of species in the survey has a negligible effect on the length of time required for computation. The time necessary for the analysis is determined solely by the number of samples used. In the north Arran survey there were 554 samples containing ninety-eight species and the analysis took 50 min. A survey, therefore, in which the whole of the British flora was contained (i.e. approximately 2000 species) would, if it were based on 10 000 samples, take approximately 15 h, which on a small computer such as an IBM 1620 would not be excessive.

Although fewer tests of interaction are performed in this method than correlations in association analysis the method is still multivariate as the effect of the presence or absence of each species on the set and non-set element potential is tested before each

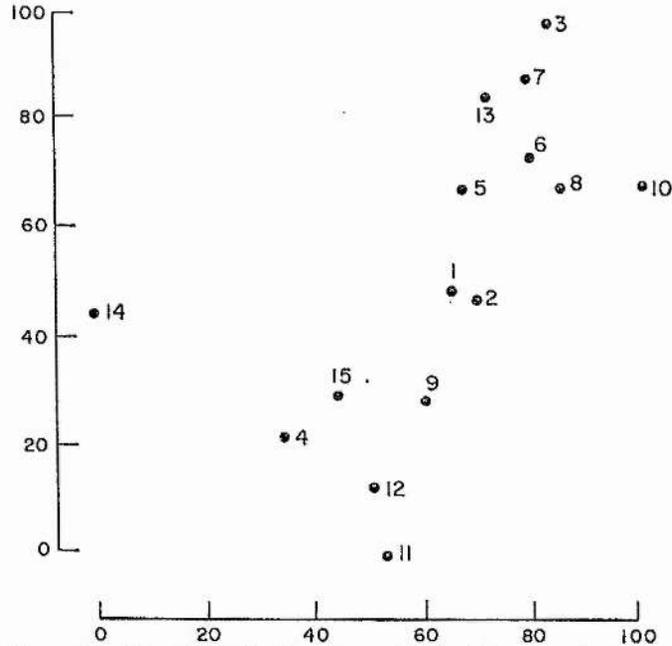


FIG. 3. Ordination of wet land vegetation types obtained by association analysis; x axis based on types 14 and 10, y axis on types 11 and 3.

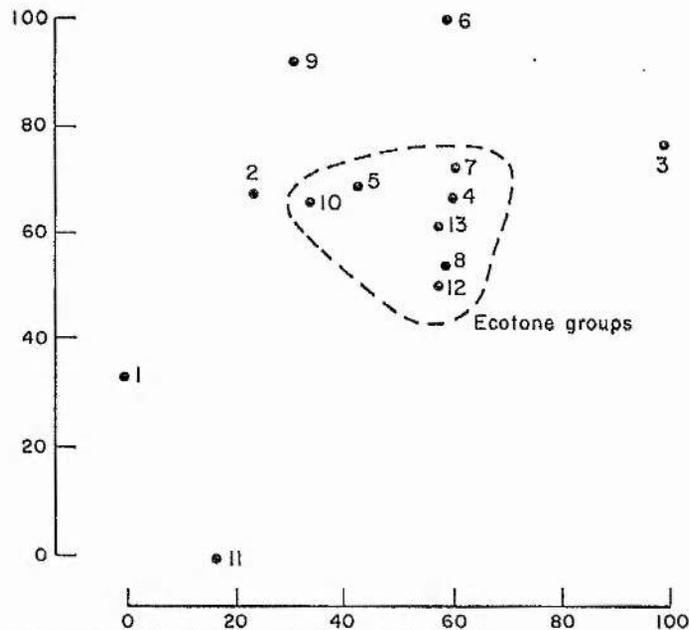


FIG. 4. Ordination of wet land vegetation types obtained by group analysis; x axis based on types 1 and 3, y axis on types 11 and 6.

and every division of the quadrats. The method of division (maximum μ'^2), however, is empirical, but is the most efficient found so far when judged by the increase in group coefficient ($C = \frac{1}{N} \sum S$) per division of the quadrats.

The classification of ecological groups

It is the definition of an ecological grouping which in this present work is most at variance with standard procedure in statistical ecology. Williams & Lambert (1959, 1960) search for communities as homogeneous groupings of quadrats. Although this homogeneity is only relative in the field, it is doubtful if the rigidity of hierarchical classification in any divisive monothetic system of data sorting can ever arrive at homogeneity. As has been pointed out by Gittins (1965), there is a tendency in association

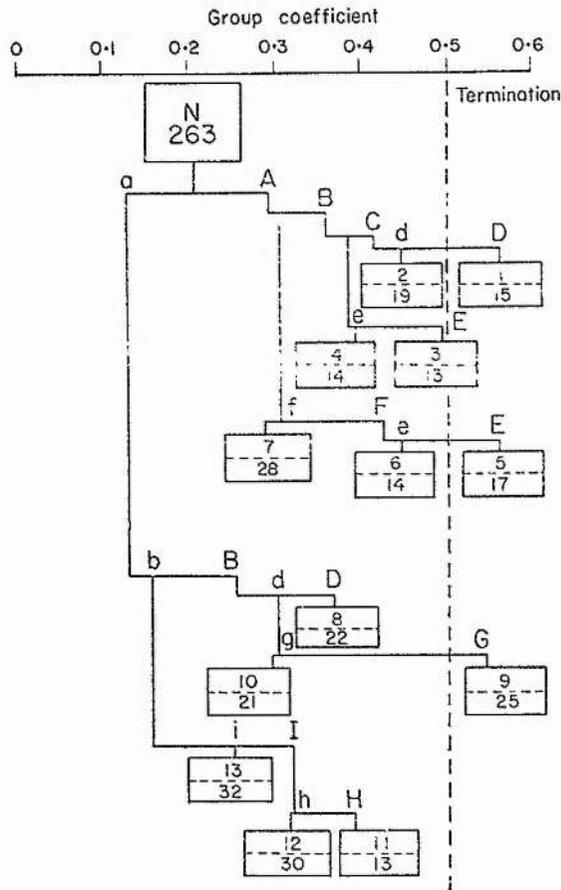


FIG. 5. Group analysis of Tentsmuir slacks based on 263 quadrats and 142 species. The lower numbers enclosed in boxes represent the number of quadrats in the terminal groups, the upper numbers are the reference numbers of these groups. The letters at the nodes represent the species involved in each division of the quadrats. Capital letters denote the presence of a species, lower case letters its absence. A, *Carex arenaria*; B, *Festuca rubra*; C, *Salix repens*; D, *Erica tetralix*; E, *Cladonia sylvatica*; F, *Hieracium pilosella*; G, *Honkenya peploides*; H, *Hydrocotyle vulgaris*; I, *Filipendula ulmaria*.

analysis for over-classification. In his study of calcareous grassland species Gittins found that association analysis produced twice as many meaningful groups as appeared justified by an ordination study. This situation is paralleled in this present investigation where in both north Arran and Tentsmuir more types are distinguished by association analysis than group analysis. This is particularly the case in the Tentsmuir survey where, as has been shown, the vegetation is changing rapidly and much of it could therefore be expected to be in a transitional stage. The process of searching for homogeneity where it does not exist may be the basic cause of this over-classification.

Table 6. *The floristic composition of dune slack types at Tentsmuir as segregated by group analysis (only species with a presence of 33% or more are listed)*

| | % presence | | % presence |
|----------------------------------|---------------|-----------------------------|---------------|
| TYPE 1 | | TYPE 2 | |
| <i>Carex arenaria</i> | 100 | <i>Carex arenaria</i> | 100 |
| <i>Erica tetralix</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Festuca rubra</i> | 100 | <i>Salix repens</i> | 100 |
| <i>Salix repens</i> | 100 | <i>Lotus corniculatus</i> | 61 |
| <i>Rhytidadelphus triquetrus</i> | 79 | <i>Hieracium pilosella</i> | 56 |
| <i>Filipendula ulmaria</i> | 71 | <i>Juncus balticus</i> | 50 |
| <i>Hylocomium splendens</i> | 57 | <i>Parnassia palustris</i> | 44 |
| <i>Cladonia sylvatica</i> | 50 | | |
| <i>Dicranum scoparium</i> | 42 | | |
| <i>Peltigera canina</i> | 43 | | |
| <i>Vicia lathyroides</i> | 36 | | |
| TYPE 3 | | TYPE 4 | |
| <i>Carex arenaria</i> | 100 | <i>Carex arenaria</i> | 100 |
| <i>Cladonia sylvatica</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Festuca rubra</i> | 100 | <i>Erica tetralix</i> | 46 |
| <i>Erica tetralix</i> | 62 | <i>Filipendula ulmaria</i> | 46 |
| <i>Dicranum scoparium</i> | 54 | <i>Ammophila arenaria</i> | 38 |
| <i>Ammophila arenaria</i> | 46 | <i>Galium palustre</i> | 38 |
| <i>Hieracium pilosella</i> | 46 | | |
| <i>Lotus corniculatus</i> | 38 | | |
| TYPE 5 | | TYPE 6 | |
| <i>Carex arenaria</i> | 100 | <i>Carex arenaria</i> | 100 |
| <i>Cladonia sylvatica</i> | 100 | <i>Hieracium pilosella</i> | 100 |
| <i>Hieracium pilosella</i> | 100 | <i>Lotus corniculatus</i> | 62 |
| <i>Ammophila arenaria</i> | 76 | <i>Salix repens</i> | 62 |
| <i>Thymus drucei</i> | 65 | <i>Ammophila arenaria</i> | 38 |
| <i>Acrocladium cuspidatum</i> | 47 | <i>Juncus balticus</i> | 38 |
| <i>Parmelia physodes</i> | 41 | <i>Parnassia palustris</i> | 38 |
| <i>Dicranum scoparium</i> | 35 | <i>Peltigera canina</i> | 38 |
| <i>Epilobium hirsutum</i> | 35 | | |
| TYPE 7 | | TYPE 8 | |
| <i>Carex arenaria</i> | 100 | <i>Erica tetralix</i> | 100 |
| <i>Salix repens</i> | 46 | <i>Festuca rubra</i> | 100 |
| <i>Lotus corniculatus</i> | 34 | <i>Cladonia sylvatica</i> | 59 |
| | | <i>Hylocomium splendens</i> | 50 |
| | | <i>Hypnum cupressiforme</i> | 50 |
| | | <i>Filipendula ulmaria</i> | 36 |
| | | <i>Salix repens</i> | 36 |
| TYPE 9 | | TYPE 10 | |
| <i>Festuca rubra</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Honkenya peploides</i> | 100 | <i>Salix repens</i> | 47 |
| <i>Rhinanthus minor</i> | 80 | <i>Filipendula ulmaria</i> | 38 |
| <i>Agrostis stolonifera</i> | 60 | | |
| <i>Juncus gerardii</i> | 56 | | |
| <i>Plantago maritima</i> | 44 | | |
| <i>Hieracium pilosella</i> | 40 | | |
| TYPE 11 | | TYPE 12 | |
| <i>Filipendula ulmaria</i> | 100 | <i>Filipendula ulmaria</i> | 100 |
| <i>Hydrocotyle vulgaris</i> | 100 | <i>Galium palustre</i> | 46 |
| <i>Lophocolea bidentata</i> | 38 | <i>Juncus effusus</i> | 46 |
| TYPE 13 | | | |
| <i>Hieracium pilosella</i> | 46 | | |
| <i>Juncus balticus</i> | 34 | | |
| <i>Lotus corniculatus</i> | 34 | | |
| <i>Salix repens</i> | 34 | | |

The parameter used here as the basis of correlation $[(P_x, V_x) / \bar{V}]$ is similar to $[2w / (a+b) \times 100]$ in that the data is continuous (in contrast to χ^2/N) and is a measure of similarity. However, in this case, it is not calculated between stands or between species,

Table 7. A comparison of the floristic similarities of the group and association analyses sets in the dune slack survey at Tentsmuir which have the highest coefficient of community (the coefficient of community between matching sets is calculated only on those species that have a presence of 33% or more)

| Group analysis set | Matching association analysis set | Coefficient of community | Species occurring in both sets with 50% presence or more |
|--------------------|-----------------------------------|--------------------------|---|
| 6 | 8 | 62 | <i>Lotus corniculatus</i> <i>Salix repens</i> <i>Carex arenaria</i> |
| 1 | 6 | 76 | <i>Erica tetralix</i> <i>Festuca rubra</i> <i>Hylocomium splendens</i> <i>Carex arenaria</i> |
| 9 | 1 | 93 | <i>Honkenya peploides</i> <i>Festuca rubra</i> <i>Rhinanthus minor</i> <i>Agrostis stolonifera</i> |

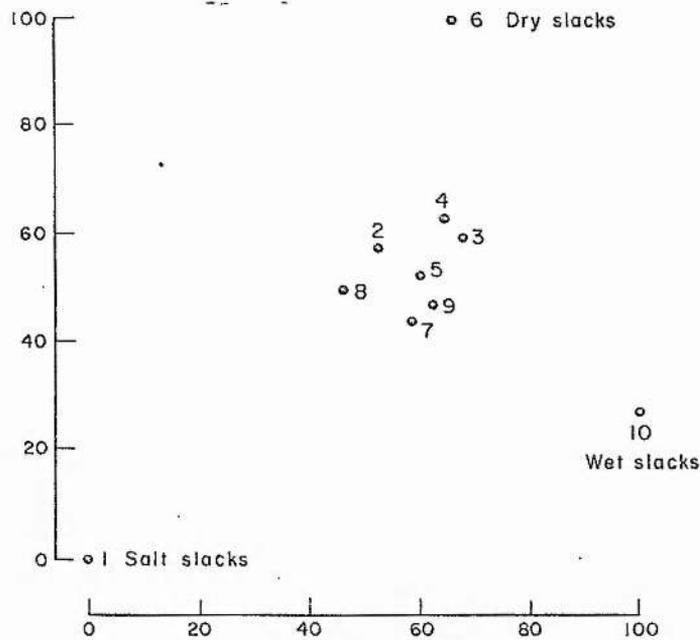


FIG. 6. Ordination of slack types obtained by association analysis; x axis based on types 1 and 10, y axis on types 1 and 6.

but between each species in turn and the entire sample population. Further this statistic is made absolute by division by \bar{V} (the mean sample density for all quadrats) so the significance of the statistic can be used not only for the classification of the vegetation concerned but for a comparison of the degree of sociability (as determined by species frequency and gregariousness) between groupings in different areas. As has been shown

with the Tentsmuir data the degree of sociability is much less in this area of rapid accretion than in the presumably more stable mires and bogs of north Arran.

The value ϕ (the arbitrary group coefficient value used to terminate the division) can be varied to suit the needs of the investigation. In the north Arran survey the first group to be segregated (group 1) contained 44% of all the quadrats sampled. This wet heath vegetation of *Molinia caerulea* and *Erica tetralix* was the most widespread type encountered on the Island and if further detail were wanted of its composition it would only be necessary to raise the value of ϕ from 0.5 to 0.6.

As the analysis involves the computation of species and quadrat attributes of a continuous nature it is evident that this method lends itself to ordination studies. By coupling

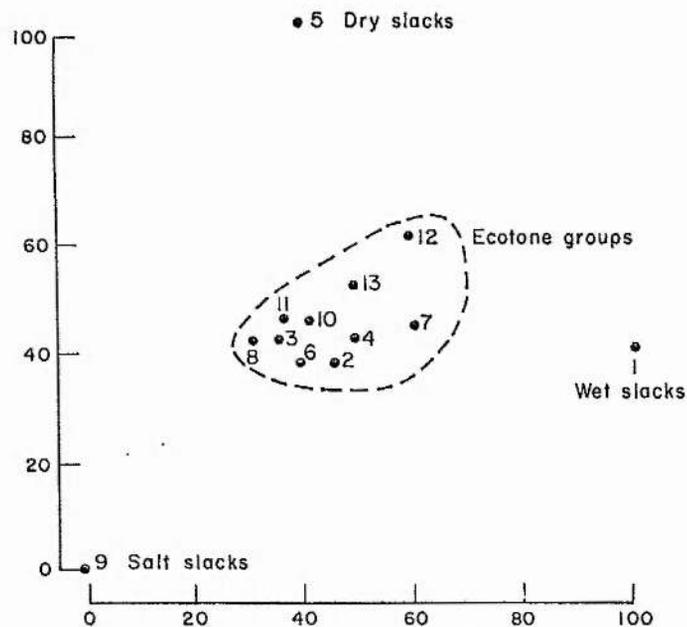


FIG. 7. Ordination of slacks obtained by group analysis; x axis based on types 9 and 1, y axis on types 9 and 5.

an on-line digital plotter to the computer it is possible to view the data in a number of different ways with great facility. If the quadrats have been taken as a grid the groups may be mapped. As a result of the evaluation of quadrat attributes for each classification unit it is possible to prepare maps showing the distribution of each ecological grouping at different group coefficient levels, thus avoiding the necessity of drawing lines on a map representing vegetational boundaries that do not exist in the field. Instead, these absolute boundaries are replaced by lines representing the probability of the area belonging to any particular ecological group.

It must be emphasized that although the distribution of the quadrat sets can be mapped directly this is not what the method is intended to do and will only yield a minimum of information. For mapping purposes it is the quadrat attributes that should be used. It is therefore possible that any one quadrat will have attributes that relate it to more than one vegetation type as determined by the process of classification. This will be particularly the case if the quadrats are large. It is hoped to illustrate the application of direct digital plotting in vegetation mapping and stand ordination in a subsequent communication.

ACKNOWLEDGMENTS

We are much indebted to Mr J. R. Gray and Mr R. L. Constable for statistical advice and to Miss H. Crane, Miss M. C. MacNaughtan, Mr R. M. Campbell, and the Honours Botany Class, 1966, for assistance in collecting the data.

SUMMARY

A method is suggested for the rapid analysis of large ecological surveys by computers with limited high speed storage. Coincidence of occurrence rather than homogeneity is taken as the fundamental property of ecological groupings. The species and the quadrats that contain them are classified according to their similarity with the population sample as a whole or with a sub-set of this population.

Using the statistic $\mu'^2 = \Sigma(o_i - e_i)^2$ a test is made of species interaction on the group properties of the population sample and on each sub-set. The quadrats are then divided depending on whether or not they contain the species with the highest interaction. As each species is tested for its interaction value the method is multivariate. The process is repeated on the sub-sets and a hierarchical division is made of the data.

A statistic is used to define an ecological group and quadrat sets that fail to reach the desired level of significance are classified as ecotone or transitional units. As the method is based on the calculation of species and quadrat attributes of a continuous nature the method lends itself to ordination studies as well as to classification.

The analysis system, written in Fortran II D, is available on application to the authors.

REFERENCES

- Austin, M. P. & Orloci, L. (1966). Geometric models in ecology. II. An evaluation of some ordination techniques. *J. Ecol.* 54, 217-27.
- Bray, J. R. & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325-49.
- Crawford, R. M. M. & Wishart, D. (1966). A multivariate analysis of the development of dune slack vegetation in relation to coastal accretion at Tentsmuir, Fife. *J. Ecol.* 54, 729-43.
- Gittins, R. (1965). Multivariate approaches to a limestone grassland community. III. A comparative study of ordination and association analysis. *J. Ecol.* 53, 411-25.
- Greig-Smith, P. (1961). Ecological terminology. *Encyclopedia of Biological Science* (Ed. by P. Gray), pp. 322-4. New York.
- Greig-Smith, P. (1964). *Quantitative Plant Ecology*, 2nd edn. London.
- Ivimey-Cook, R. B. & Proctor, M. C. F. (1966). The application of association analysis to phytosociology. *J. Ecol.* 54, 179-92.
- Williams, W. T. & Lambert, J. M. (1959). Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* 47, 83-101.
- Williams, W. T. & Lambert, J. M. (1960). Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. *J. Ecol.* 48, 689-710.
- Williams W. T. & Lambert, J. M. (1961). Multivariate methods in plant ecology. III. Inverse association-analysis. *J. Ecol.* 49, 717-29.

(Received 6 December 1966)

A RAPID CLASSIFICATION AND ORDINATION METHOD AND ITS APPLICATION TO VEGETATION MAPPING

BY R. M. M. CRAWFORD AND D. WISHART

*Department of Botany and Computing Laboratory,
The University, St Andrews*

INTRODUCTION

In a previous communication (Crawford & Wishart 1967) a rapid multivariate method was described for the classification of ecological data by a monothetic divisive process. The method differed from standard numerical taxonomy techniques in that it was designed to detect sets of quadrats in terms of groups of co-incident species and not, as is more usual, in terms of quadrat homogeneity. The stopping rule applied therefore (see Macnaughton-Smith 1965) was determined by the degree of co-occurrence between species and not the attaining of a set level of similarity or absence of dissimilarity, as in most agglomerative and divisive methods (see Sokal & Sneath 1963; Williams & Dale 1965). Owing to the manner in which the species group correlations are carried out the method is very rapid even with large surveys and is relatively unaffected by the number of species in the survey. The time necessary for analysis is dependent solely on the number of samples to be analysed and increases linearly with the sample number. However, in common with all other monothetic divisive methods, no indication is obtained of the relationships between the various terminal groups; and further there is always the danger of misclassification due to the chance occurrence, or erroneous diagnosis of a dividing species.

This paper describes firstly, a rapid agglomerative method that can be used after the initial divisive process to check for any misclassifications, and secondly, a means of representing the variance both within and between the terminal groups by an ordination procedure. The need in ecological surveys for such a representation of the results of classification is demonstrated by the present dichotomy in the use of ordination and classification methods. A review of the problems involved in this divergence is given by Greig-Smith (1964). The recently devised polythetic agglomerative methods of Jancey (1966) and Orloci (1967) certainly give a solution to these problems, in that the ordination of the group centroids displays the relative distances between the resulting groups. However, as developed at present these methods are difficult to use with large surveys owing to the quadratic relationship between computation time and sample number. In an attempt to overcome this difficulty the methods described in this paper achieve the economy in core storage which enables them to be used with large surveys by applying the same strategy as that used in the previous communication. The basis of this approach was to look for the occurrence of groups produced by dividing the set of quadrats on the presence or absence of each species in turn. Thus each division of the data involved only N tests, where N^2 would have been necessary using existing methods.

Thirdly, a method is presented for mapping vegetation that has been sampled on a grid. Maps of distribution of vegetation types usually involve the drawing of lines on paper that may or may not represent a discrete vegetation boundary in the field. In this present study the conventional mapping of several discrete vegetation types is replaced by a

method which assesses the potential of any quadrat for membership of any classified type. Because the classification process is based on the grouping potentials of species, from which are calculated the group potentials of the quadrats that contain them, it is possible to retrieve these individual quadrat potentials for the purposes of map making. Thus, for each group it is possible to draw contours delimiting areas of equal potential with respect to each vegetation type and thus obtain a map of varying group potential rather than discrete vegetation-type boundaries.

To avoid the tedium that is often associated with vegetation ordination and mapping studies, all the methods used in this paper have been devised for direct plotting by computer with the use of an on-line digital plotter.

Finally, the use of this method is explained for the identification and classification of further samples in the field without resort to additional computation.

METHODS

Classification

In the monothetic divisive method already described (Crawford & Wishart 1967), the mean grouping potential of a set of quadrats (SEP values) determined from species coincidence formed the basis of the classification. As these mean values are obtained from individual quadrat values it follows that it is possible to make a geometric representation of the affinities between these groups and also of the variation within the sets, in terms of their set element potentials (SEP values).

Consider a survey in which M quadrats are sampled for the presence or absence of N species. Thus the i th quadrat may be represented by a point in N dimensional space with co-ordinates

$$\alpha_i = (\alpha_{i1}, \alpha_{i2} \dots \alpha_{ij} \dots \alpha_{iN})$$

where

$$\alpha_{ij} = 1 \text{ if quadrat } i \text{ possesses species } j$$

and

$$\alpha_{ij} = 0 \text{ if quadrat } i \text{ lacks species } j.$$

The number of species possessed by the i th object

$$V_i = \sum_{j=1}^N \alpha_{ij}$$

is called the quadrat density. For a subset of M quadrats every point α_i is weighted according to the density V_i of the corresponding quadrat, and the centroid of the resulting array of weights,

$$W = \left(\frac{1}{M} \sum_{i=1}^M \alpha_{i1} V_i, \frac{1}{M} \sum_{i=1}^M \alpha_{i2} V_i, \dots, \frac{1}{M} \sum_{i=1}^M \alpha_{iN} V_i \right) \quad (1)$$

is called the characteristic vector of the subset. This is identical with the calculation of the product statistic W'_x used in the divisive method.

The relationship between each quadrat contained in the subset and the subset's overall characteristics is considered, and provides the criterion of similarity between the quadrat and the subset as a whole. Thus a generalized approach of this type will yield a much faster analysis than the more widely accepted strategy by which similarities are considered for all pairs of quadrats. For this analysis the similarity function is obtained by measuring the distance from the origin which the i th object projects onto the characteristic vector.

The scalar product of two vectors is defined as

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| \times |\mathbf{b}| \cos \theta$$

where θ is the angle between the vectors \mathbf{a} and \mathbf{b} (see Fig. 1) and since $OA \cos \theta = OA'$

$$\mathbf{a} \cdot \mathbf{b} = OA' \times OB$$

Hence $OA' = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|}$ or the projection of α_i onto \mathbf{W} is given by

$$\alpha'_i = \frac{\alpha_i \cdot \mathbf{W}}{|\mathbf{W}|} = \frac{1}{|\mathbf{W}|} \sum_{j=1}^N \alpha_{ij} \omega_j \quad (2)$$

In order to obtain an absolute value for the similarity function so that characteristic vectors of different length can be compared, the maximum projection that can be

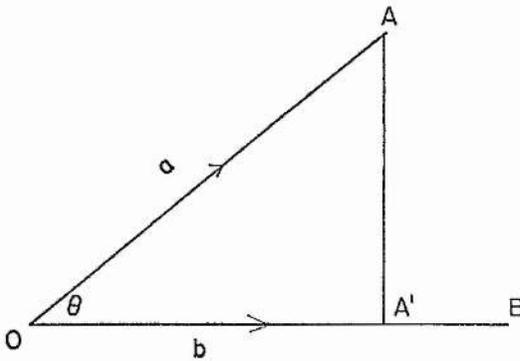


FIG. 1. See text.

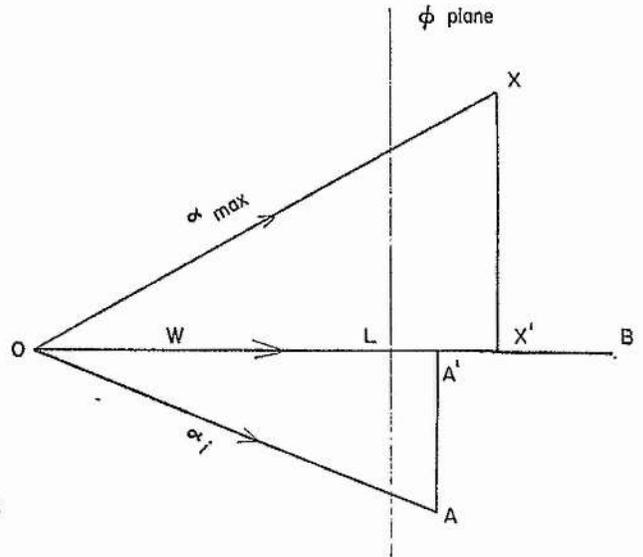


FIG. 2. See text.

obtained along \mathbf{W} is considered. This again is only a geometric representation of the value \mathbf{W} (the absolute set element potential as described in the previous communication). This absolute value is obtained from (2) when $\sum_{j=1}^N \alpha_{ij} \omega_j$ is a maximum for quadrat α_i with $\alpha_{ij} = 1$ for each non-zero ω_j . Alternatively when $\alpha_{ij} = 1$ for all j , the same result is derived, namely,

$$\alpha'_{\max} = \frac{1}{|\mathbf{W}|} \sum_{j=1}^N \omega_j \quad (3)$$

and the quadrat potential is defined as

$$S_i = \frac{\alpha'_i}{\alpha'_{\max}} = \frac{OA'}{OX'} \quad (\text{see Fig. 2}) \quad (4)$$

where OX is the optimum quadrat vector with $\alpha_{ij} = 1$ for all j .

Hence from (2) and (3), (4) becomes

$$S_i = \frac{\sum_{j=1}^N \alpha_{ij} \omega_j}{\sum_{j=1}^N \omega_j} \quad (5)$$

When a limiting value of the potential of a quadrat is selected, say $S_i = \phi$, we have

$$\phi = \frac{OL}{OX'}$$

and clearly any point α_i which lies in or beyond the plane through L which is orthogonal to W will satisfy

$$S_i \geq \phi$$

The plane is referred to as the ϕ plane for W, and the meaning of a ϕ -cluster for W is defined, as the set of quadrats $\{\alpha_i\}$ such that $S_i \geq \phi$. The cluster may be thought of as the area in N space which is bounded by the ϕ plane and the N -dimensional cuboid of side 1 (all points with the earlier definition of co-ordinates lie at vertices of the cuboid of side 1 that forms the positive quadrant of the space).

In the earlier communication (Crawford & Wishart 1967) a more empirical approach to this concept is discussed and a method introduced which used the interaction statistic μ'^2 in a monothetic divisive strategy to derive approximate ϕ clusters. The group element potential (GEP) of a species by this method is proportional to the j th coefficient of the characteristic vector and the set element potential (SEP) corresponds to the quadrat potential used here.

The divisive strategy resolves groups of quadrats for which the membership of a group is defined by a vector of conditions (e.g. A, b, defines the set of quadrats which possess species A but lack species B). In the case where the characteristics of a group are determined by the presence of a large number of species, such generalizations will undoubtedly lead to the misclassification of quadrats which although they have the overall characteristics of the group fail one of the conditions.

In order to correct the temporary classification obtained by the divisive strategy, one approach might be to consider the potential obtained by each quadrat α_i , with respect to every characteristic vector (W_k), $k = 1, K$, select the highest S_{ik} and reclassify the i th quadrat with the provisional group K . Such an approach would however be inefficient since a quadrat with a high potential value with respect to any parent group would normally be reclassified into the same group. It seems therefore only necessary to consider for reclassification those quadrats which appear to be misclassified, namely the quadrats which have a low potential value with respect to their parent groups. Such quadrats are termed misfit quadrats defined at a level β as those quadrats (S_i) which have a potential $S_i < \beta$ with respect to their parent groups. This is equivalent to defining a β plane orthogonal to each characteristic vector and parallel to the corresponding ϕ plane. The misfit quadrats will be those whose points lie on the origin side of the β plane which corresponds to the quadrat's parent group. The potential of each misfit quadrat is calculated with respect to all the characteristic vectors and the quadrat is reclassified according to the group whose characteristic vector yields the highest potential value. (This value need not necessarily be greater than β and the misfit quadrat need not necessarily switch to a new group.)

When all the misfit quadrats have been reclassified, a better classification will have been obtained and the identities of the groups will have been altered. It is therefore necessary to recalculate the characteristic vectors of groups that have been modified, and update the potential values of the quadrats involved.

The modification of the characteristic vectors and the changes in potential values which will result will reveal new misfit quadrats and the procedure will return, to attempt a

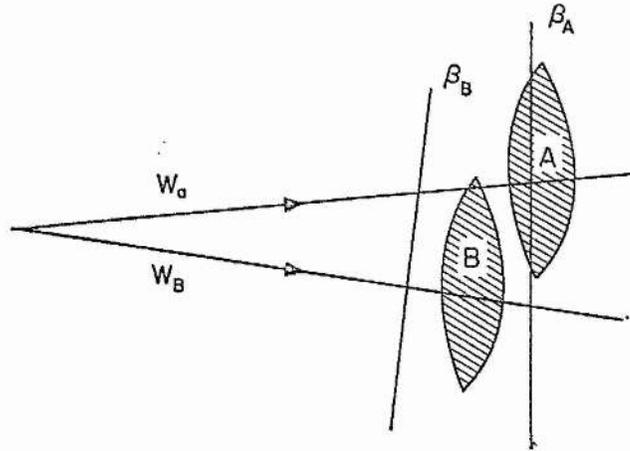


FIG. 3. See text.

reclassification where appropriate. This iterative procedure continues until either the degree of accuracy which is demanded is achieved or else no further misfit quadrats change parent groups and the convergence is complete. In order to satisfy the conditions imposed by the divisive strategy on the size of the resultant groups, it is necessary to examine group size at each iteration and absorb the quadrats from any group whose size is less than the limit L . This lower limit of group size (usually ten quadrats) is always set

Table 1. *Summary of agglomerative corrective analyses as carried out on the North Arran survey using different levels of β (level of testing for misfits—see text)*

| Value of β | No. of final clusters obtained | No. of iterations required for convergence | Time taken on IBM 1620 (min) |
|------------------|--------------------------------|--|------------------------------|
| 0.20 | 12 | 4 | 42 |
| 0.25 | 12 | 5 | 55 |
| 0.35 | 10 | 10 | 108 |
| 0.45 | 10 | 8 | 86 |
| 0.55 | 9 | 30 | 165 |

in order to avoid the production of trivial groups that fail to attain the desired level of species coincidence.

This may be achieved by setting all the potential values of the quadrats involved equal to zero so that reclassification occurs. The elimination of small subsets is an important feature of the corrective procedure since it counteracts any overclassification that might be caused by using a high level of ϕ at the divisive stage. Suppose a classifiable group is subdivided into two subsets A and B (see Fig. 3) whose β planes are nearly parallel. The

characteristic vectors W_A and W_B will be nearly coincident and therefore similar. Clearly, at the first iteration, misfit quadrats from A will be reclassified with B and the centroid and β plane of A will move away from the origin revealing new misfit quadrats. The result is that in successive iterations the quadrats in A (Fig. 3) will be eaten away and absorbed by B until eventually subset A will be eliminated and the overclassification corrected. It follows that the higher the value of β the faster this process will be and the greater the number of spurious subsets that will be eliminated. The procedure was tested exhaustively on three surveys, and the results of the largest survey (554 quadrats) are summarized in Table 1.

Ordination

Orloci (1966) and Austin & Orloci (1966) described a method in which a principal components analysis was applied to an $M \times N$ data matrix in order to obtain an efficient graphical representation of ecological structure. It was suggested that the spatial representation of M or N dimensions for Q or R type analyses should be chosen according to the minimum of (M, N) . Suppose $P = \text{Min}(M, N)$, then the method of principal components applied to this situation may be summarized as follows (a detailed approach can be obtained in Kendall (1957)):

- (a) construct a $P \times P$ correlation matrix R ,
- (b) calculate the two principal roots (or eigen values) λ_1, λ_2 , from the characteristic equation $(R - \lambda I) = 0$,
- (c) obtain the two principal eigen vectors E_1, E_2 , which correspond to λ_1, λ_2 ,
- (d) generate for each quadrat (or species) cartesian co-ordinates

$$\sum_{j=1}^P e_{1j} \alpha_{ij}, \quad \sum_{j=1}^P e_{2j} \alpha_{ij}$$

(e_{1j} is the j th element in E_1),

- (e) the percentage of the variance explained by the two principal components will be

$$\varepsilon = \frac{1}{P} (\lambda_1 + \lambda_2).$$

In order to obtain a graphical representation of group relationships the P -space points are projected onto a plane through the space and the resultant graph has orthogonal axes which are parallel to this plane. The variance of the points in P space Φ^2 is a measure of the distribution with respect to the mean position, and the corresponding variance σ^2 of the points on the graph measures the spread of the displayed distribution. Clearly the projection of the points onto the plane produces a distortion of their original orientation and the efficiency of the graphical display can be measured by the ratio σ^2/Φ^2 . The method of principal components maximizes this ratio by:

- (1) finding the line of 'best fit' through the points (this maximizes the variance 'explained' by the first principal axis);
- (2) obtaining the line of next best fit from the family of lines orthogonal to the first axis, and thus providing the other cartesian axis of the plane.

The efficiency ratio ε is equivalent to the percentage of variance which is explained by the plane, and if ε is high it can be assumed that the resultant display is a reasonably good representation of the P -space point structure.

However, if P is large (i.e. if M and N are both large) then considerable core storage and machine time are required for the calculation of the correlation matrix R , eigen values, and eigen vectors. It is suggested therefore that an initial reduction of the N space is obtained using the classification method previously described before a principal components analysis is attempted.

Consider the vector of potentials

$$S_i = (S_{i1}, \dots, S_{ik}, \dots, S_{iK})$$

where S_{ik} is the potential of the i th quadrat α_i with respect to the k th characteristic vector W_k for the k th group.

Then the distance between the quadrats α_A, α_B in the N space is

$$d^2_N = \sum_{j=1}^N (\alpha_{Aj} - \alpha_{Bj})^2$$

and in the K space in which the quadrats are represented by their vectors S_A, S_B

$$d^2_K = \sum_{k=1}^K (S_{Ak} - S_{Bk})^2$$

$$= \sum_{k=1}^K \left[\frac{\sum_{j=1}^N \omega_{kj} (\alpha_{Aj} - \alpha_{Bj})}{\sum_{j=1}^N \omega_{kj}} \right]^2$$

then it is apparent that the j th species difference for the two quadrats $(\alpha_{Aj} - \alpha_{Bj})$ is weighted according to the j th co-ordinate of each characteristic vector. Hence the inter-quadrat relationships are biased towards the non-trivial species and the K -space representations of the quadrats indicate the quadrat relationships with the groups previously obtained.

It must be stressed that this is not regarded as a projection of the original vectors, but rather as a subjective mapping of the space (N) into (K).

The $N \times K$ matrix of potentials S is first standardized by reducing each column vector to zero mean and unit variance, in order that the origin of the co-ordinates be located at the centroid of the point distribution. Principal components analysis as previously described is then applied to the K -space swarm structure, and the resultant plane of best fit is used as base graph. The points on this graph which correspond to those quadrats of a particular group k can be compared with the points obtained from another group, since the within group heterogeneity and between group homogeneity is demonstrated by the spread and affinity of the points. This is shown symbolically by plotting a circle for each group whose centre is the mean position of the group's points and radius the standard deviation of the points' radii from the group mean. It follows that provided the efficiency ratio is reasonably high, the size of each circle provides an indication of the heterogeneity of the corresponding group while the distance between the circles shows the group's mutual homogeneity.

Geographical mapping

If the quadrats have been sampled systematically in a rectangular grid as in the *Calluna* heath studied by Williams & Lambert (1959) it is possible to mark on a graph, for each quadrat, the number of the group with which it is classified. When applied to the present

method, this yields an indication of the location of the groups on the grid, but is not necessarily a precise mapping of the extent of each ecological group. It has been shown that certain quadrats may have high potential values with respect to two or more groups. These may be regarded as quadrats that exist on the border between two communities or in the region where two communities overlap. However, since they are classified into one group, a discrete map of the type described will yield no indication of overlapping. Suppose that for group k the potential values S_{ik} with respect to k for all the quadrats are marked on the grid. The high potential values correspond to the quadrats in which the ecological type is dominant, while low potential values indicate regions where the community is absent. If a level of potential γ is chosen to determine the significant community regions, then those quadrats that have $S_{ik} \geq \gamma$ can be marked as community regions. Furthermore, when a boundary between adjacent potential values $S_{ik} > \gamma$ and $S'_{ik} < \gamma$ exists it is possible to extrapolate for the approximate location of the potential value $S = \gamma$ and draw on the grid the equipotential contour $S = \gamma$.

If a set of K maps are obtained, one for each group, then the degree to which different ecological groups overlap provides a visual representation of the physical homogeneity of the region. When $\gamma = \beta$, the value used as a significance level in the agglomerative analysis, then the amount of overlapping will also indicate the homogeneity of the classification.

By making use of a contouring programme and on-line digital plotter it is possible to draw with great facility the equipotential contours at any desired levels of γ for each of the K set of groups in the survey.

Computer programmes

The programmes used in this paper were written in Fortran IID and are available on application to the authors.

DATA

The data used in this study comprise three separate surveys, two of which, the North Arran wet land survey and the Tentsmuir dune slack survey, were used in the previous communication. The third survey, a grid transect 138 m long and 12 m wide was laid out at the Nature Conservancy's reserve at Tentsmuir, Fife (GR NO 502268). This grid ran from an *Erica tetralix** slack type to a *Filipendula ulmaria* slack as defined by an association analysis carried out on all the slack regions at Tentsmuir after the manner of Williams & Lambert (1959) and described by Crawford & Wishart (1966). The transect followed a gradient of increasing wetness (see Phot. 1) and observations were made of water table fluctuations at monthly intervals over a period of 2 years. Sampling by metre square quadrats was carried out at 3 m intervals along the length and breadth of the grid giving 230 samples containing a total of sixty-three species.

RESULTS

North Arran survey

Table 3 records the floristic composition of the wet land vegetation types before and after the application of the agglomerative process. For brevity, the four most frequent species only are recorded in each case. When the agglomeration is carried out at a level of

* Nomenclature follows Clapham, Tutin & Warburg (1962) and Richards & Wallace (1950).

Table 2. Percentage variance accounted for by the first two vectors of the principal components analysis of the North Arran survey, with changing values for the agglomerative corrective factor β

| | β | | | |
|-------------|---------|------|------|------|
| | 0.25 | 0.35 | 0.45 | 0.55 |
| Component 1 | 48.1 | 47.9 | 45.8 | 43.9 |
| Component 2 | 28.0 | 27.4 | 26.2 | 27.5 |
| Total | 76.1 | 75.3 | 72.0 | 71.4 |

$\beta = 0.55$, this resulted in thirty iterations (see Table 2) and a convergence that reduced the number of vegetation types to 9. Three of the groups which disappeared, namely 4, 8 and 13, were all considered as ecotone types with low coefficients of community. The coefficient of community as defined in the previous communication assesses the degree of species coincidence. When this value is low, the lack of species coincidences is taken as indicating the group has no definite ecological standing and is considered as a transitional or ecotone group.

Type 11, which was the salt vegetation type, converged on type 12 which now becomes a more general coastal grass area with *Festuca rubra* present in 93% of the samples and the

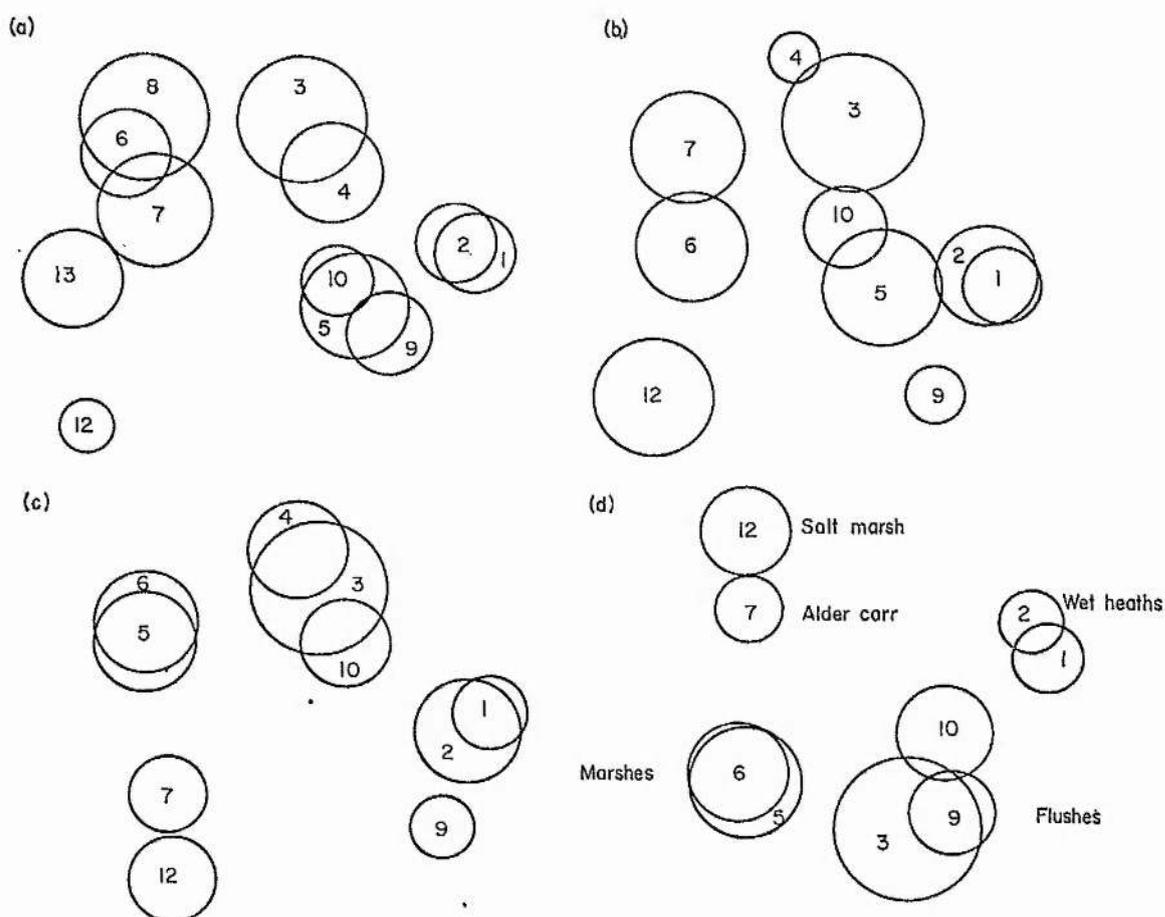


FIG. 4 (a-d). Ordination of groups obtained in the North Arran survey after agglomeration at different levels of β (β = level of testing for misfits—see text). (a) $\beta = 0.25$; (b) $\beta = 0.35$; (c) $\beta = 0.45$; (d) $\beta = 0.55$. (Ordination drawn by computer.)

Table 3. The effect of the agglomerative corrective process on the group composition of wet land vegetation types in the North Arran survey as seen in the percentage occurrence of the four most frequent species in each vegetation type

| No agglomerative correction | | Agglomerated ($\beta = 0.55$) | |
|--------------------------------|---------------|---------------------------------|---------------|
| | % presence | | % presence |
| TYPE 1 | | | |
| <i>Molinia caerulea</i> | 100 | <i>Molinia caerulea</i> | 90 |
| <i>Erica tetralix</i> | 100 | <i>Erica tetralix</i> | 89 |
| <i>Sphagnum</i> spp. | 73 | <i>Sphagnum</i> spp. | 84 |
| <i>Calluna vulgaris</i> | 66 | <i>Potentilla erecta</i> | 61 |
| TYPE 2 | | | |
| <i>Molinia caerulea</i> | 100 | <i>Calluna vulgaris</i> | 96 |
| <i>Calluna vulgaris</i> | 100 | <i>Molinia caerulea</i> | 92 |
| <i>Potentilla erecta</i> | 100 | <i>Trichophorum cespitosum</i> | 73 |
| <i>Trichophorum cespitosum</i> | 79 | <i>Potentilla erecta</i> | 69 |
| TYPE 3 | | | |
| <i>Molinia caerulea</i> | 100 | <i>Juncus acutiflorus</i> | 86 |
| <i>Juncus acutiflorus</i> | 100 | <i>Sphagnum</i> spp. | 76 |
| <i>Potentilla erecta</i> | 100 | <i>Epilobium palustre</i> | 72 |
| <i>Sphagnum</i> spp. | 65 | <i>Galium palustre</i> | 67 |
| TYPE 4 | | | |
| <i>Molinia caerulea</i> | 100 | Type 4 disappears | |
| <i>Potentilla erecta</i> | 100 | | |
| <i>Myrica gale</i> | 36 | | |
| <i>Sphagnum</i> spp. | 32 | | |
| TYPE 5 | | | |
| <i>Molinia caerulea</i> | 100 | <i>Iris pseudacorus</i> | 94 |
| <i>Sphagnum</i> spp. | 61 | <i>Mentha aquatica</i> | 81 |
| <i>Calluna vulgaris</i> | 45 | <i>Juncus effusus</i> | 69 |
| <i>Carex binervis</i> | 29 | <i>Cirsium palustre</i> | 63 |
| TYPE 6 | | | |
| <i>Anthoxanthum odoratum</i> | 100 | <i>Anthoxanthum odoratum</i> | 87 |
| <i>Holcus lanatus</i> | 100 | <i>Holcus lanatus</i> | 78 |
| <i>Conopodium majus</i> | 75 | <i>Conopodium majus</i> | 67 |
| <i>Potentilla erecta</i> | 67 | <i>Potentilla erecta</i> | 67 |
| TYPE 7 | | | |
| <i>Anthoxanthum odoratum</i> | 100 | <i>Conopodium majus</i> | 71 |
| <i>Potentilla erecta</i> | 57 | <i>Luzula sylvatica</i> | 67 |
| <i>Conopodium majus</i> | 47 | <i>Oxalis acetosella</i> | 63 |
| <i>Luzula sylvatica</i> | 30 | <i>Deschampsia cespitosa</i> | 58 |
| TYPE 8 | | | |
| <i>Rhytidiadelphus loreus</i> | 100 | Type 8 disappears | |
| <i>Festuca rubra</i> | 47 | | |
| <i>Galium saxatile</i> | 38 | | |
| <i>Potentilla erecta</i> | 38 | | |
| TYPE 9 | | | |
| <i>Calluna vulgaris</i> | 100 | <i>Molinia caerulea</i> | 96 |
| <i>Drosera rotundifolia</i> | 100 | <i>Potentilla erecta</i> | 96 |
| <i>Sphagnum</i> spp. | 100 | <i>Juncus acutiflorus</i> | 56 |
| <i>Erica tetralix</i> | 80 | <i>Cirsium palustre</i> | 44 |
| TYPE 10 | | | |
| <i>Calluna vulgaris</i> | 100 | <i>Potentilla erecta</i> | 92 |
| <i>Erica tetralix</i> | 65 | <i>Festuca ovina</i> | 84 |
| <i>Potentilla erecta</i> | 65 | <i>Pteridium aquilinum</i> | 73 |
| <i>Sphagnum</i> spp. | 40 | <i>Galium saxatile</i> | 70 |

Table 3 (continued)

| No agglomerative correction | | Agglomerated ($\beta = 0.55$) | |
|-----------------------------|---------------|---------------------------------|---------------|
| | % presence | | % presence |
| TYPE 11 | | | |
| <i>Festuca rubra</i> | 100 | | |
| <i>Glaux maritima</i> | 100 | | |
| <i>Plantago maritima</i> | 91 | Type 11 disappears | |
| <i>Armeria maritima</i> | 83 | | |
| TYPE 12 | | | |
| <i>Festuca rubra</i> | 100 | <i>Festuca rubra</i> | 93 |
| <i>Trifolium repens</i> | 65 | <i>Glaux maritima</i> | 61 |
| <i>Deschampsia flexuosa</i> | 54 | <i>Armeria maritima</i> | 57 |
| <i>Potentilla anserina</i> | 46 | <i>Plantago maritima</i> | 54 |
| TYPE 13 | | | |
| <i>Sphagnum</i> spp. | 33 | | |
| <i>Ranunculus repens</i> | 31 | | |
| <i>Epilobium palustre</i> | 26 | Type 13 disappears | |
| <i>Galium palustre</i> | 23 | | |

salt tolerant species *Glaux maritima*, *Armeria maritima* and *Plantago maritima* all occurring with a presence of 50% or more.

This convergence simplifies the classification as there are now only five main groups (see Fig. 4d). These comprise two *Molinia* heaths (types 1 and 2), two coastal types liable to sea flooding (types 7 and 12), two low-level marshes, occurring at the foots of cliffs and wet valleys (types 6 and 5) and two flushes (types 3 and 9), to which a drier heath type with *Pteridium aquilinum* (type 10) displays an affinity.

The progressive resolution of these types during the agglomerative process is illustrated in Fig. 4. Approximately four groups of overlapping or contiguous circles can be seen in Fig. 4(d) where the variation between the groups is maximized (distance between circles) and the variation within the groups minimized (radii of circles). As can be seen in Table 2, the two components used for the projection of this plane account for 71.4% of the total variation. When the position of the groups on this graph is compared with their species composition as shown in Table 3 there appear to be no significant distortions of the ecological affinities of the groups. The two *Molinia caerulea* moorland heaths are found close together, while the low lying *Molinia* type (9) is found closest to type 3 to which it is most closely related both floristically and topographically. Type 7, although it is not recorded in the table; also contained *Alnus glutinosa* and was the alder carr at the head of Lochranza. This area would be prone to occasional salt flooding as it bordered on the salt marsh area which is grouped as type 12. Types 5 and 6 are both low-level marsh types and probably belong to the most base-rich flushes found in the survey (with the exception of those prone to sea flooding). Groups 3 and 9 distinguish themselves from the other marsh types (5 and 6) by being of a more upland character, containing *Juncus acutiflorus* and *Molinia caerulea* respectively instead of *Iris pseudacorus*, *Conopodium majus* and *Anthoxanthum odoratum*. On slightly drier ground in the same areas as types 3 and 9, i.e. sloping sides of valleys, it would not be unexpected to find stands with *Pteridium aquilinum* as is found in type 10 which contains also *Festuca ovina*, *Galium saxatile* and *Potentilla erecta*.

Tentsmuir dune slacks

Table 4 records the floristic composition of the dune slack types as before, in terms of

Table 4. The effect of the agglomerative corrective process on the group composition of dune slack types at Tentsmuir as seen in the percentage occurrence of the four most frequent species in each vegetation type

| | No agglomerative correction % presence | | Agglomerated ($\beta = 0.55$) % presence |
|------------------------------|--|-----------------------------|--|
| TYPE 1 | | | |
| <i>Carex arenaria</i> | 100 | | |
| <i>Erica tetralix</i> | 100 | | |
| <i>Festuca rubra</i> | 100 | Type 1 disappears | |
| <i>Salix repens</i> | 100 | | |
| TYPE 2 | | | |
| <i>Carex arenaria</i> | 100 | <i>Salix repens</i> | 94 |
| <i>Festuca rubra</i> | 100 | <i>Carex arenaria</i> | 89 |
| <i>Salix repens</i> | 100 | <i>Festuca rubra</i> | 89 |
| <i>Lotus corniculatus</i> | 61 | <i>Lotus corniculatus</i> | 72 |
| TYPE 3 | | | |
| <i>Carex arenaria</i> | 100 | <i>Erica tetralix</i> | 100 |
| <i>Cladonia sylvatica</i> | 100 | <i>Festuca rubra</i> | 100 |
| <i>Festuca rubra</i> | 100 | <i>Cladonia sylvatica</i> | 95 |
| <i>Erica tetralix</i> | 62 | <i>Dicranum scoparium</i> | 64 |
| TYPE 4 | | | |
| <i>Carex arenaria</i> | 100 | <i>Erica tetralix</i> | 91 |
| <i>Festuca rubra</i> | 100 | <i>Filipendula ulmaria</i> | 89 |
| <i>Erica tetralix</i> | 46 | <i>Festuca rubra</i> | 80 |
| <i>Filipendula ulmaria</i> | 46 | <i>Hylocomium splendens</i> | 71 |
| TYPE 5 | | | |
| <i>Carex arenaria</i> | 100 | <i>Cladonia sylvatica</i> | 89 |
| <i>Cladonia sylvatica</i> | 100 | <i>Hieracium pilosella</i> | 86 |
| <i>Hieracium pilosella</i> | 100 | <i>Carex arenaria</i> | 81 |
| <i>Ammophila arenaria</i> | 76 | <i>Ammophila arenaria</i> | 78 |
| TYPE 6 | | | |
| <i>Carex arenaria</i> | 100 | <i>Hieracium pilosella</i> | 94 |
| <i>Hieracium pilosella</i> | 100 | <i>Lotus corniculatus</i> | 88 |
| <i>Lotus corniculatus</i> | 62 | <i>Salix repens</i> | 69 |
| <i>Salix repens</i> | 62 | <i>Ammophila arenaria</i> | 56 |
| TYPE 7 | | | |
| <i>Carex arenaria</i> | 100 | <i>Salix repens</i> | 95 |
| <i>Salix repens</i> | 46 | <i>Lotus corniculatus</i> | 79 |
| <i>Lotus corniculatus</i> | 34 | <i>Carex arenaria</i> | 68 |
| <i>Anthoxanthum odoratum</i> | 27 | <i>Juncus balticus</i> | 58 |
| TYPE 8 | | | |
| <i>Erica tetralix</i> | 100 | | |
| <i>Festuca rubra</i> | 100 | Type 8 disappears | |
| <i>Cladonia sylvatica</i> | 59 | | |
| <i>Hylocomium splendens</i> | 50 | | |
| TYPE 9 | | | |
| <i>Festuca rubra</i> | 100 | <i>Festuca rubra</i> | 88 |
| <i>Honkenya peploides</i> | 100 | <i>Honkenya peploides</i> | 82 |
| <i>Rhinanthus minor</i> | 80 | <i>Rhinanthus minor</i> | 79 |
| <i>Agrostis stolonifera</i> | 60 | <i>Agrostis stolonifera</i> | 48 |
| TYPE 10 | | | |
| <i>Festuca rubra</i> | 100 | <i>Alnus glutinosa</i> | 70 |
| <i>Salix repens</i> | 47 | <i>Festuca rubra</i> | 65 |
| <i>Filipendula ulmaria</i> | 38 | <i>Holcus lanatus</i> | 45 |
| <i>Galium palustre</i> | 27 | <i>Filipendula ulmaria</i> | 40 |

Table 4 (continued)

| No agglomerative correction | | Agglomerated ($\beta = 0.55$) | |
|-------------------------------|---------------|---------------------------------|---------------|
| | % presence | | % presence |
| TYPE 11 | | | |
| <i>Filipendula ulmaria</i> | 100 | <i>Filipendula ulmaria</i> | 100 |
| <i>Hydrocotyle vulgaris</i> | 100 | <i>Hydrocotyle vulgaris</i> | 71 |
| <i>Lophocolea bidentata</i> | 38 | <i>Acrocladium cuspidatum</i> | 46 |
| <i>Acrocladium cuspidatum</i> | 30 | <i>Lophocolea bidentata</i> | 25 |
| TYPE 12 | | | |
| <i>Filipendula ulmaria</i> | 100 | <i>Galium palustre</i> | 76 |
| <i>Galium palustre</i> | 46 | <i>Filipendula ulmaria</i> | 72 |
| <i>Juncus effusus</i> | 46 | <i>Juncus effusus</i> | 72 |
| <i>Agrostis stolonifera</i> | 30 | <i>Agrostis stolonifera</i> | 29 |
| TYPE 13 | | | |
| <i>Hieracium pilosella</i> | 46 | | |
| <i>Juncus balticus</i> | 34 | | |
| <i>Lotus corniculatus</i> | 34 | | |
| <i>Salix repens</i> | 34 | Type 13 disappears | |

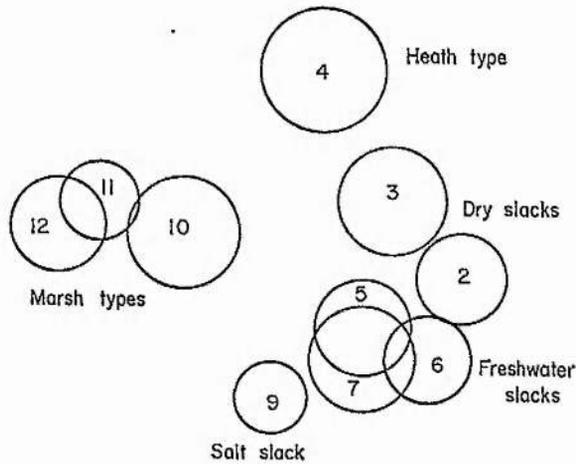


FIG. 5. Ordination of groups obtained in the Tentsmuir dune slack survey after agglomeration at $\beta = 0.55$. (Ordination drawn by computer.)

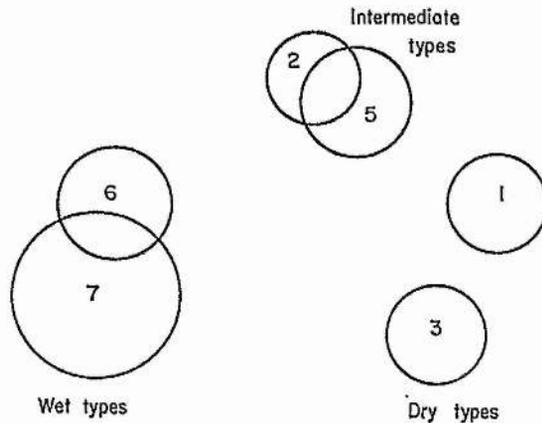


FIG. 6. Ordination of groups obtained in the Tentsmuir mapped transect after agglomeration at $\beta = 0.55$. (Ordination drawn by computer.)

the four most frequent species obtained before and after the application of the agglomerative corrective procedure. Three of the slack types obtained by the divisive process disappear, leaving ten types whose mutual affinities are represented by their positions in Fig. 5.

Table 5. *The effect of the agglomerative corrective process on the group composition of wet land vegetation types in the Tentsmuir mapped transect as seen in the percentage occurrence of the four most frequent species in each type*

| No agglomerative correction | | Agglomerated ($\beta = 0.55$) | |
|-------------------------------|---------------|---------------------------------|---------------|
| | % presence | | % presence |
| TYPE 1 | | | |
| <i>Anthoxanthum odoratum</i> | 100 | <i>Carex arenaria</i> | 100 |
| <i>Carex arenaria</i> | 100 | <i>Erica tetralix</i> | 97 |
| <i>Festuca rubra</i> | 84 | <i>Festuca rubra</i> | 97 |
| <i>Pleurozium schreberi</i> | 72 | <i>Anthoxanthum odoratum</i> | 91 |
| TYPE 2 | | | |
| <i>Carex arenaria</i> | 100 | <i>Filipendula ulmaria</i> | 100 |
| <i>Filipendula ulmaria</i> | 100 | <i>Carex arenaria</i> | 88 |
| <i>Holcus lanatus</i> | 100 | <i>Holcus lanatus</i> | 85 |
| <i>Poa pratensis</i> | 68 | <i>Festuca rubra</i> | 82 |
| TYPE 3 | | | |
| <i>Carex arenaria</i> | 100 | <i>Carex arenaria</i> | 97 |
| <i>Filipendula ulmaria</i> | 100 | <i>Pleurozium schreberi</i> | 85 |
| <i>Carex nigra</i> | 55 | <i>Galium saxatile</i> | 83 |
| <i>Festuca rubra</i> | 45 | <i>Ammophila arenaria</i> | 68 |
| TYPE 4 | | | |
| <i>Carex arenaria</i> | 100 | | |
| <i>Pleurozium schreberi</i> | 85 | Type 4 disappears | |
| <i>Galium saxatile</i> | 60 | | |
| <i>Ammophila arenaria</i> | 45 | | |
| TYPE 5 | | | |
| <i>Carex nigra</i> | 100 | <i>Filipendula ulmaria</i> | 100 |
| <i>Galium palustre</i> | 100 | <i>Pleurozium schreberi</i> | 82 |
| <i>Filipendula ulmaria</i> | 96 | <i>Poa pratensis</i> | 76 |
| <i>Acrocladium cuspidatum</i> | 80 | <i>Carex arenaria</i> | 71 |
| TYPE 6 | | | |
| <i>Carex nigra</i> | 100 | <i>Filipendula ulmaria</i> | 98 |
| <i>Filipendula ulmaria</i> | 100 | <i>Carex nigra</i> | 96 |
| <i>Poa pratensis</i> | 100 | <i>Acrocladium cuspidatum</i> | 71 |
| <i>Glyceria maxima</i> | 55 | <i>Galium palustre</i> | 55 |
| TYPE 7 | | | |
| <i>Carex nigra</i> | 100 | <i>Filipendula ulmaria</i> | 77 |
| <i>Filipendula ulmaria</i> | 100 | <i>Galium palustre</i> | 63 |
| <i>Acrocladium cuspidatum</i> | 58 | <i>Acrocladium cuspidatum</i> | 61 |
| <i>Glyceria maxima</i> | 42 | <i>Glyceria maxima</i> | 43 |
| TYPE 8 | | | |
| <i>Filipendula ulmaria</i> | 77 | | |
| <i>Acrocladium cuspidatum</i> | 71 | Type 8 disappears | |
| <i>Galium palustre</i> | 39 | | |
| <i>Glyceria maxima</i> | 39 | | |

The salt tolerant species are all found in type 9 and this is essentially the same group as that recognized in the association analysis carried out on this survey (Crawford & Wishart 1966). The floristic affinities of types 5, 6 and 7, the slacks containing *Juncus balticus*, *Lotus corniculatus* and *Salix repens* is again similar to the types recognized in the

association analysis. As the slacks age (the area is accreting rapidly and the older slacks are now some 600–800 m from the sea) there is a final divergence into types 10, 11 and 12, the marsh types, and types 3 and 4, the drier *Erica tetralix* slacks. Although the relative positions of these types as shown in Fig. 5 are comparable with the Bray and Curtis ordination carried out after association analysis (see Crawford & Wishart 1966) the present ordination appears to avoid the accentuation of differences obtained with the earlier ordination technique in that the salt-containing type 9 is shown as having a greater affinity with the early slacks, i.e. those containing *Juncus balticus* and *Salix repens*.

Tentsmuir mapped transect

Table 5 records the floristic composition of the transect vegetation types before and after the process of agglomeration. Two types disappear, and the final list contains six vegetation types whose mutual floristic affinities are illustrated in Fig. 6. From their floristic composition and ordination positions these types can be considered as three main groups: the wet slacks (types 6 and 7), with *Filipendula ulmaria* and *Glyceria maxima*, two intermediate types, 2 and 5, both containing *Filipendula ulmaria* but in combination with other species more typical of the drier vegetation types and, finally, the drier slack types, 1 and 3, in which *Carex arenaria* is the most frequent species but occurs in combination with *Ammophila arenaria* and *Galium saxatile* in type 3 and *Erica tetralix* and *Festuca rubra* in type 1. These latter types are seen in the foreground in Phot. 1 with the intermediate and wet land types in the middle and far distance respectively. The actual distribution of the groups on the ground is illustrated in the computer drawn contour maps shown in Fig. 7 (a–d) where the contours of the six groups are drawn at increasing levels of γ from 0.2 to 0.5. Even at the low level of $\gamma = 0.2$ a distinction is already clear between the distribution of types 1, 2, 3 and 4 on one hand and types 5 and 6 on the other. The extreme positions of the summer and winter water tables measured over a 2-year period in this transect is illustrated in Fig. 8, and it can be seen that types 5 and 6 are confined to the wettest areas of the transect. As the level of γ is raised, the detail of distinction between the other types becomes apparent. Only at $\gamma = 0.5$ however, does the distinction between types 1 and 3 reveal itself with type 3, the group containing *Ammophila arenaria*, having its greatest potential values on the highest ground. With the wet land vegetation groups, type 6 appears to be the more widely spread, with type 7 only occurring within the area of distribution of type 6. It is possible to plot contours for different levels of γ on the same map but for the sake of clarity, separate maps have been prepared. However, if the contours illustrated in Fig. 7 (a–d) are superimposed on one another the gradual replacement of one vegetation type by another along the transect is easily observed.

DISCUSSION

As with the divisive method described in the previous communication the present method is aimed at computational speed, even when the number of quadrats and species is large. To achieve this end, the detailed correlations of the accepted methods, either divisive or agglomerative, are sacrificed for a much faster but more general approach. The standard methods in cluster analysis (see Macnaughton-Smith 1965) consider the relationships between all possible $\frac{1}{2}n(n-1)$ pairs, and for surveys in which n is large, this quadratic relationship will become highly demanding in computer time.

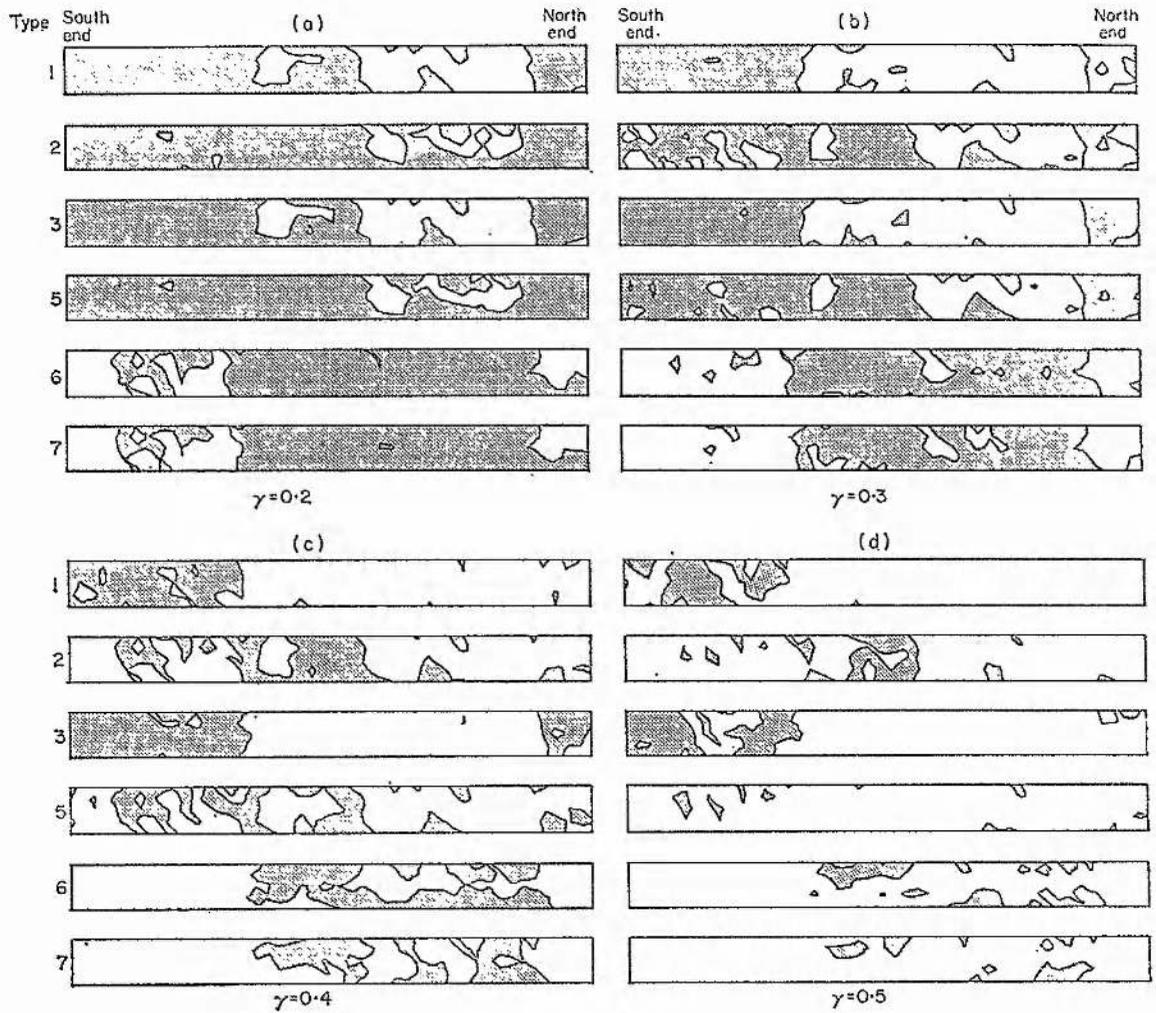


FIG. 7 (a-d). Computer-drawn contour maps showing the areas occupied by each vegetation type when delimited by different levels of γ (γ = level of potential for each quadrat for belonging to any particular type—see text).

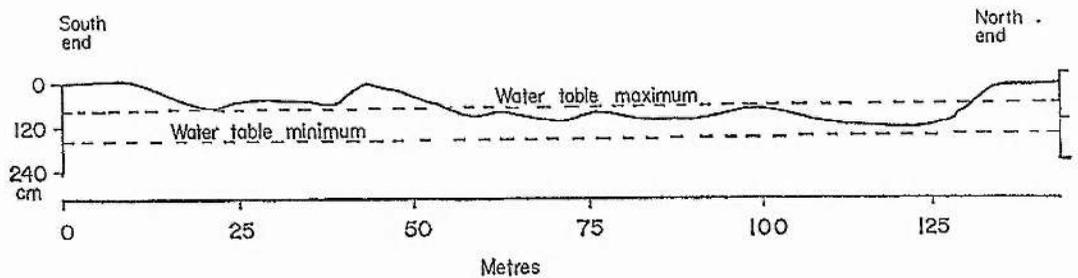


FIG. 8. Profile of the Tentsmuir mapped transect showing the minimum and maximum levels attained by the water table during 1965 and 1966 (maximum, 25 February 1966; minimum, 8 June 1965).

In a divisive process such as association analysis, the same problem occurs in that there is a quadratic relationship between species number and computation time (Williams & Lambert 1960). In discussing the relative merits of association analysis (divisive) and information analysis (agglomerative) Lambert & Williams (1966) conclude that the choice between the two methods depends on whether the survey in question contains more quadrats than species, or vice-versa. In either case one quadratic relationship with survey size is inevitable. In the method of group analysis described by Crawford & Wishart (1967) a quadratic relationship of either species or quadrat number with computation time is avoided as the number of species has little effect on the analysis time and the relationship between quadrat number and computing time is linear. As a similar strategy is used in the present agglomerative procedure the relationship between survey size and computing time appears to be linear.

Further, when carrying out the ordination procedure the dimensions are greatly reduced as the generality of N dimensional space is sacrificed for K -space (K = the number of groups obtained). Although the method is similar to that of Orloci (1967) the latter required to carry out $n(n-1)/2$ tests and fusions of all the possible entities and had the survey been much larger than the thirty-nine samples and 109 species, a much greater demand on storage space would have been made, and without access to a large computer might have been difficult to carry out. In the present investigation for both the Arran survey with 554 quadrats and ninety-eight species and the Tentsmuir dune slack survey with 263 quadrats and 142 species the analyses were carried out in under 3 h each on the IBM 1620.

The mapping principle of plotting the potential of each quadrat for belonging to any particular type affords a more detailed picture of vegetation groupings than a direct plot of the classification of the quadrats. With the latter the accuracy of the map decreases as the size of the quadrat increases. The larger the quadrat the more likely it is to encompass two or more different groups. This possibility is particularly evident when the sample areas of vegetation are large, as in the study of the distribution of British liverwort associations by Proctor (1967). In this study the sample units are the individual vice-counties of the British Isles and these are classified by both normal and inverse association analysis. As the author points out, in the normal association analysis a number of vice-counties will inevitably be misplaced. A contouring process, as used in the present study, would overcome this difficulty as borderline counties with affinities for more than one region would be readily distinguished and represented as such.

If the contours for varying potential shown in Fig. 7 are superimposed on each other it is seen that the species groups show no distinct boundaries but mutually replace one another. These gradations in ecological groups contrast sharply with the appearance of the vegetation if it is examined in terms of physiognomic dominants. As can be seen in Photos. 1 and 2 there appear to be clearly definable lines delineating zones of *Erica tetralix*, *Glyceria maxima*, *Carex nigra* and *Juncus effusus*. This example illustrates the nature of arguments that have been made about the reality of community boundaries. While it is always possible to draw a boundary marking the limits of distribution of one particular species irrespective of whether these limits are obtained from an examination of physiognomic dominants or by a monothetic divisive process, no boundary can be drawn with any precision for any vegetation type that is defined on the basis of the probable occurrence of a number of species. Greig-Smith (1964), in discussing the relative merits of ordination and classification, points out the over-emphasis on discontinuities obtained in classifications, concludes that ecologically ordination is a sounder approach, and does not

accept the argument that the results of an ordination analysis cannot be mapped. This present study would support this view in that provided the results of the analysis are appropriately expressed it is possible to classify, ordinate and map the resultant data.

The classification of vegetation obtained with this method of group analysis can be compared with the diagnosis of a disease in terms of a syndrome, where a number of symptoms occurring in one patient enable his condition to be pathologically classified. In this study the species can be considered the symptoms of an ecological situation and the greater the coincidence of their occurrence the more clearly is the condition defined. In principle, this is little more than the 'Kennarten' of Braun-Blanquet (1964) but because it is determined objectively on a number of species it is more likely to be of universal application.

The method is designed for the detection of major groups in large surveys and should there be significant groups of size only marginally greater than L (the subsidiary stopping rule for subset size in the divisive process) then these will not be detected.

The classification or identification of further material is a desirable feature in any method of numerical taxonomy and in this the monothetic divisive methods have an advantage over polythetic agglomerative methods as the identity of each sample is determined by a number of conditional vectors. Polythetic methods, however, require a certain amount of reprocessing with the entire survey in order to find the group of best fit. In

Table 6. *Summary table recording the co-ordinates of the characteristic vectors for a hypothetical example of three groups and seven species*

| | Species | | | | | | | |
|-------|---------|-----|-----|-----|-----|-----|-----|------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | ΣW |
| W_1 | 0.2 | 0.8 | 0.1 | 0.6 | 0.2 | 0.1 | 0.1 | 2.4 |
| W_2 | 0.8 | 0.1 | 0.1 | 0.1 | 0.7 | 0.6 | 0.1 | 2.5 |
| W_3 | 0.1 | 0.2 | 0.8 | 0.1 | 0.2 | 0.1 | 0.6 | 2.1 |

the present method identification of further material in the field is possible if the co-ordinates of the characteristic vectors for each group have been retained. Table 6 illustrates such a record for a hypothetical example of seven species and three groups. If a new quadrat contains species 3, 5 and 7 only, then it will have potential values with respect to each group as follows:

$$\text{Group 1 } \frac{0.1+0.2+0.1}{2.4} = 0.166$$

$$\text{Group 2 } \frac{0.1+0.7+0.1}{2.5} = 0.36$$

$$\text{Group 3 } \frac{0.8+0.2+0.6}{2.1} = 0.76$$

The new quadrat is therefore associated with group 3.

ACKNOWLEDGMENTS

We are much indebted to Dr A. J. Cole for his subroutines to plot grid contours and compute eigen values and eigen vectors from the correlation matrix. We are also most grateful

to the staff of the St Andrews computing laboratory for their assistance in the preparation and processing of the data, and to Miss M. C. MacNaughtan for the use of the records from the mapped transect at Tentsmuir.

SUMMARY

An agglomerative method is described for the rapid checking of possible misclassifications that are inevitable in a monothetic divisive process such as the method of group analysis described in an earlier communication. Using this corrected data an ordination method is suggested for displaying graphically the variation within and between the classified groups. The method is designed for speed of computation even when the survey is large and achieves this by replacing the calculation of N -dimensional space with K -space (K = the number of groups obtained).

The mapping by computer, of vegetation sampled systematically on a grid, is described, using the potentials of the quadrats for belonging to the recognized groups. This method allows for the possibility of one quadrat having an equal affinity with more than one vegetation type and the map is drawn by contours delimiting areas of equal potential for each vegetation type.

Finally, the identification of further material in the field without resorting to additional computation is described.

All the graphic displays of the data are obtained directly from an on-line digital plotter, thus eliminating much of the tedium associated with vegetation ordination and mapping studies.

The method is designed for the detection of major groups in large surveys and should there be significant groups of size only marginally greater than L (the subsidiary stopping rule for sub-set size in the divisive process) then these will not be detected.

The analysis systems written in Fortran IID are available on application to the authors.

REFERENCES

- Austin, M. P. & Orloci, L. (1966). Geometric models in ecology. II. An evaluation of some ordination techniques. *J. Ecol.* 54, 217-27.
- Braun-Blanquet, J. (1964). *Pflanzensoziologie*, 3rd edn. Vienna.
- Clapham, A. R., Tutin, T. G. & Warburg, E. F. (1962). *Flora of the British Isles*, 2nd edn. Cambridge.
- Crawford, R. M. M. & Wishart, D. (1966). A multivariate analysis of the development of dune slack vegetation in relation to coastal accretion at Tentsmuir, Fife. *J. Ecol.* 54, 729-43.
- Crawford, R. M. M. & Wishart, D. (1967). A rapid multivariate method for the detection and classification of groups of ecologically related species. *J. Ecol.* 55, 505-24.
- Greig-Smith, P. (1964). *Quantitative Plant Ecology*, 2nd edn. London.
- Jancey, R. C. (1966). Multidimensional group analysis. *Aust. J. Bot.* 14, 127-30.
- Kendall, M. G. (1957). *A Course in Multivariate Analysis*. London.
- Lambert, J. M. & Williams, W. T. (1966). Multivariate methods in plant ecology. VI. Comparison of information-analysis with association-analysis. *J. Ecol.* 54, 635-64.
- Macnaughton-Smith, P. (1965). *Some statistical and other numerical techniques for classifying individuals*. (Home Office: Studies in the causes of delinquency and the treatment of offenders 6). London.
- Orloci, L. (1966). Geometric models in ecology. I. The theory and application of some ordination methods. *J. Ecol.* 54, 193-215.
- Orloci, L. (1967). An agglomerative method for classification of plant communities. *J. Ecol.* 55, 193-206.
- Proctor, M. C. F. (1967). The distribution of British liverworts: a statistical analysis. *J. Ecol.* 55, 119-35.
- Richards, P. W. & Wallace, E. C. (1950). An annotated list of British mosses. *Trans. Br. bryol. Soc.* 1, i-xxx.
- Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco.

Rapid classification and ordination

- Williams, W. T. & Dale, M. B. (1965). Fundamental problems in numerical taxonomy. *Adv. bot. Res.* 2, 35-68.
- Williams, W. T. & Lambert, J. M. (1959). Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* 47, 83-101.
- Williams, W. T. & Lambert, J. M. (1959). Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. *J. Ecol.* 47, 689-710.

(Received 24 July 1967)

A NUMERICAL ANALYSIS OF HIGH ALTITUDE SCRUB VEGETATION IN RELATION TO SOIL EROSION IN THE EASTERN CORDILLERA OF PERU

BY R. M. M. CRAWFORD, D. WISHART AND R. M. CAMPBELL

Department of Botany, The University, St Andrews

INTRODUCTION

This study of the scrub vegetation which occurs immediately below the tree line over wide stretches of the Andes was carried out during a 3-month expedition to south-eastern Peru in 1967. In this region of South America, because of the extensive removal of natural tree cover by felling and grazing, soil erosion is an increasing problem. The high altitude scrub vegetation growing about the level of the tree line is one of the few remaining associations of natural species which can maintain the stability of the soil and prevent the rapid acceleration of erosion.

In spite of the important role this scrub formation plays in maintaining the fertility of upland soils no study has been made of its species composition or structure. Tosi (1960), in an account of the vegetation zones of Peru, based on Holdridge's (1947) system of climatic classification; describes some thirty vegetation formations. One of these, *Bosque Seco Montano Bajo* corresponds closely to the scrub described in this paper. As Tosi points out, this formation is commonly found in those densely settled regions of the *Sierra* (mountain regions) where the principal towns are to be found, many of which are of great antiquity, such as Huancayo, Ayacucho, Andahuaylas, Abancay, Urabamba and Paucartambo (see Fig. 1). A list of genera commonly found in this formation includes *Kageneckia*, *Cassia*, *Barnadesia*, *Agave*, *Spartium* and *Schinus*, all of which occurred in the scrub surveyed by the expedition. According to Tosi the formation of this scrub is brought about by an interaction of human activity with the edaphic and climatic conditions.

Although the scrub is of little direct economic use, its presence in these mountain areas is of great importance in relation to soil conservation. When the scrub is removed, the soil is deprived of a large part of its permanent vegetation cover and is no longer capable of holding sufficient water to prevent a rapid acceleration of sheet erosion. It was observed in this present study that at the position on the valley sides where the scrub died out arable agriculture also ceased (see Phot. 3). Therefore a knowledge of the composition and distribution of the scrub is of crucial importance to the maintenance of soil stability and very relevant to the agricultural problems of the area.

A wide variety of current numerical techniques has been used in this investigation. The current exploratory stage of numerical taxonomy requires a comparative approach which makes use of several methods, if the species groupings are to be demonstrated as biological entities and not just artifacts of the particular method employed.

It is hoped therefore that this investigation, as well as being an examination of the last bastion of phanerophyte vegetation on the eroding slopes of the Andes, may serve also as a practical comparison of the application of many of the numerical taxonomy methods currently in use.

DESCRIPTION OF AREA

The survey was restricted to one locality, as it was thought that a detailed study of the distribution of the vegetation in relation to altitude and aspect would be more informative than a diffuse sampling over a wider area. The object of the investigation was to determine if a pattern of species distribution existed in this scrub in relation to altitude and aspect. If this could be established at one pilot site then it is not unlikely that similar results would be obtained elsewhere.

The area chosen for the investigation was in the valley of the Vilcanota (a tributary of the Urabamba river) at Urco, 2 miles (3 km) north-east of Calca (72°0'W, 13°10'S; see Fig. 1). The area is typical of those described by Tosi as the principal sites for this type of high altitude scrub. This region of the Vilcanota valley is only some 30 miles (48 km)

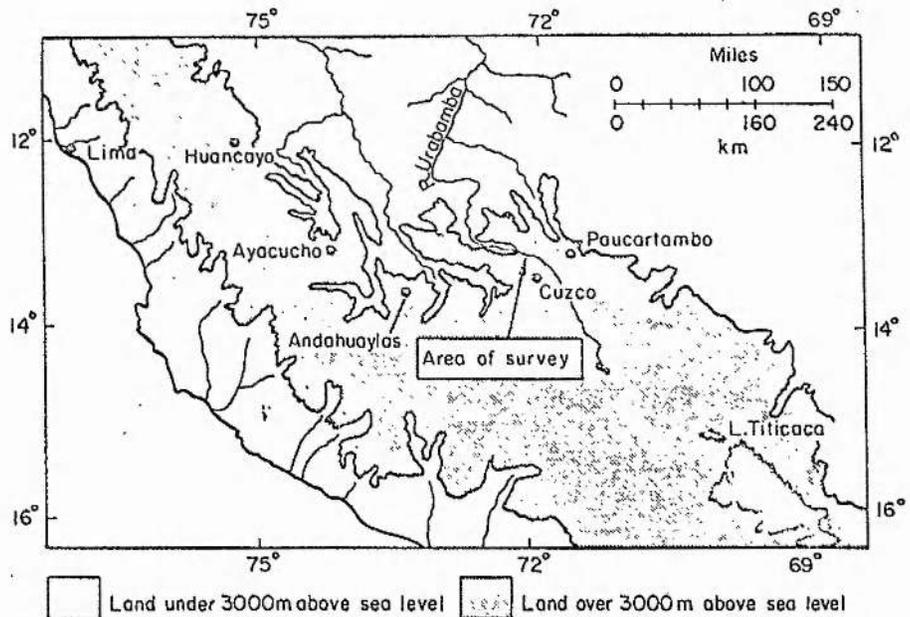


FIG. 1. Map of southern Peru showing the area of survey.

from Cuzco, the ancient capital of the Inca empire. The equitable climate of the valley, resembling perpetual spring, made the region a popular summer residence with the Incas and there is sufficient archaeological evidence to indicate that it has been densely settled for a long period.

The valley floor at the point where the survey was carried out lies at an altitude of 3000 m (9800 ft) and is approximately 1 mile (1.6 km) wide. The valley runs east-west, and on either side the mountain slopes rise steeply to approximately 4000 m (13000 ft) from where the table land of the *altiplano* extends, interrupted only by the peaks of the Eastern Cordillera. A view across the valley looking north is shown in Phot. 1, and illustrates the extensive scrub cover found on the mountain sides as well as the bare eroding soils at the higher altitudes seen in the far distance. Apart from a patchwork of fields this scrub covers the floor and sides of the valley up to the tree-line which lies here at an altitude of 3650 m (12000 ft). Phot. 2 is taken from the eroding region seen in the far distance in Phot. 1 and shows the mountain side at the upper limit of scrub growth with the treeless expanse of the Eastern Cordillera beyond (see also Phot. 3.).

In this region the tree-line is somewhat lower than that of 4000 m reported by Ellenberg (1968). This may well be related to the dense agricultural population that has long been settled in the valley. In a re-examination of the natural climax vegetation of the high altitude steppe lands (*altiplano*) of Peru, Ellenberg concludes that the natural tree-line lies at an altitude of 4500 m and that the present line at 4000 m is artificially low. Much of the land over 4000 m in Peru has been treeless for a long time. The vast grasslands of the *altiplano* made a strong impression on the earliest Spanish *conquistadores* (Prescott 1847) so that the antiquity of this feature, together with the high altitude, has led ecologists to consider the area as belonging to a natural steppe formation. However, it seems probable that the grazing of these upland pastures has been intensive ever since the rise of the Inca empire in the eleventh century A.D. We know from historical accounts made at the time of the Spanish conquest (1528-31) that the herds of llamas and alpacas were so large that insufficient grazing could be found for them (Garcilaso de la Vega 1608). It seems probable therefore that Ellenberg's conclusions on the biotic rather than climatic determination of the present tree-line are justified.

Land shortage and overgrazing have thus been constant features of life in the Sierra for many centuries. Recently, however, the problem has become even more acute. The population of Peru at the time of the Spanish conquest has been estimated to have been in the region of 3 million (Rowe 1946). Recent census figures give estimates varying between 12 and 16 million. The successful eradication of many diseases, particularly malaria, during the last 30 years has led to a very large increase in the agricultural population despite large migrations to the towns.*

The Vilcanota valley is one of the most densely populated agricultural areas, and an area where the human influence on vegetation has been felt for many centuries. The problems of land shortage are likely to increase, and in this the high altitude scrub vegetation has a vital role to play in the maintenance of soil fertility and the prevention of further erosion.

METHODS

Sampling

The vegetation was sampled at random within a strip 1 km wide running across the floor of the valley and up either side to the tree-line. The species were listed on the basis of their presence or absence in a 5 m square quadrat. As it was the structure of the scrub vegetation that was being examined and not the ground flora, only perennial species having a height of 30 cm or more were listed. In addition, the altitude of each quadrat above sea level was recorded using an altimeter, as well as the aspect of the sample area. For the latter, an eight-point compass scale was used, with 0 being recorded for quadrats taken on level ground.

In all, 450 samples of vegetation were taken, giving a list of thirty-seven different species. These scrub species were all readily identifiable and the nomenclature used follows that of Herrera (1941) and Vargas (1966). The identifications were checked by Professor C. Vargas against specimens in the herbarium of the University of Cuzco. In contrast with much of the flora of Peru, which is not well known botanically, this scrub vegetation presented few taxonomic difficulties. Most of the species appear also to be recognized by the local population as the majority have common names in both Spanish and Quechua (the native Indian language).

* *Peruvian Times*, 8 January 1960.

Numerical analyses

A summary of the methods used is given in Table 1, together with the relevant coefficients and references. Since the IBM 1620 II computer restricts some of the programs to 400 individuals, a subset of 400 quadrats selected by a pseudo-random number generator was used consistently throughout the entire study. As expected, the species composition for the subset did not differ significantly from that of the original survey.

Table 1. List of numerical methods and coefficients used in the analysis of the survey

| Method | Coefficient | Remarks |
|---|--------------------------------|---------------------------|
| AGGLOMERATIVE | | |
| 1. Single linkage | d^{2*} | Fail to produce clusters |
| 2. Single linkage | A/M^* | |
| 3. Furthest neighbour | d^2 | |
| 4. Furthest neighbour | A/M | |
| 5. Group average | d^2 | |
| 6. Group average | A/M | |
| 7. Centroid sorting | d^2 | |
| 8. Centroid sorting | $2\Delta I$ (information gain) | |
| 9. Centroid sorting | Pearson's ϕ coefficient | |
| 10. Centroid sorting | non-metric coefficient | |
| 11. Ward's error sum method | | |
| 12. Lance and Williams' flexible method | | |
| 13. Group analysis (agglomerative stage) | | |
| DIVISIVE | | |
| 14. Association analysis | $\Sigma \chi^2$ | Fails to produce clusters |
| 15. Association analysis | $\Sigma \sqrt{\chi^2}$ | |
| 16. Association analysis | $\Sigma (AD-BC)^{2*}$ | |
| 17. Maximum information fall | $2\Delta I$ | |
| 18. Maximum centroid distance | | |
| 19. Maximum decrease error sum of squares | | |
| 20. Group analysis (divisive stage) | | |

* d^2 , Euclidean distance; A/M , Russell & Rao's similarity coefficient, where A, B, C, D, refer to the usual 2×2 table, $M = A+B+C+D$, the total number of species.

References: 1, 2, Sokal & Sneath (1963), Lance & Williams (1967); 3, 4, Sørensen (1948), Sokal & Sneath (1963), Johnson (1967), Lance & Williams (1967); 5, 6, Sokal & Michener (1958), Sokal & Sneath (1963), Lance & Williams (1967); 7-10, Lance & Williams (1966, 1967); 11, Ward (1963), Orloci (1967), Wishart (1969a); 12, Lance & Williams (1967); 13, Crawford & Wishart (1967); 14-16, Lance & Williams (1965), Macnaughton-Smith (1965), Gower (1967); 17, Macnaughton-Smith (1965), Lance & Williams (1968); 18, 19, Gower (1967); 20, Crawford & Wishart (1968).

Some of the methods used did not resolve satisfactory clusters, notably owing to the chaining of individual quadrats or small groups onto one large predominant cluster. This is illustrated by the dendrogram for centroid sorting using Euclidean distance (d^2) shown in Fig. 4(a). A discussion of chaining effects in relation to classification techniques has been given by Williams, Lambert & Lance (1966) and Wishart (1969a). In this present study chaining was found in agglomerative methods 1-7 (see Table 1) and in one of the divisive methods, viz. maximum centroid distance. As these procedures were unfruitful in producing any meaningful classification of the data they have been omitted from further discussion in the presentation of the results.

Details of the floristic composition for each cluster can be obtained at any stage of fusion or division together with data on height and aspect. In most cases this was restricted to the last ten fusions or first four divisions of the population. An on-line digital

plotter was used for drawing the dendrograms (Fig. 4) as well as for plotting the ordination given in Fig. 7(a-c).

Table 2. *The species composition of the scrub vegetation of the valley floor and mountain sides at Urco as determined from 450 random quadrats*

| Species | % presence | Species | % presence |
|-------------------------------------|------------|----------------------------------|------------|
| <i>Barnadesia horrida</i> | 33 | <i>Caesalpinia tinctoria</i> | 4 |
| <i>Psila boliviana</i> | 30 | <i>Polyepis incana</i> | 4 |
| <i>Schinus molle</i> | 30 | <i>Proustia pungens</i> | 4 |
| <i>Baccharis cassinaefolia</i> | 30 | <i>Franseria artemisioides</i> | 4 |
| <i>Berberis boliviana</i> | 24 | <i>Hyptis arborea</i> | 2 |
| <i>Cassia hookeriana</i> | 22 | <i>Nicotiana glauca</i> | 2 |
| <i>Astragalus garbancillo</i> | 21 | <i>Minthostachys glabrescens</i> | 2 |
| <i>Marrubium vulgare</i> | 21 | <i>Agave americana</i> | 1 |
| <i>Colletia spinosa</i> | 15 | <i>Chenopodium ambrosioides</i> | 1 |
| <i>Eupatorium pentlandianum</i> | 14 | <i>Eucalyptus globulus</i> | 1 |
| <i>Berberis commutata</i> | 13 | <i>Eremocharis triradiata</i> | 0.7 |
| <i>Puya longistyla</i> | 12 | <i>Alonsoa acutifolia</i> | 0.7 |
| <i>Baccharis salsifolia</i> | 9 | <i>Psoralea glandulosa</i> | 0.7 |
| <i>Solanum pulverulentum</i> | 9 | <i>Psittacanthus cuneifolius</i> | 0.4 |
| <i>Croton ruizii</i> | 9 | <i>Hypericum cespitosum</i> | 0.4 |
| <i>Citharexylum argentidentatum</i> | 7 | <i>Kageneckia lanceolata</i> | 0.4 |
| <i>Spartium junceum</i> | 7 | <i>Sambucus peruviana</i> | 0.2 |
| <i>Gynoxys nitida</i> | 6 | | |
| <i>Ephedra americana</i> | 4 | | |

Table 3. *List of species with the greatest altitudinal range in their distribution (for actual occurrences in altitude see Fig. 2)*

| | Altitudinal range in ft (m) | % presence |
|---------------------------------|-----------------------------|------------|
| <i>Eupatorium pentlandianum</i> | 2230 (680) | 14 |
| <i>Berberis boliviana</i> | 2230 (680) | 24 |
| <i>B. commutata</i> | 2230 (680) | 13 |
| <i>Marrubium vulgare</i> | 2175 (665) | 21 |
| <i>Barnadesia horrida</i> | 2090 (635) | 33 |
| <i>Baccharis cassinaefolia</i> | 2050 (625) | 30 |
| <i>Astragalus garbancillo</i> | 2035 (620) | 21 |
| <i>Cassia hookeriana</i> | 2010 (615) | 22 |

Table 4. *List of species with the most restricted range in altitudinal distribution (for actual occurrences in altitude see Fig. 2)*

| | Altitudinal range in ft (m) | % presence |
|---------------------------------|-----------------------------|------------|
| <i>Sambucus peruviana</i> | 0 (0) | 0.2 |
| <i>Kageneckia lanceolata</i> | 0 (0) | 0.4 |
| <i>Psoralea glandulosa</i> | 100 (30) | 0.7 |
| <i>Hypericum cespitosum</i> | 120 (35) | 0.4 |
| <i>Proustia pungens</i> | 170 (50) | 4.0 |
| <i>Chenopodium ambrosioides</i> | 280 (85) | 1.0 |
| <i>Croton ruizii</i> | 300 (90) | 9.0 |

Computer programs

A numerical classification program package has now been developed for use with computers having FORTRAN compilers and magnetic disc peripherals. The first release,

entitled CLUSTAN I, is written in FORTRAN II for the IBM 1620 II (Wishart 1969b) and clusters by eight hierarchical methods using any one of nineteen similarity coefficients. One of these programs, called HIERAR, was used for analyses 1-7, 11 and 12 in Table 1.

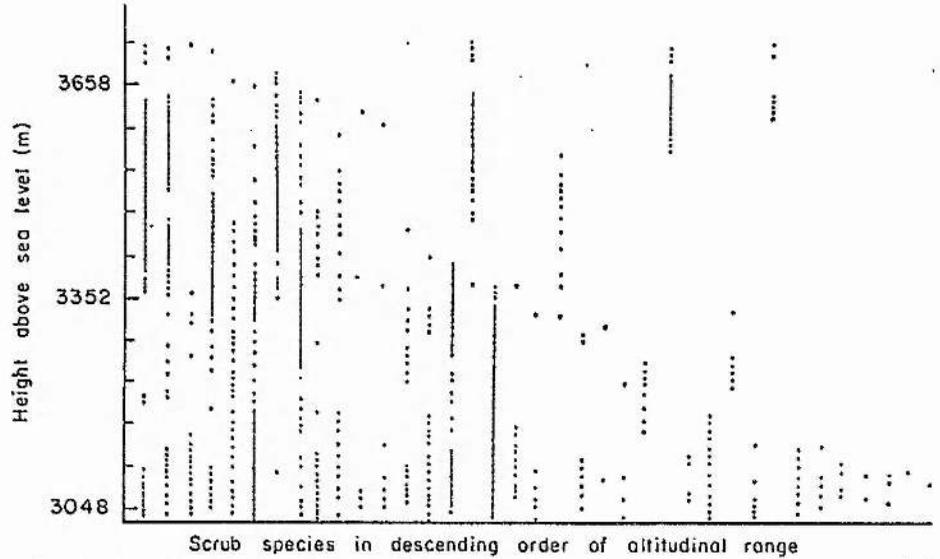


FIG. 2. Plot for each species of all occurrences in relation to altitude. The dots are joined by a line where the density of species occurrences is too great to be represented by individual dots. The species are plotted from left to right in descending order of altitudinal range.

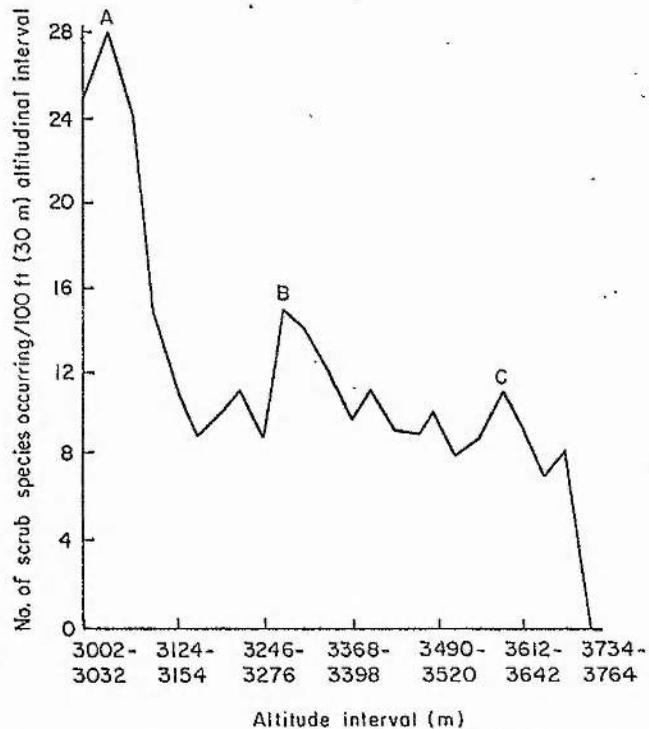


FIG. 3. Species density/100 ft (30 m) altitudinal interval.

The second release, entitled CLUSTAN IA and still unpublished, contains program CENTRO (for analyses 8-10 in Table 1), DIVIDE (for analyses 14-20 in Table 1) and RELOC (for analysis 13 in Table 1). All these programs are available in FORTRAN II or FORTRAN IV

and have been recently increased to accommodate up to 1000 individuals. Modified versions of CLUSTAN I are currently in use on the IBM 1620 II, KDF9, ICL 1909, ICL 4/75 and IBM 360 computers. Details are available on application to D. Wishart.

RESULTS

The species composition of this high altitude scrub as taken from the percentage frequencies calculated from the entire survey is given in Table 2. From this it can be seen

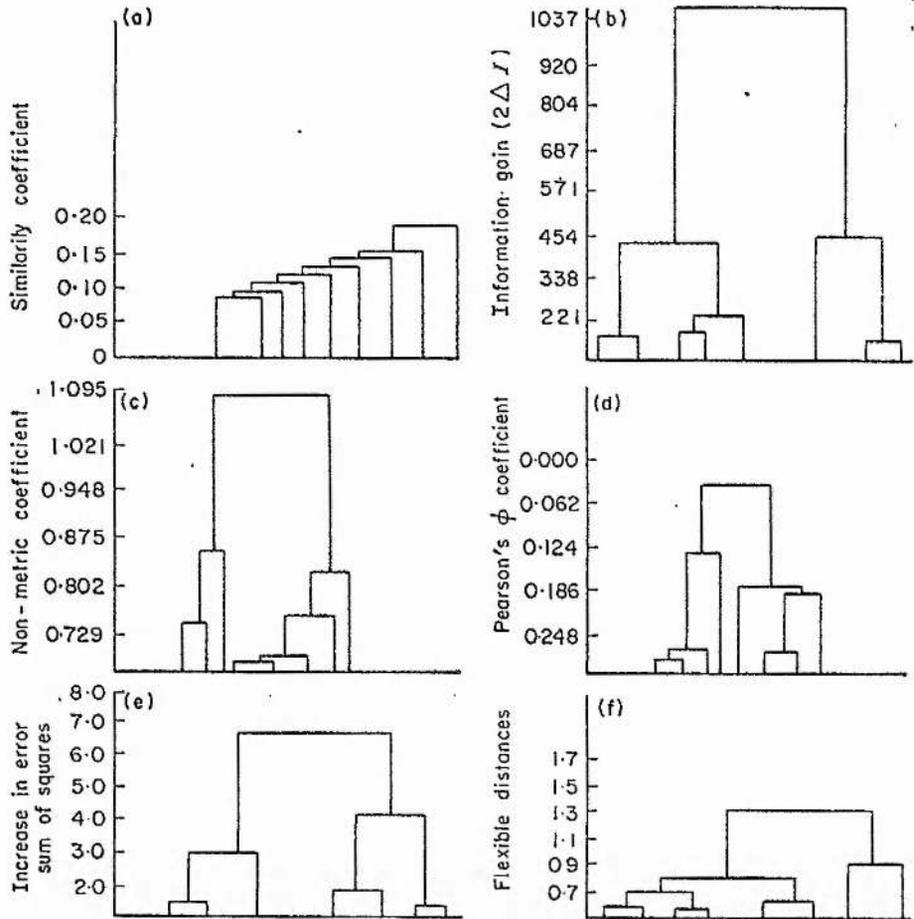


FIG. 4. Dendrograms obtained with agglomerative methods (see text). (a) Centroid sorting, Euclidean distance; (b) information analysis; (c) centroid sorting, non-metric coefficient; (d) centroid sorting, Pearson's ϕ coefficient; (e) Ward's error sum method; (f) Lance & Williams' flexible method.

that no one species dominates the composition of the scrub. The highest percentage frequency recorded is 33% for *Barnadesia horrida*. The altitudinal range for those species with the greatest and smallest ranges respectively is listed in Tables 3 and 4. The actual occurrences of every species in relation to altitude are shown in Fig. 2. Although the species are not named, the diagram is given to illustrate the complexity of the distribution of the vegetation. Although several species combine a wide range of altitudinal occurrence with a discontinuous distribution, others can be seen mutually to replace one another in the upper and lower altitudes.

With increase in altitude there is a decrease in the number of species present. Fig. 3

plots the total number of species occurring in each 30 m (100 ft) interval of the survey. Although the most noticeable feature is the fall in species density with altitude, there are three peaks (labelled A, B and C) which might suggest optimal altitudes for specific species associations.

Table 5. Comparison of the species composition at the two cluster stage of the analysis for all agglomerative methods which did not give rise to chaining effects (only the four most frequent species are listed in each case)

| Cluster | Species | % presence | | | | | |
|-------------------|---------------------------------|--------------------|------|------|------|------|-----|
| | | Method of analysis | | | | | |
| I | | 8 | 10 | 9 | 11 | 12 | 13* |
| | <i>Schinus molle</i> | 57 | 52 | 57 | 65 | 50 | 45 |
| | <i>Psila boliviana</i> | 52 | 50 | 57 | 53 | 50 | 43 |
| | <i>Barnadesia horrida</i> | 50 | 52 | 56 | 52 | 50 | 100 |
| | <i>Baccharis cassinaefolia</i> | 40 | 41 | 35 | — | 42 | 34 |
| | <i>Acalypha aronioides</i> | — | — | — | 33 | — | — |
| | No. of quadrats | 205 | 222 | 203 | 174 | 230 | 130 |
| Mean altitude (m) | 3082 | 3064 | 3059 | 3037 | 3077 | 3107 | |
| II | <i>Astragalus garbancillo</i> | 43 | 47 | 42 | 37 | 49 | 29 |
| | <i>Cassia hookeriana</i> | 41 | 40 | 35 | 37 | 42 | 26 |
| | <i>Marrubium vulgare</i> | 30 | 34 | 31 | 28 | 30 | — |
| | <i>Berberis boliviana</i> | 28 | 26 | — | — | 28 | 26 |
| | <i>Eupatorium pentlandianum</i> | — | — | 25 | — | — | — |
| | <i>Baccharis cassinaefolia</i> | — | — | — | 27 | — | 28 |
| | No. of quadrats | 195 | 178 | 197 | 226 | 170 | 270 |
| Mean altitude (m) | 3426 | 3436 | 3406 | 3374 | 3436 | 3281 | |

* For the methods of analyses refer these numbers to those in Table 1.

Table 6. Summary of the division pattern up to the four cluster stage of the analysis for all monothetic divisive processes not giving rise to chaining effects

| Method | Cluster | | | |
|---|---------|----|-----|----|
| | I | II | III | IV |
| Association analysis—max. $\Sigma\chi^2$ | AD | Ad | aC | ac |
| Maximum information fall | AD | Ad | aC | ac |
| Association analysis max. $\Sigma(AD-BC)^2$ | AD | Ad | aF | af |
| Max. decrease error sum of squares | AB | Ab | bF | bf |
| Group analysis—divisive stage | BA | Ba | bC | bc |
| Association analysis—max. $\Sigma\sqrt{\chi^2}$ | AE | Ae | aC | ac |

The presence of a species is denoted by an upper case letter and its absence by a lower case letter. A, *Schinus molle*; B, *Barnadesia horrida*; C, *Psila boliviana*; D, *Acalypha aronioides*; E, *Berberis commutata*; F, *Astragalus garbancillo*.

Numerical analyses

From the twenty different forms of analyses listed in Table 1, six of the agglomerative and six of the divisive methods achieved a division of the quadrats without producing any chaining effects. The pattern of fusion for the last eight clusters is shown for the agglomerative methods in Fig. 4(b-f). The agglomerative process in group analysis is arrived at somewhat differently and is therefore not included here.

The pattern of division with the divisive methods is summarized in Table 6. The forty-

two different divisions of the population (seven divisions for each of the six methods) involve only six different species. It can be seen in Table 6 that the course of division is very similar for most of the methods, and in some cases is identical, e.g. association analysis using $\max \Sigma \chi^2$ and maximum information fall. To facilitate a comparison of the analytical methods employed the species composition and distribution of the subsets are presented at the two, four and eight cluster stage of the classification. The effects of

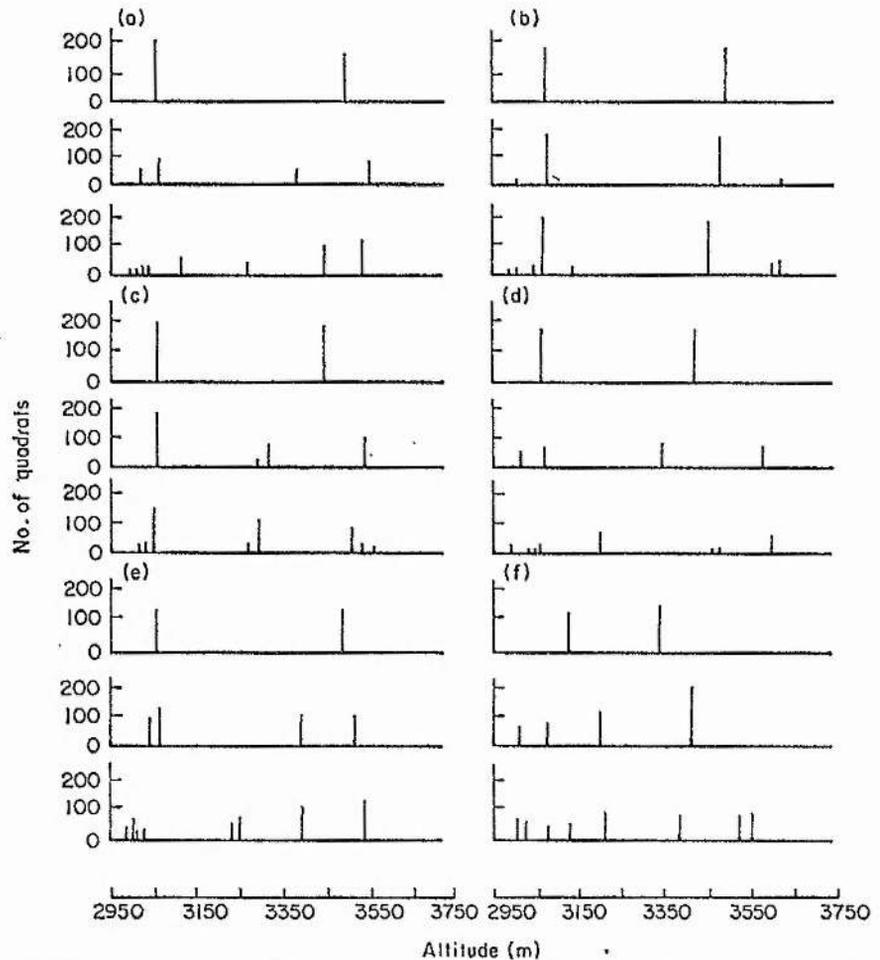


FIG. 5. Schematic representation of mean altitude and cluster size at the two, four and eight cluster stage of the *agglomerative* analyses. (a) Information analysis; (b) centroid sorting, non-metric coefficient; (c) centroid sorting, Pearson's ϕ coefficient; (d) Ward's error sum method; (e) Lance & Williams' flexible method; (f) group analysis (*agglomerative*).

division and fusion in producing these clusters can be followed for the *agglomerative* processes in Fig. 5 and for the *divisive* processes in Fig. 6.

Two cluster stage

At this stage of the analysis all the methods, *divisive* and *agglomerative*, divide the population into two more or less equal subsets, one occurring in the lower altitudinal ranges, the other in the higher ranges. For the *agglomerative* methods a summary of the floristic composition at this stage is given in Table 5, together with the mean altitude and

size of the subsets. Even at this stage in the analysis, after many separate fusions, there is a large measure of agreement between all the methods.

With the divisive methods the similarity of results between the various methods is even greater for, as can be seen from Table 6, with one single exception, all the methods make the first division of the population on the presence or absence of *Schinus molle*. Group analysis differs in splitting on *Barnadesia horrida*. However, when the results

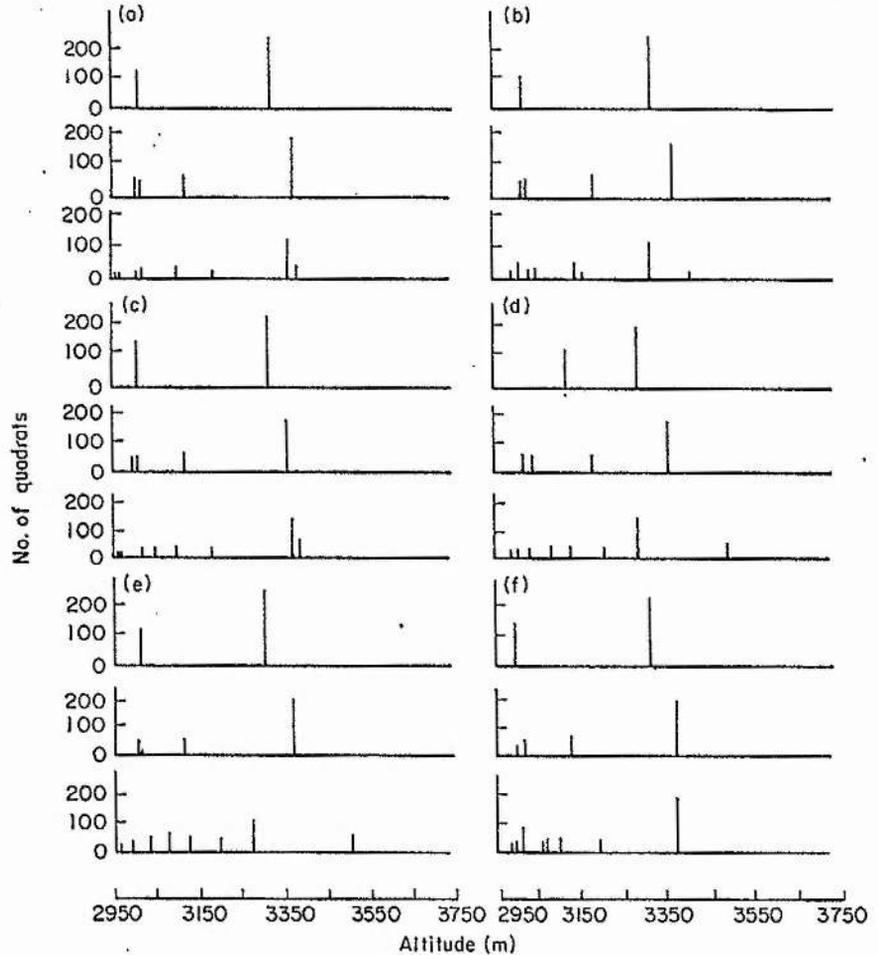


FIG. 6. Schematic representation of the mean altitude and cluster size at the two, four and eight cluster stage of the *divisive* analyses. (a) Association analysis; (b) maximum information fall; (c) association analysis max. $\Sigma (AD-BC)^2$; (d) group analysis (divisive); (e) maximum decrease error sum of squares; (f) association analysis max. $\Sigma \sqrt{\chi^2}$.

are viewed in terms of the four most frequent species in each subset (see Tables 5 and 7) it is seen that all the methods both agglomerative and divisive (with the exception of one species when using Ward's method) are identical.

Four cluster stage

At this stage of the analysis all the methods, with the exception of centroid sorting, using Pearson's ϕ coefficient, produce one high altitude group, one middle altitude group and two low altitude groups. This is most clearly seen in Figs. 5 and 6. In one case—centroid sorting using the non-metric coefficient—one of the low altitude groups is of trivial size. When the analysis is carried a stage further to the eight cluster stage the

tendency to produce trivial groups increases greatly. It appears therefore that the clearest stage for interpreting the results ecologically is at the four cluster level.

Table 7. Summary of the species composition at the two cluster stage of the analysis for all divisive methods which did not give rise to chaining effects (only the four most frequent species are listed for each cluster)

| Species | % presence | | | |
|--------------------------------|--------------------|----|-----|----|
| | Method of division | | | |
| | A | a | B | b |
| <i>Schinus molle</i> | 100 | - | 45 | - |
| <i>Barnadesia horrida</i> | 51 | - | 100 | - |
| <i>Psila boliviana</i> | 49 | - | 44 | - |
| <i>Baccharis cassinaefolia</i> | 35 | 28 | 39 | 26 |
| <i>Astragalus garbancillo</i> | - | 30 | - | 29 |
| <i>Cassia hookeriana</i> | - | 29 | - | 26 |
| <i>Marrubium vulgare</i> | - | 28 | - | - |
| <i>Berberis boliviana</i> | - | - | - | 26 |

Upper case letters denote the presence of a species, lower case letters its absence. A, *Schinus molle*; B, *Barnadesia horrida*.

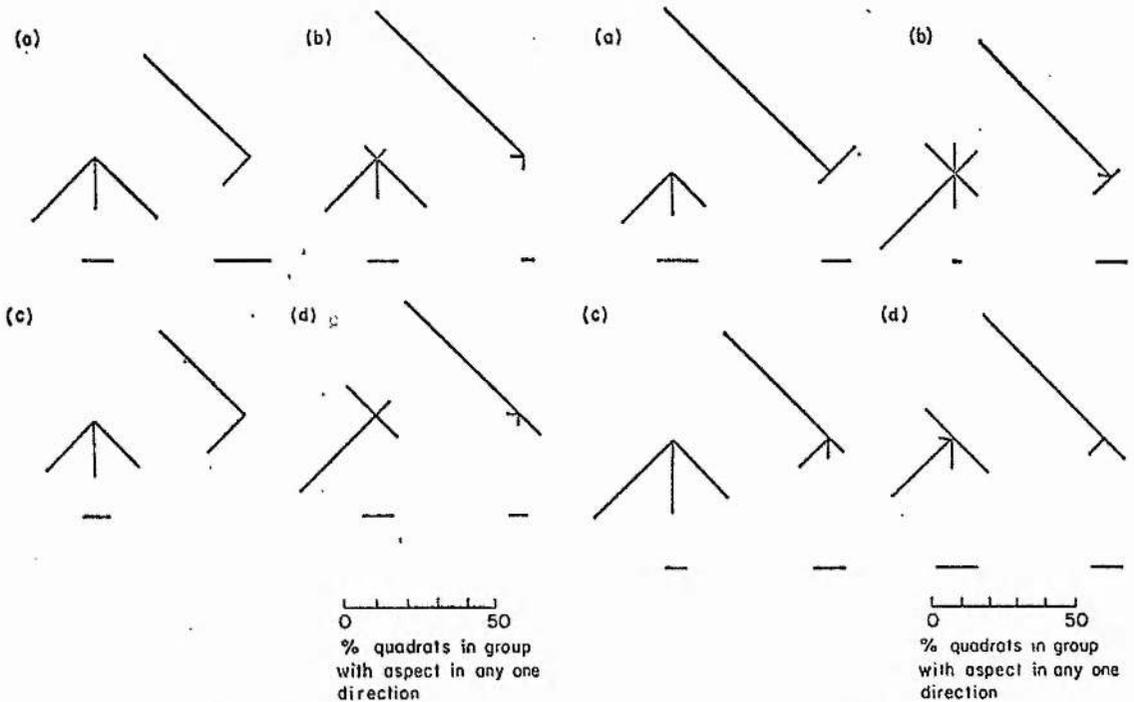


FIG. 7

FIG. 8

FIGS. 7 and 8. Distribution of quadrats in relation to aspect from pairs of low altitude groups detected at the four cluster stage as represented by a star diagram for eight compass points. The length of the line is proportional to the percentage of quadrats for that cluster with that particular aspect. The line below each star represents the percentage of quadrats in each cluster that occur on level ground.

FIG. 7. Agglomerative analyses. (a) Information analysis; (b) Lance & Williams' flexible method; (c) Ward's error sum method; (d) group analysis (agglomerative).

FIG. 8. Divisive analyses. (a) Association analysis, max. $\Sigma\chi^2$; (b) group analysis (divisive); (c) association analysis, max. $\Sigma\sqrt{\chi^2}$; (d) maximum decrease error sum of squares.

The distinction in altitude between the upper and middle range clusters is well marked but provides no distinction in the case of the two lower subsets. However, when these are

Numerical analysis of high altitude scrub

Table 8. Summary of species composition at the four cluster stage of analysis for all groups obtained by both divisive and agglomerative methods

| Species | Analytical method | | | | | | | | | | | | | | |
|---------------------------------|-------------------|------|------|------|------|------|-------------------|------|------|------|------|------|------|------|--|
| | 10 | 9 | 11 | 15 | 16 | 19 | 8 | 12 | 13 | 10 | 14 | 17 | 20 | 12* | |
| (a) > 3350 m | | | | | | | % presence | | | | | | | | |
| <i>Gynoxys nitida</i> | 91 | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| <i>Eupatorium pentlandianum</i> | 73 | 35 | 36 | - | 29 | 29 | 55 | 41 | 33 | - | - | - | - | - | |
| <i>Astragalus garbancillo</i> | - | 62 | 61 | 39 | 100 | 100 | 71 | 39 | 63 | 50 | 39 | 39 | 38 | 39 | |
| <i>Marrubium vulgare</i> | - | 43 | 50 | 30 | 36 | 36 | 64 | 55 | 44 | 36 | 30 | 30 | 26 | - | |
| <i>Berberis boliviana</i> | - | 36 | 41 | - | 27 | 27 | 45 | 48 | 34 | 28 | - | - | 26 | - | |
| <i>Cassia hookeriana</i> | - | - | - | 30 | - | - | - | - | - | 43 | 30 | 30 | 30 | 80 | |
| <i>Baccharis cassinaefolia</i> | - | - | - | 28 | - | - | - | - | - | - | 28 | 28 | 29 | - | |
| <i>Berberis commutata</i> | - | - | - | - | - | - | - | - | - | - | - | - | - | 20 | |
| No. of quadrats | 11 | 122 | 107 | 218 | 84 | 84 | 119 | 85 | 131 | 167 | 218 | 218 | 205 | 85 | |
| Altitude (m) | 3553 | 3498 | 3488 | 3487 | 3483 | 3483 | 3481 | 3479 | 3472 | 3429 | 3376 | 3376 | 3359 | 3393 | |
| | | | | | | | Analytical method | | | | | | | | |
| (b) 3350-3120 m | | | | | | | % presence | | | | | | | | |
| <i>Cassia hookeriana</i> | 93 | 73 | 93 | 73 | 67 | 31 | 31 | 54 | - | 26 | 26 | 26 | 26 | - | |
| <i>Astragalus garbancillo</i> | 23 | - | 23 | - | 27 | - | - | - | - | - | - | - | - | - | |
| <i>Baccharis cassinaefolia</i> | 32 | 48 | 32 | 48 | 43 | 35 | 35 | 50 | 34 | - | - | - | - | - | |
| <i>Psila boliviana</i> | - | - | - | - | - | 33 | 33 | 38 | 44 | 100 | 100 | 100 | 100 | - | |
| <i>Barnadesia horrida</i> | - | - | - | - | - | 32 | 32 | 41 | 100 | 47 | 47 | 47 | 47 | - | |
| <i>Berberis boliviana</i> | - | - | - | - | - | - | - | 27 | 27 | 29 | 29 | 29 | 29 | - | |
| <i>Colletia spinosa</i> | 23 | - | 23 | - | 27 | - | - | - | - | - | - | - | - | - | |
| No. of quadrats | 76 | 75 | 76 | 75 | 119 | 200 | 200 | 112 | 71 | 66 | 66 | 66 | 66 | 66 | |
| Altitude (m) | 3338 | 3289 | 3338 | 3289 | 3279 | 3250 | 3250 | 3225 | 3181 | 3125 | 3125 | 3125 | 3125 | 3125 | |

| | | Analytical method | | | | | | | | |
|---------------------------------------|------|-------------------|------|------|------|------|------|------|------|------|
| | | 13 | 14 | 17 | 20 | 16 | 19 | 15* | | |
| | | % presence | | | | | | | | |
| (c) <3120 m—northern aspect | | | | | | | | | | |
| | 12 | 8 | 11 | 13 | 14 | 17 | 20 | 16 | 19 | 15* |
| <i>Barnadesia horrida</i> | 63 | 63 | 81 | 71 | 82 | 82 | 82 | 100 | 100 | 100 |
| <i>Acalypha aronioides</i> | 38 | 46 | 63 | 59 | 100 | 100 | 100 | 53 | 53 | 53 |
| <i>Schinus molle</i> | 49 | 50 | 61 | 75 | 100 | 100 | 100 | 100 | 100 | 100 |
| <i>Psila boliviana</i> | - | - | - | 43 | 42 | 42 | 42 | 46 | 46 | 46 |
| <i>Baccharis cassinaefolia</i> | 49 | 51 | 38 | - | - | - | - | - | - | - |
| No. of quadrats | 136 | 119 | 83 | 98 | 38 | 38 | 38 | 59 | 59 | 59 |
| Altitude (m) | 3088 | 3057 | 3054 | 3037 | 3022 | 3022 | 3022 | 3018 | 3018 | 3018 |
| (d) <3120 m—southern aspect | | | | | | | | | | |
| | 12 | 20 | 13 | 8 | 11 | 14 | 17 | 16 | 19 | 15* |
| Analytical method | | | | | | | | | | |
| % presence | | | | | | | | | | |
| <i>Schinus molle</i> | 51 | 46 | 70 | 65 | 68 | 100 | 100 | 100 | 100 | 100 |
| <i>Psila boliviana</i> | 57 | 100 | 68 | 63 | 57 | 53 | 53 | 53 | 53 | 64 |
| <i>Berberis commutata</i> | 47 | 42 | 66 | 50 | 47 | - | - | - | - | 100 |
| <i>Baccharis cassinaefolia</i> | - | - | - | - | - | 39 | 39 | 39 | 39 | - |
| <i>Barnadesia horrida</i> | - | - | - | - | - | 36 | 36 | 36 | 36 | - |
| <i>Berberis boliviana</i> | 48 | - | 48 | 42 | 44 | - | - | - | - | - |
| <i>Colletia spinosa</i> | - | 34 | - | - | - | - | - | - | - | 54 |
| No. of quadrats | 94 | 65 | 59 | 86 | 91 | 78 | 78 | 78 | 78 | 28 |
| Altitude (m) | 3104 | 3068 | 3022 | 2024 | 3023 | 3012 | 3012 | 3012 | 3012 | 3010 |

* For the methods of analyses refer these numbers to those in Table 1. For brevity only the four most frequent species are listed in each case. Species with frequencies of less than 20% are also omitted. The groups are arranged in order of decreasing mean altitude: (a) greater than 3350 m; (b) between 3350 and 3120 m; (c) below 3120 m and having a predominantly northern aspect; (d) below 3120 m and having a predominantly southern aspect.

examined in relation to aspect as in Figs. 7 and 8 a marked difference is seen between them. In every case where two low altitude clusters are delimited, one member of the pair is predominantly north-west in aspect while the other has a pronounced southerly inclination. Peru being in the southern hemisphere, this will mean that the north-facing quadrats are more likely to be in the drier positions, and in contrast, owing to the steepness of the valley sides, the degree of shading on the south-facing slopes is considerable.

In Table 8 the subsets are arranged in order of descending altitude. A further distinction for north- and south-facing quadrats is made for the lower altitude sets in sections (c) and (d) of the Table. Even although the four most frequent species are listed only, it is possible to see that there is a large measure of agreement on the species composition of the high, middle and low altitude clusters. *Eupatorium pentlandianum*, *Astragalus garbancillo*, *Marrubium vulgare* and *Berberis boliviana* are consistently present in the high altitude clusters. In the middle altitude *Cassia hookeriana* and *Baccharis cassinaefolia* are prominent in the upper ranges, gradually being replaced at the lower levels by *Psila boliviana* and *Barnadesia horrida*.

In the lower altitude groups *Schinus molle* and *Psila boliviana* are common to both north- and south-facing quadrats. The north-facing quadrats, however, are characterized by the presence of *Acalypha aronioides* and *Barnadesia horrida*. This latter species always occurs with high percentage frequency in the north-facing groups and although present in some of the south-facing clusters its frequency is much less.

Although the ecological relationship is most easily seen at the four subset stage of the analysis, the percentage frequency of some of the characteristic species, particularly in the residual groups produced by the divisive processes, is somewhat low. A better resolution of the subsets in terms of species frequency is found at the next stage of the analysis.

Eight cluster stage

At this point in the analysis many of the methods tend to produce one or two large groups with a larger number of groups of trivial size. A satisfactory resolution of the subsets was, however, obtained with group analysis. This method, as it combines an ordination technique (see Crawford & Wishart 1968) along with the classification, facilitates the interpretation of the results. After the divisive stage of the analysis has produced eight subsets, every quadrat is tested by a polythetic agglomerative process to determine if it is in its group of best fit. This polythetic check is used to correct the misclassifications that are inherent in any monothetic divisive process. In this case the operation of the check has reduced the eight clusters to seven, one being absorbed into the remaining subsets. An ordination of these subsets after principal components analysis after the method of Crawford & Wishart (1968) is given in Fig. 9(a-c). (The two and four cluster sets are also given for comparison.) The centre of the circles represents the mean position of the points of each cluster and their radii the standard deviation of the points' radii from the cluster mean. The species composition of the eight cluster set after being reduced to seven groups is given in Table 9.

Fig. 9(b) shows the clear distinction between the high and middle altitude subsets and the overlap at the lower altitudes, already mentioned with the results at the four cluster stage of the analysis. When the analysis is resolved further to the eight cluster stage two groups appear in the high altitude range. The higher of these contains *Gynoxys nitida* and *Eupatorium pentlandianum*; the other contains the same species as those found at the four cluster stage of the analysis with the high altitude subset. The middle altitude set

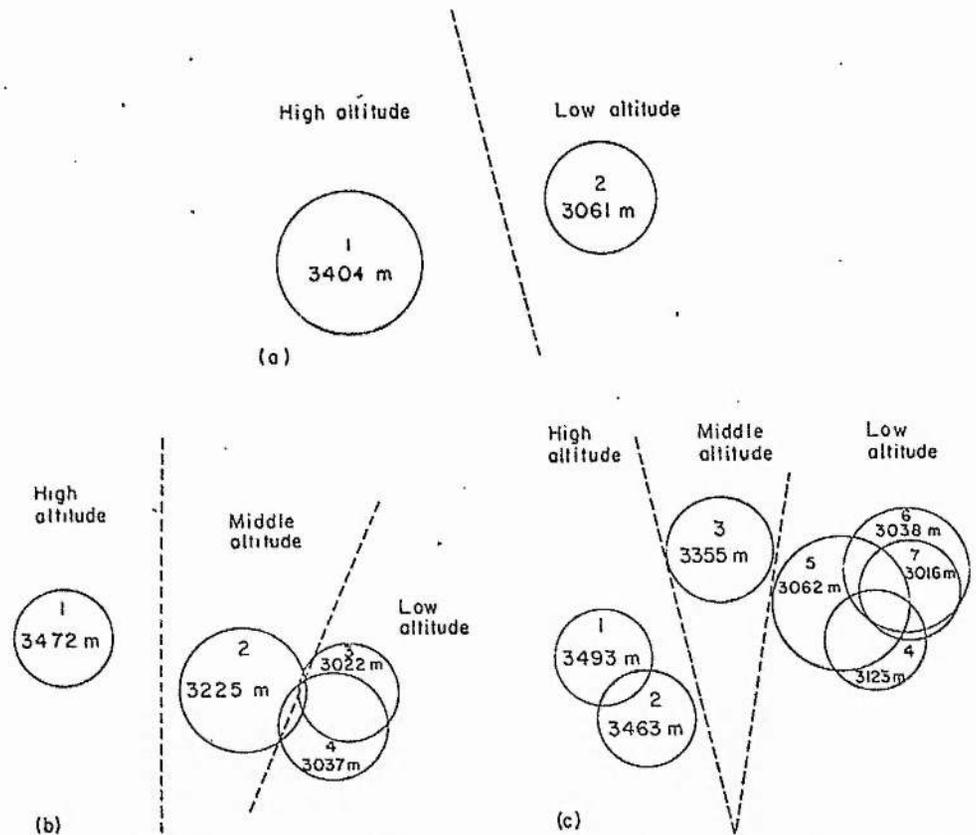


FIG. 9. Ordination drawn by computer after agglomeration and principal components analysis at (a) two, (b) four and (c) eight cluster stage of group analysis.

Table 9. Summary of the species composition of the clusters obtained after agglomeration at the eight cluster stage of the analysis by group analysis (see text)

| Species | Cluster number | | | | | | |
|-------------------------------------|----------------|------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | % presence | | | | | | |
| <i>Schinus molle</i> | - | - | - | - | 33 | 78 | 80 |
| <i>Berberis commutata</i> | - | - | - | - | - | - | 78 |
| <i>Croton ruizii</i> | - | - | - | - | - | - | 44 |
| <i>Psila boliviana</i> | - | - | - | 74 | 28 | 49 | 70 |
| <i>Acalypha aronioides</i> | - | - | - | - | - | 75 | - |
| <i>Barnadesia horrida</i> | - | - | - | 86 | - | 71 | - |
| <i>Baccharis cassinaefolia</i> | - | - | 29 | - | 46 | - | - |
| <i>Berberis boliviana</i> | - | 69 | - | 43 | 100 | - | - |
| <i>Colletia spinosa</i> | - | - | - | 31 | - | - | - |
| <i>Cassia hookeriana</i> | - | - | 100 | - | - | - | - |
| <i>Astragalus garbancillo</i> | 66 | 45 | 26 | - | - | - | - |
| <i>Baccharis salsifolia</i> | - | - | 18 | - | - | - | - |
| <i>Marrubium vulgare</i> | 23 | 80 | - | - | - | - | - |
| <i>Citharexylum argentidentatum</i> | - | 35 | - | - | - | - | - |
| <i>Eupatorium pentlandianum</i> | 69 | - | - | - | - | - | - |
| <i>Gynoxys nitida</i> | 21 | - | - | - | - | - | - |
| No. of quadrats | 61 | 51 | 77 | 35 | 61 | 69 | 46 |
| Altitude (m) | 3493 | 3463 | 3355 | 3123 | 3062 | 3038 | 3016 |

The clusters are listed from left to right in order of decreasing mean altitude. An ordination of these clusters is given in Fig. 9.

(No. 3) is clearly related to groups 4 and 5 as there is a gradual change in species composition with reduction in altitude. These lower altitude groups are also seen to be more variable, for their radii are greater than those found at a higher altitude. Groups 6 and 7 are also closely related, as can be seen in both the ordination diagram and species list. Group 6 has a predominantly northern aspect and this is again matched with the presence of *Barnadesia horrida* and *Acalypha aronioides*.

DISCUSSION

The initial aim of this survey was to determine if any structure existed in the distribution of high altitude scrub vegetation. The agreement obtained in the elimination of the

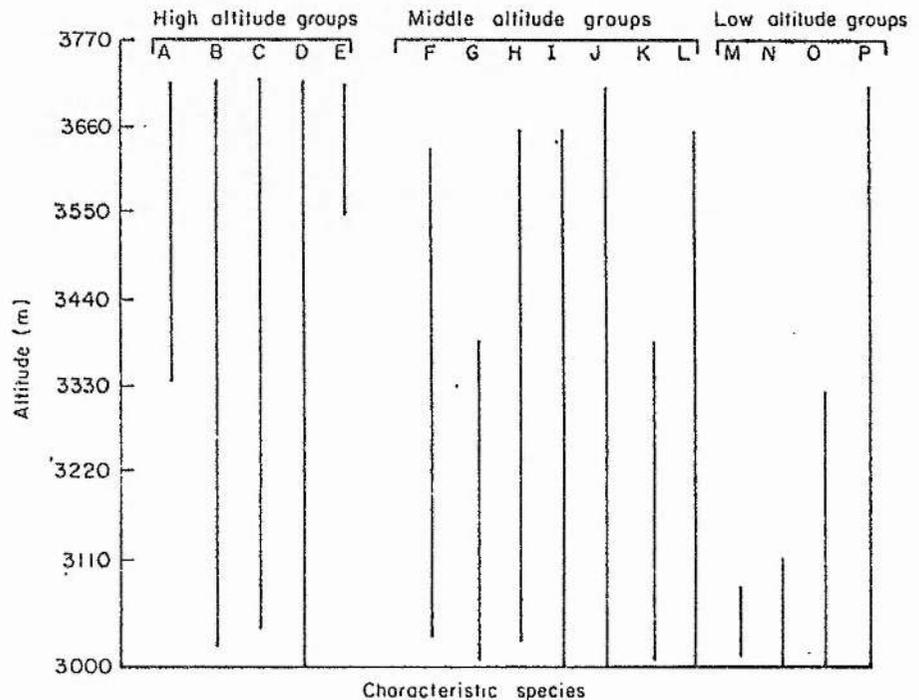


FIG. 10. Altitudinal range of characteristic species in high, middle and low altitude clusters. A, *Citharexylum argentidentatum*; B, *Marrubium vulgare*; C, *Astragalus garbancillo*; D, *Eupatorium pentlandianum*; E, *Gynoxys nitida*; F, *Baccharis salsifolia*; G, *Colletia spinosa*; H, *Cassia hookeriana*; I, *Baccharis cassinaefolia*; J, *Berberis boliviana*; K, *Acalypha aronioides*; L, *Barnadesia horrida*; M, *Croton ruizii*; N, *Schinus molle*; O, *Psila boliviana*; P, *Berberis commutata*.

subsets, particularly when examined at the two and four cluster level, establishes clearly a definite pattern of species association in relation to altitude and aspect. From the twenty different forms of numerical analysis employed, only eight had to be rejected as unsuitable. This was not because of any misclassifications or inaccuracies, but due to a total failure to resolve any clusters whatsoever. This result was to be expected with some methods owing to chaining effects and a discussion of the problem has already been given by Lambert & Williams (1966) and Wishart (1969a).

The most clearly defined pattern in the distribution of the scrub is seen at the four cluster stage of the analysis (see Fig. 9b). Here there are a distinct high altitude group and two low altitude groups with a marked similarity but differing in aspect. Between the low

and high altitude groups a gradually changing ecotone type occurs occupying the intermediate positions in relation to altitude.

It is of interest to examine the overall distribution of the species which characterize these four different groups, as it provides an indication of how the changes in scrub distribution are brought about. The altitudinal range and occurrences of these characteristic species (the four most frequent species in each of the clusters) are shown in Fig. 10.

The only species restricted exclusively to the upper ranges of the valley and characteristic of high altitude subsets are *Gynoxys nitida* and *Citharexylum argentidentatum*. These species are present only in the highest of the subsets in relation to altitude, which are only clearly seen at the eight cluster stage of the analysis. Most of the high altitude sets are characterized by the presence of *Astragalus garbancillo*, *Marrubium vulgare* and *Eupatorium pentlandianum*. It can be seen from Fig. 10 that these species have a wide range in altitude and are in fact all listed in Table 3 as belonging to the most altitudinally wide ranging species found in the survey.

Table 10. *List of species found on the survey possessing spines or thorns*

Agave americana
Barnadesia horrida
Berberis boliviana
B. commutata
Proustia pungens

A more restricted pattern of distribution is found when the characteristic species of the low level subsets is examined. Here the characteristic species are clearly restricted to the lower levels of the valley. This is especially the case with the species from the subsets found on south-facing slopes, where presumably the soils are least prone to drought (see p. 186). The north-facing quadrats of the low altitude subsets have a species composition which shares a slightly greater affinity with the middle range vegetation (i.e. presence of *Barnadesia horrida*).

Although it is beyond the scope of this descriptive enquiry to account for the mechanisms which give the varying species their different altitudinal tolerances it is of interest to note the relationship between the possession of spines and altitudinal distribution. A list of species encountered on the survey and possessing spines is given in Table 10. Here it can be seen that the shrub species in the upper vegetation types are no more spiny than those in the lower. In fact this aspect of the vegetation appears to have little influence on the composition of the scrub.

The problem of vegetation boundaries within any one life form is undoubtedly a complex one and further study would be needed before any reasons could be advanced for the different scrub zonations detected in this investigation.

The question has to be asked if, having established the existence of an ecological structure for this high altitude scrub, any immediate suggestions can be made in relation to the conservation of the vegetation and the preservation of the soil. An examination of the upper and middle altitude subsets shows two species, *Astragalus garbancillo* and *Cassia hookeriana*, both members of the Leguminosae, to be constant features of the vegetation. *Astragalus garbancillo* is by no means a conspicuous plant and without the analysis the extent to which it contributes to the vegetation, particularly at high altitudes, would not have been obvious. It could therefore be suggested that in an attempt to preserve this scrub vegetation these two leguminous species should be encouraged.

Not only will they probably fix nitrogen, but as their altitude range is as great as any of the other species encountered on the survey, they should prove both hardy and responsive to attempts to increase their distribution.

It is of interest to note the relatively high frequency of occurrence of two introduced species, namely *Spartium junceum* and *Marrubium vulgare*. The latter species is particularly well established at the higher altitudes while *Spartium junceum* is found only in the lower regions of the valley. *Eucalyptus globulus* is also not native to South America. Although this species was encountered in the area (relative frequency 1%) it is more typically established below the level at which this survey was carried out.

As was pointed out in the Introduction, the number of numerical methods used in this investigation allows some comparison to be made of their relative efficiency. In a survey of this type where there are a large number of samples but only a few species the divisive methods lend themselves to the most rapid computation of the data. However, when the relative homogeneity of the resulting subsets is examined in terms of relative species frequency the results obtained are not as good as those with the agglomerative methods. The latter, however, demand considerably more computing time. A useful compromise is found in group analysis (Crawford & Wishart 1967, 1968) where the divisions are carried out rapidly irrespective of the number of species in the survey and the results, as seen here, compare closely with those obtained by other divisive methods. The polythetic agglomerative check, already mentioned (p. 186), gives the added advantage of the greater accuracy in cluster definition of polythetic methods, without taking up much computing time as would be required with a normal agglomerative analysis.

ACKNOWLEDGMENTS

We are much indebted to Professor Cesar Vargas for invaluable assistance in the field and in naming the specimens in the herbarium in Cuzco. Our thanks are also due to Mr Douglas Gifford (Spanish Department, University of St Andrews) whose organization and leadership of the expedition made this investigation possible. Financial assistance from the Carnegie Trust for the Universities of Scotland is also gratefully acknowledged.

SUMMARY

A survey of the distribution of high altitude scrub vegetation growing at the tree-line was carried out in a densely settled valley in south-eastern Peru. A lengthy period of human settlement has reduced the tree-line by nearly 850 m (2800 ft) below the climatic optimum. The scrub vegetation left covering much of the valley floor and mountain sides is thought to be essential to the maintenance of adequate water reserves in the soil for agriculture as well as for preventing a rapid acceleration of erosion. No previous ecological study has been reported for this scrub and this investigation uses two different numerical methods in an attempt to relate the distribution of the scrub species associations to altitude and aspect. Characteristic scrub types are found in the upper and lower regions of the valley. These are described together with transitional types found in the middle altitudinal ranges and suggestions are made in relation to conservation of the scrub and the prevention of further erosion.

REFERENCES

- Crawford, R. M. M. & Wishart, D. (1967). A rapid multivariate method for the detection and classification of groups of ecologically related species. *J. Ecol.* 55, 505-24.
- Crawford, R. M. M. & Wishart, D. (1968). A rapid classification and ordination method and its application to vegetation mapping. *J. Ecol.* 56, 385-404.
- Ellenberg, H. (1958). Wald oder Steppe? Die natürliche Pflanzendecke der Anden Perus. I and II. *Umschau*, 645-8, 679-81.
- Garcilaso de la Vega (1608). *The Royal Commentaries of the Incas*. English edition by the Hakluyt Societies, London 1869.
- Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 23, 623-37.
- Herrera, F. L. (1941). *Sinopsis de la Flora del Cuzco*. Lima.
- Holdridge, L. R. (1947). Determination of world plant formations from simple climatic data. *Science* N.Y., 105, 367-8.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32, 241-54.
- Lambert, J. M. & Williams, W. T. (1966). Multivariate methods in plant ecology. VI. Comparison of information analysis and association analysis. *J. Ecol.* 54, 635-64.
- Lance, G. N. & Williams, W. T. (1965). Computer programs for monothetic classification ('Association analysis'). *Comput. J.* 8, 246-9.
- Lance, G. N. & Williams, W. T. (1966). Computer programs for hierarchical polythetic classification ('Similarity analyses'). *Comput. J.* 9, 60-4.
- Lance, G. N. & Williams, W. T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.* 9, 373-80.
- Lance, G. N. & Williams, W. T. (1968). Note on a new information statistic classificatory program. *Comput. J.* 11, 195.
- Macnaughton-Smith, P. (1965). Some statistical and other numerical techniques for classifying individuals. Home Office: *Studies in the Causes of Delinquency and the Treatment of Offenders*. Vol. 6 pp. 1-33. London.
- Orlaci, L. (1967). An agglomerative method for classification of plant communities. *J. Ecol.* 55, 193-200.
- Prescott, W. H. (1847). *History of the Conquest of Peru with a Preliminary View of the Civilization of the Incas*. London.
- Sokal, R. R. & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *Kans. Univ. Sci. Bull.* 38, 1409-38.
- Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco.
- Sørensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation. *Biol. Skr.* 5, 1-34.
- Rowe, J. H. (1946). Inca culture at the time of the Spanish conquest. *Handbook of the South American Indians*. Bull. 143, Vol. II. Washington.
- Tosi, J. A. (1960). *Zonas de Vida Natural en el Peru*. Bull. 5 Instituto Interamericano de Ciencias Agrícolas de la OEA Zona Andina.
- Vargas, C. (1966). Síntesis de la flora de las provincias de Canas, Espinar y Chumbivilcas. *Revta Univ. Cuzco Universitaria*, No. 126-29.
- Ward, H. H. (1963). Hierarchical grouping to optimise an objective function. *J. Am. statist. Ass.* 58, 236-44.
- Williams, W. T., Lambert, J. M. & Lance, G. N. (1966). Multivariate methods in plant ecology. V. Similarity analyses and information analysis. *J. Ecol.* 54, 427-45.
- Wishart, D. (1969a). An algorithm for hierarchical classification. *Biometrics*, 25, 165-70.
- Wishart, D. (1969b). Fortran II programs for 8 methods of cluster analysis (CLUSTAN I). *Kans. Geol. Comp. Contr. Ser.* 38.

(Received 3 March 1969)

Methods of Deriving Multi-Factor Uniform Regions

D. C. D. POCOCK, M.A., PH.D.

(Lecturer in Geography, University of Dundee)

AND D. WISHART, B.SC.

(Computer Programmer, University of St. Andrews)

Revised MS. received 4 June 1968

ABSTRACT—This paper introduces a new method of obtaining multifactor uniform regions. Those fusion techniques such as 'centroid' and 'group average', which are based on the imposition of minimum-variance constraints, may well generate artificial classifications, while the step-wise clustering procedure and its corresponding dendrogram facility is inefficient when dealing with large data sets. The new method, termed the dense-space method, searches initially for dense spheres which signal the presence of important uniform regions or dense space, and then derives distinct regions by linking any dense spheres which intersect. Three classification levels are suggested: nuclear, basic and complete. The nucleus of each distinct region is described by a set of intersecting dense spheres of radius $\frac{1}{2}r$, each of which has the property that it does not intersect any dense sphere from any other distinct region. The subset of sample points inside the dense spheres are termed nuclei points and constitute a cluster. Those points which lie outside the dense spheres but at a distance not greater than r from the centre of a dense sphere are included with the classification for the sphere at the basic level. Points which are unclassifiable at the basic level are relatively remote and their classification at the complete level, into the regions which contain the points' nearest dense spheres, should only be used when a best fit is demanded for every sample.

The dense-space method achieves more 'natural' classifications and demands less computation and memory storage. Its advantages over other methods are shown by a reworking of U.S.A. census data, first used by B. J. L. Berry, and by reference to an urban survey of Middlesbrough.

THE PROBLEM of regional classification in geography is essentially the classification problem common to all the behavioural sciences. A set of samples (observation units) is divided into a small number of subsets or clusters so that each subset represents a grouping of samples which have a basic common similarity with respect to the survey variables. Classifications involving total uniformity in the cluster samples, that is, every variable having uniform values for each subset, are derived by imposing constraints on the cluster's overall variance. Two such techniques, centroid and group-average, are compared here before introducing a third method, that of dense space. The new method, it is claimed, achieves more 'natural' classifications and is suitable for rapid analysis of large surveys.

Computation details are given for data used by B. J. L. Berry in a recent paper¹ from nine census divisions of the U.S.A., and the speed facility of the dense-space method when used on large surveys is demonstrated with reference to a 231-sample urban survey. A formal presentation of the methods is given in a mathematical appendix.

Data Preparation

In general, n samples are classified according to their values measured for m variables, which for this article are of the numerical type. In Table I six variables (the numbers of service

establishments per 1000 population for six categories) are measured for nine samples (census divisions of the U.S.A.).

An essential notion for any cluster analysis method, by which similar samples are grouped together, is the measure of the similarity between two samples, or two groups. Many coefficients have been proposed; the coefficient adopted here, and also used by Berry, is the 'squared Euclidean distance', d^2_{ik} , which is discussed in detail elsewhere.² The coefficient can be obtained by summing over all m variables the squared differences between each variable observation for the i -th and k -th samples, but when the raw data are used, a bias is introduced in favour of those variables with high variance. This bias may be eliminated by an initial standardization

TABLE I
Services per Thousand Population for U.S.A. Census Divisions, 1954

| <i>Census division</i> | 1 <i>Personal</i> | 2 <i>Business</i> | 3 <i>Auto. repair</i> | 4 <i>Misc. repair</i> | 5 <i>Amusement</i> | 6 <i>Hotels, etc.</i> |
|------------------------|----------------------|----------------------|------------------------------|------------------------------|-----------------------|--------------------------|
| 1. New England | 2.56 | 0.57 | 0.53 | 0.69 | 0.43 | 0.46 |
| 2. Middle Atlantic | 2.70 | 0.72 | 0.54 | 0.72 | 0.41 | 0.25 |
| 3. E.N. Central | 2.10 | 0.50 | 0.52 | 0.68 | 0.46 | 0.30 |
| 4. W.N. Central | 2.11 | 0.47 | 0.71 | 0.84 | 0.56 | 0.53 |
| 5. S. Atlantic | 1.74 | 0.38 | 0.49 | 0.53 | 0.42 | 0.42 |
| 6. E.S. Central | 1.38 | 0.25 | 0.38 | 0.41 | 0.33 | 0.22 |
| 7. W.S. Central | 2.04 | 0.45 | 0.68 | 0.80 | 0.45 | 0.40 |
| 8. Mountain | 1.92 | 0.57 | 0.70 | 0.78 | 0.55 | 1.24 |
| 9. Pacific | 2.37 | 0.87 | 0.82 | 0.87 | 0.51 | 0.63 |

Source: Statistical Abstract of the United States (1959)

of the variable distributions. Of the various standardization procedures, the most appropriate here is the reduction of each variable distribution to unit variance and zero mean. The resultant standard scores are given in Table IIIb.

The samples are now represented by points with standard coordinates in standardized Euclidean space, and the redefined distance d^2_{ik} is adopted as the coefficient of similarity between samples U_i and U_k . A notion fundamental to 'dense space' is that, if a significant distance limit, r say, is chosen, then two samples are said to be 'similar' if $d^2_{ik} \leq r^2$.

The Role of Principal Components Analysis

Berry suggests that a transformation to principal component scores will eliminate the redundancies incurred when several variables display a single pattern of concomitant variation. Each pattern of correlated variables is replaced by a single component which represents the pattern, and the point distribution can be described approximately in terms of a smaller number of uncorrelated component variables. It is certainly true that, as Berry claims, the transformation will in some instances save considerable computation, especially when a large number of initial variables is used. The analysis is also of interest in its own right, because inevitably any classification obtained from the data will be a function of the initial variables, and the isolation of the major factors present as a result of the choice of variables gives an indication of the terms of reference to which the classification applies (the classifications obtained from quadrat

sampling of a town on socio-economic variables may be completely different from those derived, for example, from health variables). It cannot be stressed too strongly that the results

TABLE II
Factor Loadings (Eigenvectors) obtained by Principal Components Analysis for U.S.A.
Census Data used in Table I

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------|-------|-------|-------|-------|------|-------|
| Factor: 1 | 0.30 | 0.40 | 0.46 | 0.48 | 0.43 | 0.31 |
| 2 | -0.63 | -0.40 | 0.14 | -0.05 | 0.33 | 0.54 |
| 3 | -0.16 | -0.42 | 0.31 | 0.38 | 0.24 | -0.69 |
| 4 | 0.54 | -0.49 | -0.50 | 0.08 | 0.41 | 0.14 |
| 5 | -0.21 | 0.43 | -0.24 | -0.39 | 0.67 | -0.31 |
| 6 | 0.37 | -0.23 | 0.58 | -0.67 | 0.08 | -0.02 |

TABLE III
(a) Factor Scores for Six Factors of Nine U.S.A. Census Divisions

| Census Division (Sample) | Factors | | | | | |
|--------------------------------|---------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | -0.03 | -1.1 | -0.50 | 0.59 | -0.23 | 0.11 |
| 2 | 0.20 | -2.1 | -0.39 | 0.11 | 0.00 | -0.06 |
| 3 | -0.62 | -0.35 | 0.30 | 0.28 | 0.36 | -0.17 |
| 4 | 1.4 | 0.76 | 1.0 | 0.47 | 0.20 | 0.07 |
| 5 | -1.9 | 0.58 | -0.15 | -0.03 | 0.21 | 0.15 |
| 6 | -4.1 | 0.59 | -0.11 | -0.45 | -0.04 | -0.03 |
| 7 | 0.24 | 0.13 | 0.89 | -0.21 | -0.57 | -0.04 |
| 8 | 1.9 | 2.1 | -1.0 | 0.20 | -0.10 | -0.07 |
| 9 | 2.8 | -0.55 | -0.09 | -0.98 | 0.17 | 0.04 |

(b) Standard Scores for Six Variables of Nine U.S.A. Census Divisions

| Census Division (Sample) | Variables | | | | | |
|--------------------------------|-----------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1.19 | 0.22 | -0.51 | -0.08 | -0.40 | -0.11 |
| 2 | 1.55 | 1.09 | -0.43 | 0.12 | -0.69 | -0.83 |
| 3 | 0.00 | -0.18 | -0.58 | -0.15 | 0.03 | -0.66 |
| 4 | 0.02 | -0.35 | 0.87 | 0.97 | 1.48 | 0.12 |
| 5 | -0.94 | -0.87 | -0.82 | -1.2 | -0.54 | -0.25 |
| 6 | -1.87 | -1.63 | -1.66 | -2.07 | -1.85 | -0.94 |
| 7 | -0.16 | -0.47 | 0.64 | 0.69 | -0.11 | -0.32 |
| 8 | -0.47 | 0.22 | 0.79 | 0.55 | 1.34 | 2.55 |
| 9 | 0.69 | 1.96 | 1.71 | 1.19 | 0.75 | 0.46 |

obtained from any classification technique are dependent on the original choice of variables, and therefore the derivation of 'meaningful' principal components can be extremely helpful where clarification of the frame of reference is required.

In principal components analysis,³ the original coordinate axes are rotated to a new set of orthogonal axes so that the major axis (factor) is the line of best fit through the point swarm (that is, it accounts for the maximum amount of variance), and successive factors are similar lines of best fit subject to the constraint that they must be, in each case, orthogonal to each of their predecessors. The result can be demonstrated by a set of points which lie in a plane through a three-dimensional space. The first axis will lie along the line of best fit, the second is orthogonal to the first, and the third which must be orthogonal to the plane that contains the points is therefore redundant. If factor scores (coordinates) are obtained for the first two principal axes, then the distance between any two points under this system will be identical to the distance measured in the original three-variable system. In the general case, the first few components will usually account for a large proportion of the overall variance in the point distribution, and when scores, computed on these factors alone, are used for measuring similarity distances, good approximations to the true distances in m -space are achieved. This reduction from m variables to a few (f , say) factors corresponds to a projection of the point swarm from m -space into f -space with the minimum possible distortion of the point orientation. Factor loadings, or eigenvectors, obtained from the correlation matrix for the six variables are shown in Table II, and the corresponding transformations to factor scores are given in Table IIIa.

The contributions of components 5 and 6 to the distance coefficient for any two samples may be seen to be very small by comparison with the four major components. There is little difference between the distance measures obtained using (a), all six standard scores and (b), the first four component scores. When distances are obtained using all six factor scores (as adopted by Berry), the results are identical to those derived from the standard scores, and in this instance, the principal components analysis is ineffective.

There seems to be no clear rule for determining the number of factors that should be chosen to define the components' subspace. H. F. Kaiser⁴ suggests a rule for principal components analysis where significant components are those which account for an eigenvalue not less than unity, but whether this rule should be adopted for classification methods is doubtful, for, on the basis of the unity rule, only the first two factors obtained from the present census data would be adopted (Table IV). While it is true that combined they account for 88.5 per

TABLE IV
Analysis of Variance of U.S.A. Census Data

| Factor | Variation Explained | | |
|--------|---------------------|------------|-----------------------|
| | Total | Percentage | Cumulative Percentage |
| 1 | 3.94 | 65.7 | 65.7 |
| 2 | 1.37 | 22.8 | 88.5 |
| 3 | 0.39 | 6.5 | 95.0 |
| 4 | 0.22 | 3.6 | 98.6 |
| 5 | 0.07 | 1.2 | 99.8 |
| 6 | 0.01 | 0.2 | 100.0 |

cent of the overall variance, the sizeable contributions to the distance measures of the scores for factors 3 and 4 in Table IIIa suggest that, for classification purposes, the unity rule would involve an over-simplification and create excessive distortion. However, the dichotomy that exists concerning the choice of relevant components is not present in this instance, for clearly

too many components cannot be selected. The suggestion by D. F. Morrison⁵ that components should be chosen which together explain some arbitrary percentage of the total variance seems to be more pertinent to classification methods, and a level of 90 or 95 per cent of the variances would be reasonable.

More recently, Berry⁶ has proposed an additional standardization of the component scores prior to classification. This effectively destroys all relationship between distance measures obtained using the original standard scores and the new component scores. The components are standardized in such a way that they have equal importance, a state which is clearly not substantiated by the many applications of principal components analysis in this subject. Problems now arise concerning the number of components which should be used, and the incorporation of components which do not have meaningful interpretations. But what is most important is that the technique has the effect of creating a 'synthetic' frame of reference in which inter-sample similarities no longer correspond to the observed relationships. Classification methods which derive similarity measures from eigenvectors normalized in this way would appear to be invalid, and the procedure should be avoided.

The adoption by D. M. Ray and Berry⁷ of an additional rotation from the principal components solution to a normal Varimax frame of reference has certain advantages concerning the interpretation of factors. This may be adopted when meaningful principal components cannot be derived and a factor analysis appraisal of the regional structure patterns is desired. It is not clear, however, whether Ray and Berry use scores computed from Varimax factors for their similarity measures. If this is the case, and the only axes rotated are those corresponding to the f eigenvectors which would otherwise be used to compute distance similarities, then the distances using the f Varimax factor scores will be the same as those derived from the f major component scores. On the other hand, if more than f axes are rotated to a Varimax solution, then more dimensions will usually be required to compute accurate distances since the Varimax rotation does not result in an optimal variance solution as obtained by principal components. The effect of Varimax is to share out the large variance explained by the major components among the lesser components in order to obtain factors which lend themselves to easier interpretation. Distances calculated from the major Varimax factors are still good approximations to those obtained using standard scores, but are less accurate than those derived using principal components loadings.

It is therefore recommended that, when a reduction in the number of dimensions used to compute distance similarity coefficients is desired, then factor scores obtained from those eigenvectors associated with the major principal components, which together account for an arbitrary proportion of the overall variance, should be used. A Varimax solution may be obtained as an auxiliary investigation but should not be used in conjunction with classification procedures.

General Comments on Classification Techniques

Cluster analysis methods can be grouped into two general categories according to whether one is classifying small sets or large sets of samples. When the number of samples is small, the analyst is often interested in the relationships between individual samples. Several linkage, or agglomerative, methods have been proposed along the lines that the population of samples is progressively fused into a diminishing number of groups, so that the relative similarity between individual samples is indicated by the order of their fusion (or the fusion of the groups which contain the individual samples). Groups are compared pair-wise by some notion of inter-group

similarity usually based on a specific similarity coefficient measured for sample pairs, one from each group. The two most similar groups are found, combined to form one single group, and thus the procedure passes to the next fusion step. The order of fusion can be represented graphically by a dendrogram or 'linkage tree', whereby the fusion of two groups at a particular stage is shown by a joint or node connecting the two sub-branches which represent the groups. The samples that constitute a group can be easily seen as a growth of branches from the sub-branch representing the group back to the original sample points. Fusions are shown in chronological order from bottom to top, and for some analyses the rise from one fusion level to the next is used as an indication of the loss of homogeneity in the new group, caused by its formation. This will be discussed in greater detail below. Some writers base their comparisons of different fusion methods on the resulting dendrogram structures and have shown that several fusion methods are variants of a single general system involving four parameters.⁸

The dendrogram method of representing individual fusions becomes clumsy when dealing with large sets of data, where the analyst is more interested in the latter stages of a procedure when the survey set has been reduced to a few fairly large groupings of samples. The fusion methods are no less effective for these applications, but the calculation involved becomes tedious, inefficient and demands a large computer memory store. To date, no method of the fusion type can accommodate a survey of more than about 500 samples using existing computers without implementing magnetic backing store as an auxiliary memory and thereby increasing the computation time out of all proportion. With this in mind, the writers have been concerned with the development of alternative methods which require both less computation and less memory storage, with a resulting economy and increase in the maximum size of survey that can be accommodated. One such method, introduced here and termed the dense-space method, sacrifices the step-wise clustering technique and its corresponding dendrogram facility and, instead, generates classifications which in this instance are very similar to the sample groupings obtained at the latter stages of some fusion processes. This is demonstrated here by a comparison of the results obtained using two fusion methods, centroid and group-average, with the classifications derived by dense space for both the American census data and a 231-case study of Middlesbrough. The dense-space method is introduced here for numerical variables only, although applications involving binary data are possible, and an attempt is made to establish a theoretical approach to the classification problem. The specific problem involving binary data for which attributes have equal importance is examined elsewhere.⁹

The Method of Dense Space

(a) General

In recent years the emphasis in numerical classification has been directed towards the derivation of clusters which have some minimum-variance partition properties. This concept, referred to here as space-conservation, can be illustrated by the synthetic two-dimensional scatter distributions in Figure 1A. The three distinct clusters of points have the property that their overall variances (the average of their squared deviations from each point to its cluster mean) are small. The reason for this is that the two-variable variances for each cluster distribution are significantly smaller than their variances for the overall population, and consequently each cluster appears as a small spherical grouping of points. The variables are termed 'diagnostic', that is, their values for a cluster are uniform and therefore contribute towards the isolation of the character of the cluster. J. H. Ward¹⁰ proposes a hierarchical fusion method for

obtaining such clusters by minimizing an 'error sum of square objective function' which is the sum of the squared distances from each point to its parent cluster mean. Other writers¹¹ have concentrated on defining iterative solutions to achieve the same ends. M. J. Shepherd¹² seems to have been the first to recognize that the important region of a cluster is its densest part or nucleus. He isolates cluster nuclei by applying the accepted technique 'single-linkage' at a high level of similarity and then rejects those members of a cluster for which the average of their

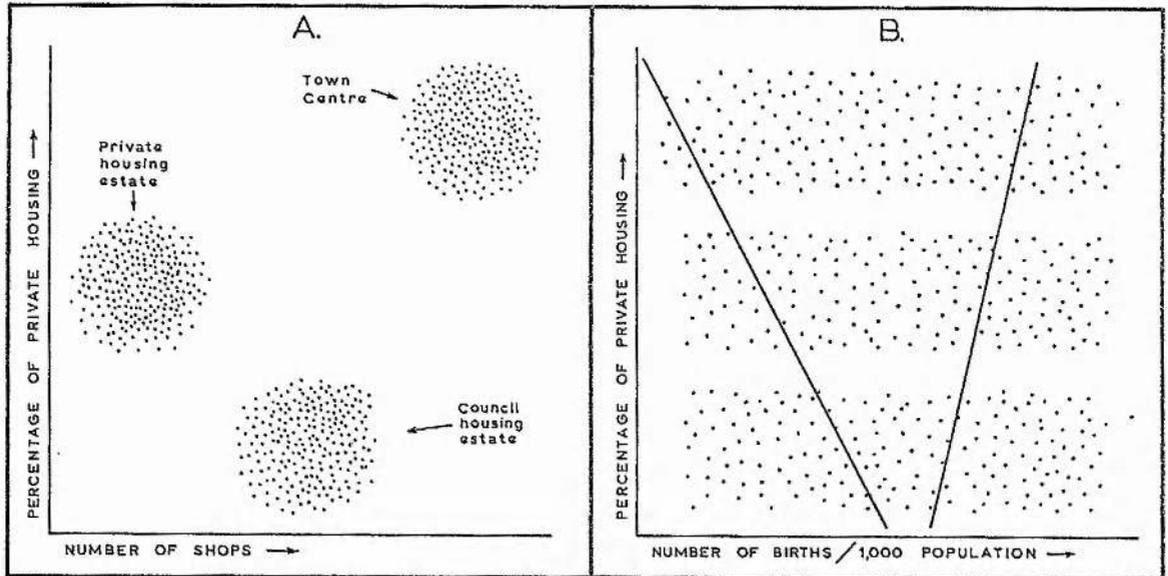


FIGURE 1—A hypothetical sample distribution showing the elongation effects on well-defined clusters owing to the introduction of an unrelated variable

similarities with the other cluster members is less than a critical threshold. When the coefficient d^2_{ik} is used, it is easily shown that rejection occurs for those points whose distance from the cluster centre exceeds the cluster's standard deviation, that is, if σ_c^2 is the variance of a cluster, then those members of the cluster which lie outside the sphere whose centre is the cluster mean and radius σ_c are thrown out. The resultant constraint on cluster variance is clear.

The question which now arises is what happens when a 'natural' cluster cannot be diagnosed or characterized by all the variables. Suppose a variable is added to the data which is completely unrelated to the study, for example, the number of births per 1000 population is measured for each unit of a town in addition to the variables of Figure 1A. The scatter diagram of housing and birth of Figure 1B shows that the three previous spherical clusters are transformed into elongated swarms of points which no longer possess a minimum-variance property. The three-dimensional scatter diagram would repeat this pattern, and the conclusion to be drawn is that a 'natural' cluster should not necessarily have a minimum-variance property, but would appear as a connected swarm of points which is separated from any other such swarm. Furthermore, those methods which impose minimum-variance constraints on such a distribution may well generate artificial classifications; for example, the two partition lines in Figure 1B might yield three clusters having variances which are smaller than the three 'natural' cluster swarms.

This notion that natural clusters occur as elongated swarms of points is not just valid when irrelevant variables are present. Classification techniques are most often applied to data involving several or many variables and the assumption, when using a space-conserving method,

that each resultant cluster or region should have a significantly small variance for every variable seems unreasonable, however careful the original choice of variables. It is to be expected that, while one region may be characterized by uniform values for a sub-group of the variables, another region is identified by a different sub-group of variables or the same sub-group of variables taking different values. In either case, some variables can be expected to be unrelated to the character of the region concerned and reflect the general variation of the population as a whole. The theoretical principle proposed is that, when a set of points in the sample space forms a single dense connected swarm (mode), each point is one instance of a single phenomenon (region, taxon, etc.). When several modes are present, each represents one phenomenon having a complex of interactions (in the ecological sense) with the others, thus sustaining the system as a whole. It is the objective of the dense-space classification method to resolve these distinct phenomena as dense-sample distribution modes, and it is the job of the analyst to interpret their characteristics and interrelationships.

(b) *Theory*

A distinct region is considered a group of samples for which some of the variables are uniform. Such variables are denoted diagnostic and are identified by the significantly small variance of their distributions for the group's sub-set of sample points. The remaining non-diagnostic variables are permitted to have distributions with high variance, the criterion being that a cluster need not be identified by uniform values for every variable. A uniform region, on the other hand, is the special case of a distinct region for which all variables are diagnostic, and it is this concept together with some derived results which forms the basis of the method.

Associated with a uniform region is the idea of a uniform space, which in general is a set of points X (of the standardized space) such that every point in X is similar to every other point in X . In this context, X does not refer to a finite sub-set of sample points, but rather the region of space which may contain the sub-set. For applications involving numerical variables, X will be an infinite set of points for non-trivial similarity definitions, but in other cases (for example, involving binary data) X may be finite.

Under the similarity coefficient d^2_{ik} in the standardized space, a sphere K of radius $\frac{1}{2}r$ satisfies the definition of a uniform space by virtue of the fact that the distance between any two points in K never exceeds r . Two points U_i, U_k contained in K are therefore similar since $d^2_{ik} \leq r^2$.

A sub-set of points which constitutes a distinct region for h diagnostic variables can be projected from m -dimensional standardized space into the h -space of the diagnostic variables to form a uniform region. The definition of a distinct region is thus a set of points X which, for the sub-space of h diagnostic variables, has the property that every point in X is similar to every other point in X . Such a projection has the effect of selecting from the full variable list only those variables which have characteristic or uniform values for the region.

(c) *Analysis Objectives*

Because the analyst is concerned with finding the general characteristics of a group of points, an important factor is the number of points that make up the group. For the spherical uniform space discussed here, a sphere containing only a few points cannot be regarded as as important as one which has a much higher density. A uniform region which is 'important', one which represents a large sub-set of samples, can be detected by a uniform space (a sphere) which contains a large number of points. The spherical structure of a uniform region is not, however,

a property of a distinct region in m -space because the $m-h$ non-diagnostic variables may have high variance and hence the spatial distribution of points is not constrained. In fact, no general structure can be expected or should be induced, and the only generalizations that can be made are that the points representing a distinct region constitute a band or amorphous swarm which is continuously dense throughout and at the same time separated from any other such grouping by non-dense space. The following points may be added to clarify what is meant by natural and artificial classifications. A distinct region should not be partitioned at any point where its distribution of sample points is dense. Two distinct regions which are separated by non-dense space should not be combined. Finally, individual points which do not occur in dense space must be deemed unclassifiable and can only be associated with the nearest distinct region in a loose sense.

The dense-space method searches initially for dense spheres, which signal the presence of important uniform regions or dense space, and then derives distinct regions by linking any dense spheres which intersect. In this way, any orientation of points which is sufficiently dense can be described with no restrictions being imposed on scatter in the distribution. The first step is to select a threshold value r (the critical similarity level) and construct a grid of n' spheres such that every sample point lies inside at least one sphere. The density of each sphere is found and a critical density level k selected. Those spheres having a density less than k are rejected as being insufficiently dense to constitute an important uniform space, and from the remainder any spheres which intersect are linked to form the nuclei of distinct regions. The nucleus of each distinct region is therefore eventually described by a set of intersecting dense spheres, each of which has the property that it does not intersect any dense sphere from any other distinct region. Three classification levels are now suggested: nuclei, basic and complete classifications.

The sub-set of sample points which lie inside the dense spheres of a distinct region constitutes a cluster and the points are termed nuclei points. Any points which lie outside the dense spheres are deemed non-nuclei or unclassifiable at this level. However, the clusters obtained are often very small and may not serve any useful purpose other than to delimit the very central positions of distinct regions. Attention is therefore turned to basic classifications.

A certain dichotomy exists between the definitions of a uniform space and a uniform region. In this context, a sphere of radius $\frac{1}{2}r$ containing a sub-set X of sample points may exclude points close to its perimeter which satisfy the definition of a uniform region with respect to X . In Figure 2, each point inside the sphere has a distance from the point P which is less than r , and the inclusion of P in the uniform region would therefore be in order. To extend the classification to account for such irregularities would require an examination of the relationships between every non-nucleus point with every sub-set of nuclei points which represents a

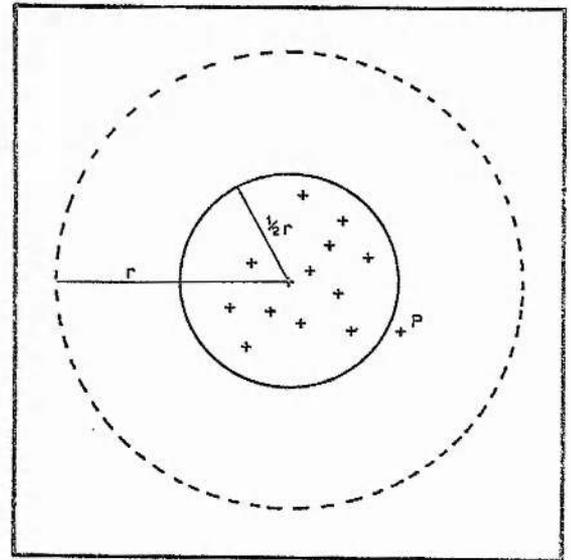


FIGURE 2—Misclassification generated by dense space at nucleus level. Dashed line delimits cluster spherical space at basic level.

uniform space, and such an investigation would defeat the main asset of the method—namely its speed. A compromise solution is therefore proposed whereby any points which lie a distance not greater than r from the centre of a dense sphere are included within the classification for the sphere (that is, within the sphere of radius r dotted in Figure 2). In some instances, non-nuclei points will be classified by this means with dense spheres from more than one distinct region, that is, when the distance between two dense spheres from two distinct regions is less than $2r$. In such a case, the point is always associated with its nearest dense sphere. Any remaining points which lie a distance greater than r from their nearest dense spheres are again denoted unclassifiable at the basic level.

For the level of complete classification, the above restriction is removed, that is, every non-nucleus point is classified into the region which contains the point's nearest dense sphere. It should be stressed that points which are unclassifiable at the basic level should be regarded as relatively remote and their classification at the complete level should only be used where a best fit is demanded for every sample.

(d) Cluster Diagnosis

At each level of classification distinct regions are resolved as sub-sets of sample points, and in order to determine a region's diagnostic variables (if indeed any exist) the variances of the standardized variable values for the sub-set distribution must be obtained. A significance level S^* is selected, and diagnostic variables are those whose variances S^2 , do not exceed S^* .

(e) The Dense-Space Method Applied to the U.S.A. Census Data

Using North American census data, the sphere radius $(\frac{1}{2}r)^2 = 5.1$ is selected, necessitating the construction of five spheres to enclose the nine standardized sample points. The grid of spheres is formed by choosing a starting point (in this case point 1) as centre of the first sphere, that point which is farthest from point 1 as centre of the second sphere, and so on, at each stage the next sphere being constructed about the point which is farthest from its nearest sphere. The process is concluded when the maximum distance from any point to its nearest sphere is less

TABLE V
Dense Space Analysis of U.S.A. Census Data for Five Spheres with $(\frac{1}{2}r)^2 = 5.1$. (d^2_i = Squared Distances of Sample Points from Point i ; d^2_{min} = Squared Distances of Sample Points from Nearest Sphere)

| Spheres | | | | | | | | | | Classifications for $k = 2$ | | | |
|--|---------|---------|---------------|---------|---------------|---------|---------------|---------|---------------|-----------------------------|----|----------|----------|
| | 1 | 2 | | 3 | | 4 | | 5 | | DENMIN | SP | ICLUS(N) | ICLUS(B) |
| | d^2_1 | d^2_6 | $d^2_{min_A}$ | d^2_8 | $d^2_{min_B}$ | d^2_9 | $d^2_{min_C}$ | d^2_4 | $d^2_{min_D}$ | | | | |
| Sample point 1 | 0.0 | 20.9 | 0.0 | 15.1 | 0.0 | 11.5 | 0.0 | 8.4 | 0.0 | 0.0 | 1 | 1 | 1 |
| 2 | 1.5 | 26.9 | 1.5 | 22.3 | 1.5 | 11.1 | 1.5 | 12.6 | 1.5 | 1.5 | 1 | 1 | 1 |
| 3 | 2.1 | 14.1 | 2.1 | 14.9 | 2.1 | 14.1 | 2.1 | 6.2 | 2.1 | 2.1 | 1 | 1 | 1 |
| 4 | 8.3 | 33.3 | 8.3 | 6.7 | 6.7 | 7.3 | 6.7 | 0.0 | 0.0 | 0.0 | 4 | 1 | 1 |
| 5 | 7.2 | 5.1 | 5.1 | 18.7 | 5.1 | 25.3 | 5.1 | 13.2 | 5.1 | 5.1 | 6 | 2 | 2 |
| 6 | 21.0 | 0.0 | 0.0 | 40.9 | 0.0 | 50.5 | 0.0 | 33.3 | 0.0 | 0.0 | 6 | 2 | 2 |
| 7 | 4.4 | 20.7 | 4.4 | 11.1 | 4.4 | 9.5 | 4.4 | 2.9 | 2.9 | 2.9 | 4 | 1 | 1 |
| 8 | 15.1 | 40.8 | 15.1 | 0.0 | 0.0 | 10.4 | 0.0 | 6.7 | 0.0 | 6.7 | 4 | 0 | 1 |
| 9 | 11.6 | 50.5 | 11.6 | 10.4 | 10.4 | 0.0 | 0.0 | 7.2 | 0.0 | 7.2 | 4 | 0 | 1 |
| Sphere density $(\frac{1}{2}r)^2 = 5.1$ | 4 | 2 | | 1 | | 1 | | 2 | | | | | |

than $(\frac{1}{2}r)^2$, that is, when every point is enclosed. In Table V the arrays d_i^2 contain the distances of the points from each sphere, and d_{\min}^2 , updated at each formation, shows the current distances of the points from their nearest spheres. Termination occurs when the maximum value in d_{\min}^2 is not greater than 5.1. The density of each sphere is the number of points which the sphere encloses (those points with d_i^2 not greater than 5.1). In the example (Table V) the five spheres are constructed about points 1, 6, 8, 9 and 4 and have densities 4, 2, 1, 1 and 2 respectively. A density level $k = 2$ is chosen and the third and fourth spheres are rejected as being insufficiently dense. From the remainder, those spheres about points 1 and 4 are found to intersect (since $d_{14}^2 = 8.3$ is less than $r^2 = 20.4$) and delimit the nucleus of cluster 1, while the sphere about point 6 constitutes the cluster 2 nucleus.

Table V also shows the distance DENMIN of each point from its nearest dense sphere (SP), and nuclei classifications ICLUS (N) are obtained for those points which are enclosed by the dense spheres. Note that points 8 and 9 are excluded at this level. For basic classification, ICLUS (B), the sphere radius is doubled to $r^2 = 20.4$ and points 8 and 9 are found to be classified with cluster 1. In the event that any distances DENMIN exceed 20.4, complete classification would be required to group such points and is achieved by removing the distance restriction on DENMIN. In this instance basic classification is sufficient, although for $k = 3$, sample point 6 fails to be classified, except at the complete level.

(f) Storage Requirements

The minimum computer storage commitment for this procedure when only one level of density is tested is seen to be d^2 , d_{\min}^2 , DENMIN, SP and ICLUS. ICLUS can share space with d^2 , so that in general, when p density levels are investigated (DENMIN and SP must be duplicated at each level), the total 'large' storage commitment is $2n(1+p)$ values.

The Fusion Methods

(a) Centroid

For this method, similar samples are fused together and replaced by new synthetic points, at the samples' centroid, in a succession of steps until the whole population is replaced by a single group-individual at its centroid. The technique, first suggested in the context of socio-economic regionalization by Ray and Berry,¹³ can be summarized in the following stages.

First, the triangular matrix D of squared distances d_{ik}^2 is computed for all $\frac{1}{2}n(n-1)$ possible pairs of points in the standardized space. Secondly, to each point unit weight is allocated and the two nearest points U_i , U_k (corresponding to the smallest d_{ik}^2 in D) found. These two points are replaced by a new synthetic individual at their centroid with a weight of 2. The rows and columns in D corresponding to U_i and U_k are combined and replaced by distances from every other point to the new centroid point, and the dimension of D has effectively been reduced from n to $(n-1)$. The process is then repeated, the two nearest points at each stage being fused to form a new group-individual. This can be achieved by a transformation in D from the distances of every point U_r from U_i and U_k to the distance of U_r from their centroid $U_{(ik)}$ (Appendix 6). The matrix D can be modified by the transformation so that no actual computation of centroid coordinates is required.

With the American census data, the original matrix of squared distances D is given in Table VI; the first fusion concerns points 1 and 2. The new point, midway between the two original points, is coded 1 and has a weight of 2. The distances of the other points from this new

point replace the column corresponding to the original point 1 (Table VI, penultimate column), and the column and row corresponding to point 2 becomes redundant. The simplest way of retaining the valid sections of the matrix D is to use a weight vector $W(i)$ which is updated at

TABLE VI
U.S.A. Census Data: Matrix of Squared Distances between Samples' Points

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | First Fusion Modification | |
|---|-------|-------|-------|-------|-------|-------|-------|-------|------------------------------|------------------|
| | | | | | | | | | Centroid | Group Average |
| 2 | 1.54 | | | | | | | | — | — |
| 3 | 2.10 | 4.73 | | | | | | | 3.03 | 3.41 |
| 4 | 8.38 | 12.58 | 6.19 | | | | | | 10.10 | 10.49 |
| 5 | 7.19 | 12.47 | 3.05 | 13.18 | | | | | 9.45 | 9.83 |
| 6 | 20.95 | 26.97 | 14.09 | 33.31 | 5.07 | | | | 23.58 | 23.97 |
| 7 | 4.38 | 7.49 | 2.48 | 2.93 | 6.77 | 20.70 | | | 5.55 | 5.94 |
| 8 | 15.10 | 22.26 | 14.92 | 6.72 | 18.68 | 40.87 | 11.05 | | 18.30 | 18.69 |
| 9 | 11.58 | 11.08 | 14.06 | 7.26 | 25.28 | 50.54 | 9.48 | 10.38 | 10.94 | 11.33 |

each fusion to contain the weight of each current point, or zero if the point has become redundant. In this case, after the first fusion $W(1) = 2$, and $W(2) = 0$, the remaining values being unity. In computing the least d_{ik}^2 at subsequent fusion steps only elements of D corresponding to nonzero $W(i)$ and $W(k)$ are considered.

(b) *Group Average*

This method is identical to centroid except that the measure of distance between two groups of points is the average of the squared distances between all possible pairs of points, one from each group. The technique is attributed to R. R. Sokal and C. D. Michener¹⁴ and was first adopted for the problem of economic regionalization by Berry.¹⁵ The formula for modifying the array of squared distances D on fusion of points U_i, U_k , for another U_r , is given in Appendix 6.

The first step, using group average for the U.S.A. census data, again fuses points 1 and 2, but the vector of average distances from the other points to the new group (final column, Table VI) shows the difference between the methods. In fact, results obtained by the two methods differ very slightly for the census data.

In addition to weight vector W , and the triangular matrix of similarities D , an array ICLAS must be stored to keep track of the sample groupings obtained by either fusion method. The storage commitment for these methods is therefore $\frac{1}{2}n(n+3)$ values.

(c) *Dendrograms*

The complete analysis of a small data set can be described comprehensively using a dendrogram, as discussed earlier. In the instances of centroid and group-average, the least-distance measure at each fusion can be used as an indication of the homogeneity of the resulting group. This is shown on the dendrogram by setting a vertical distance scale and marking every node on the diagram at the distance level corresponding to the squared distances between the two points whose fusion the node represents. The dendrograms for centroid and group-average fusions on the U.S.A. census data are shown in Figure 3.

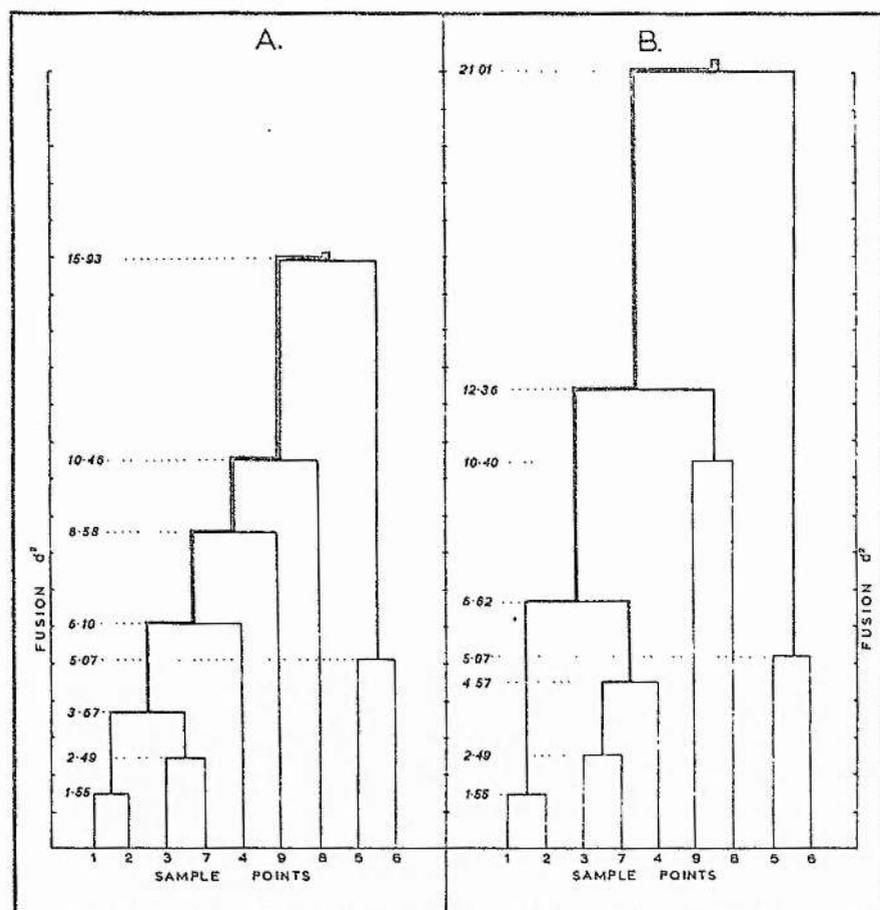


FIGURE 3—Dendrogram of U.S.A. census data for (A) centroid and (B) group-average methods

Remarks

The calculation procedure for dense space using the census data shows the inadequacy of the method when applied to small data sets. Although the classifications obtained for $k = 2$ are the same as those derived by both centroid and group-average at the seventh fusion (when the two major groups remain), it is easily shown that, given different starting conditions for the sphere construction, other results are obtained. When large surveys are analysed, however, such irregularities are reduced and more acceptable classifications derived.

W. T. Williams, J. M. Lambert and G. N. Lance¹⁶ suggest that a classification method should derive 'well marked groups at well separated levels'. At first glance it seems that the group average method gives the best classification of the U.S.A. census data, deriving three distinct sample groupings at the sixth fusion. From the dendrograms, however, it can be seen that a high rise from one fusion level to the next marks a considerable increase in the heterogeneity of the resulting group. This scaling of the dendrogram can help the analyst decide at which stage the fusion process should be stopped. It is now clear from the dendrograms that the relationships of samples 8 and 9 to the group 1, 2, 3, 7, 4 are fairly similar. Centroid fuses the samples to the group individually, while group-average forms the group 8, 9 first and then fuses the two groups. Since the level of fusion for the two groups by group-average is a

relatively small increase on the previous level, it is clear that the validity of 8, 9 as a separate group is questionable. It is also obvious, from inspection of the two dendrograms, that the two final groupings 1, 2, 3, 7, 4, 8, 9 and 5, 6 are mutually dissimilar and probably constitute the realistic classification.

The dense-space method can be considered as initially performing a local centroid-type grouping in so far as points are grouped inside small spherical similarity regions and lose their individual identities. The major difference between the two techniques is that, once this localized grouping is complete, dense space imposes no further constraint on the scatter of a cluster of points. The action of centroid or group average on a long dense band of points can be visualized as a condensation process by which, at an intermediate stage, the band has degenerated into a number of spaced-out globules. Such partitioning of a sample point swarm is considered here to constitute an artificial classification of the data, and by comparison, provided the swarm is continuously dense throughout, dense space would isolate it in its entirety.

The computation time-saving feature of dense space takes effect when the number of samples n is very large, greater than perhaps 200. For surveys of this order, the calculation of the $\frac{1}{2}n(n-1)$ values of the similarity matrix D becomes tedious and the storage commitment ratio $2n(1+p)/\frac{1}{2}n(n+3) = \frac{4(1+p)}{3+n}$ demonstrates that when a small number of density levels is investigated, the dense-space method is considerably more economical with computer space and can therefore accommodate much larger surveys.

Example of Dense-Space Classification with a Large Set of Data

The results of the different methods of obtaining uniform regions when dealing with a large number of samples will be illustrated by a study of the urban character of Middlesbrough, a town with a population of about 157,000 in the North Riding of Yorkshire. The settlement was founded as a new town in 1830, as a suitable point for exporting coal by the river Tees, but owed its substantial growth to the success of its iron, and subsequently, steel industry. In the last third of the last century, ten separate iron-smelting plants were in operation within the present town boundaries; in 1961, there remained only two pig-iron producers and one integrated iron and steel works, although the industry was still the single most important employer of labour.

The general outline of the urban structure derived from the industrial growth has been relatively simple (Fig. 4). The town has grown progressively southward from the initial settlement by the Tees, being in the early years confined between ironworks both to the west and east. By 1914, there was a broad advance as far south as Albert Park. While areas adjacent to the perimeter of the park, together with the wedge of most southerly development at Linthorpe, consisted of villas, the bulk—including the formerly separate town of North Ormesby and small growth at Cargo Fleet—was formed of terracing, compactly laid out on a grid-type system. In the north-central section of this early development is the town centre, which is cruciform in shape. Development since the first world war has been progressively away from the amenities and services of the town centre but, in return, open space has been more generously provided. Local authority housing, dispersed on several separate estates in the inter-war period, has since been concentrated in the eastern part of the borough, while the most extensive and select area of recent private development is in the extreme south-west.

Aspects of the urban character have been extracted in eight chosen variables, the data having

been compiled on a grid basis. The origin of the lattice was located in relation to the national grid, each kilometre square being divided into sixteen. That is to say, the 231 sub-areas or samples exhibiting a degree of residential function measured 250 x 250 m, or one-sixteenth of a 1:2500 Ordnance Survey plan, from which much of the information was compiled.

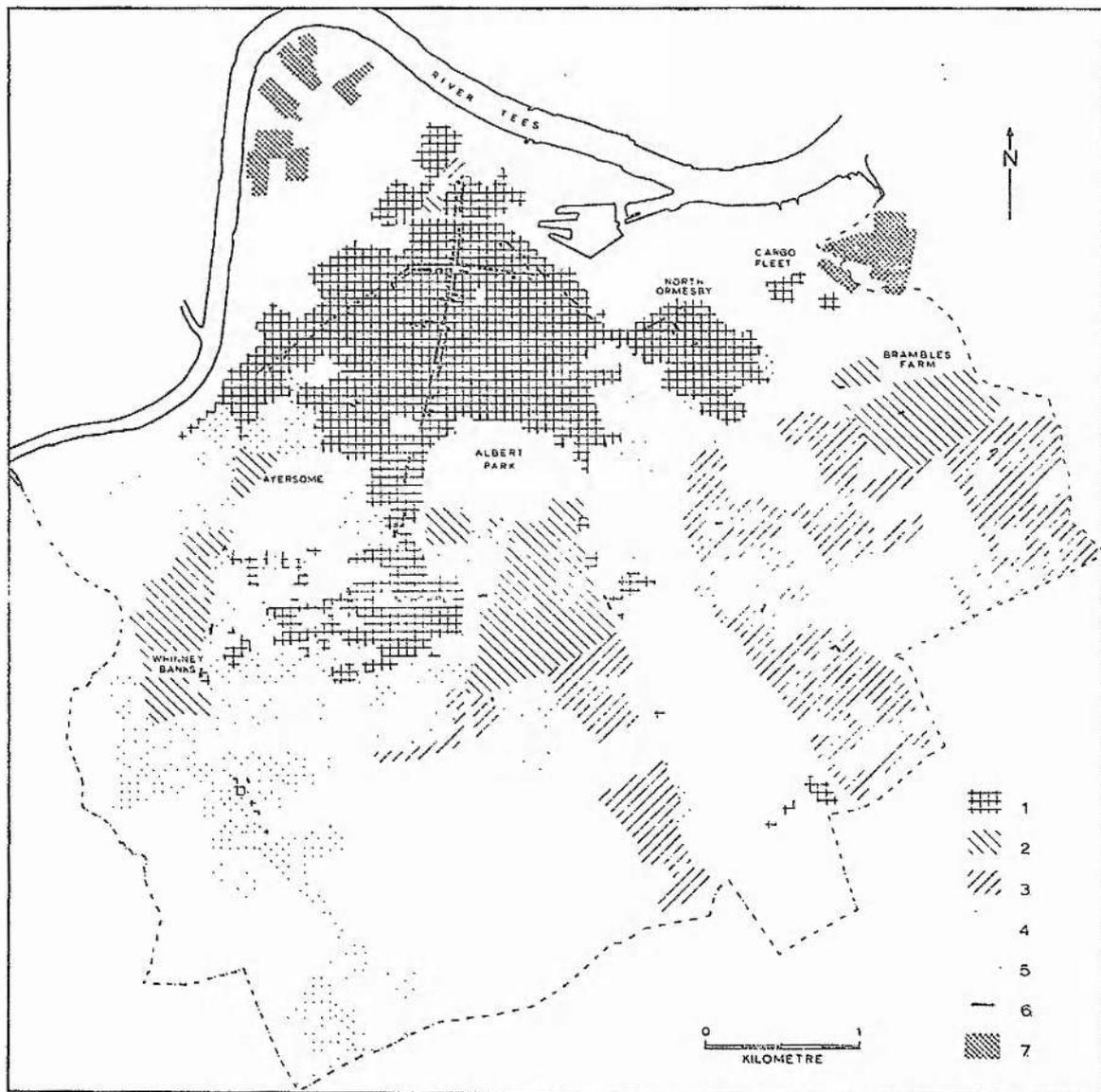


FIGURE 4—Some features of the urban structure of Middlesbrough, showing locations mentioned in the text. 1—Pre-1918 building; 2—inter-war corporation housing; 3—post-war corporation housing; 4—inter-war private housing; 5—post-war private housing; 6—main shopping streets; 7—iron and steel works

An initial descriptive measure of the pair-wise relationship between the eight variables is given in the correlation matrix (Table VII). Half of the twenty-eight coefficients have absolute values greater than 0.5, only four of 0.25 or under. The extent of intercorrelation is shown by the resultant factor values (Table VIII). The first factor, accounting for over half of the total variability, emphasizes the age, density and distance variables (Table IX). The second factor,

TABLE VII
Matrix of Correlation Coefficients for Middlesbrough Data

| | Age | Per cent non-municipal housing | Per cent terraced housing | Accessibility to open space | Population density | Accessibility to shops | Distance to town centre | Distance to iron or steel works |
|--|-------|--------------------------------|---------------------------|-----------------------------|--------------------|------------------------|-------------------------|---------------------------------|
| Age ¹ | 1.0 | | | | | | | |
| Per cent non-municipal housing | -0.57 | 1.0 | | | | | | |
| Per cent terraced housing | -0.34 | -0.08 | 1.0 | | | | | |
| Accessibility to open space ² | 0.57 | -0.25 | -0.39 | 1.0 | | | | |
| Population density | -0.58 | 0.19 | 0.53 | -0.42 | 1.0 | | | |
| Accessibility to shops ³ | -0.72 | 0.40 | 0.40 | -0.43 | 0.65 | 1.0 | | |
| Distance to town centre (miles) | 0.84 | -0.42 | -0.46 | 0.62 | -0.64 | -0.72 | 1.0 | |
| Distance to iron or steel works (miles) | 0.53 | 0.04 | -0.52 | 0.49 | -0.70 | -0.49 | 0.67 | 1.0 |

Notes: ¹ Years from 1830; ² Open space within $\frac{1}{4}$ mile; ³ number within $\frac{1}{4}$ mile.

TABLE VIII
Factor Values for Middlesbrough Data

| Factor | Variation Explained | | |
|--------|---------------------|-----------------------|-------|
| | Percentage | Cumulative Percentage | |
| I | 4.53 | 56.6 | 56.6 |
| II | 1.37 | 17.1 | 73.7 |
| III | 0.64 | 8.1 | 81.8 |
| IV | 0.51 | 6.3 | 88.1 |
| V | 0.33 | 4.2 | 92.3 |
| VI | 0.32 | 3.9 | 96.2 |
| VII | 0.17 | 2.1 | 98.3 |
| VIII | 0.13 | 1.7 | 100.0 |

TABLE IX
Factor Loadings for Middlesbrough Data

| | I | II | III | IV | V | VI | VII | VIII |
|-----------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| Age | 0.41 | -0.27 | 0.02 | -0.06 | 0.19 | 0.33 | -0.57 | 0.53 |
| Per cent non-municipal housing | -0.19 | 0.71 | 0.06 | -0.23 | 0.46 | -0.14 | -0.41 | -0.03 |
| Per cent terraced housing | -0.28 | -0.46 | -0.01 | -0.79 | 0.05 | -0.23 | -0.08 | -0.07 |
| Accessibility to open space | 0.33 | 0.00 | 0.83 | -0.02 | -0.09 | -0.43 | 0.02 | 0.04 |
| Population density | -0.38 | -0.19 | 0.39 | 0.12 | 0.59 | 0.37 | 0.36 | 0.14 |
| Accessibility to shops | -0.38 | 0.13 | 0.38 | -0.09 | -0.57 | 0.52 | -0.28 | -0.09 |
| Distance from town centre | 0.43 | -0.09 | 0.07 | -0.10 | 0.24 | 0.34 | -0.08 | -0.78 |
| Distance from iron or steel works | 0.36 | 0.37 | -0.03 | -0.52 | -0.11 | 0.32 | 0.53 | 0.27 |

MULTI-FACTOR UNIFORM REGIONS

the only other one to account for more than unit variance, has a strong bias to private housing, which is associated negatively with the extent of terraced housing.

Although principal components analysis is ideally suited to an initial stage of inquiry by the collapse of numerous variables into a smaller number of independent factors showing the extent of common variance, some authors have taken the individual factor weightings as a basis for regional division, using the quartiles for a simple divisive technique.¹⁷

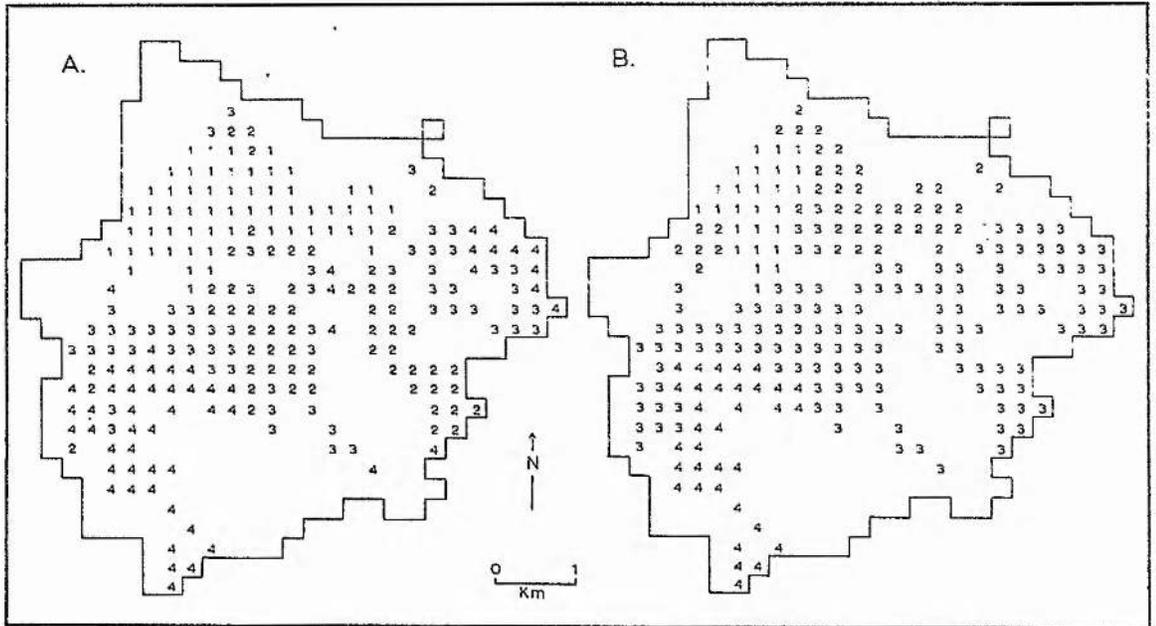


FIGURE 5—Division of Middlesbrough 1961 into urban regions by principal components analysis: (A) first component, quartile division; (B) first and second component

In a map of Middlesbrough based on the first component (Fig. 5A) the northern part of the town is clearly separated from the progressively younger development to the south. In the eastern quarter, however, the presence of regions for three of the quartile divisions suggests a surprising range in an area entirely of corporation development, largely post-war in age. The emphasis of the component on the distance variables¹⁸ accounts for the distribution of 4s around the southern perimeter, grouping as one both private and corporation development of varied age. Emphasis on type of housing in the second component confines the region of 4s to the south-west quarter in a map derived from a visually acceptable grouping of a dispersion of values for the first two components (Fig. 5B). The division, however, is less satisfactory than the previous one; not only is the northern portion split into two, but the broad central belt of 3s gives an exaggerated impression of homogeneity.

Divisions based on the centroid and group-average methods, both using similarity distance coefficients measured on the six major component scores, are shown for the last nine groupings (Fig. 6), this being the lowest acceptable fusion level in both instances. Three regions are identical on both maps: a town centre (region 1) and the inter-war and post-war private estates in the south-west (3 and 4 respectively). The centroid method, however, has three groupings which account collectively for a mere four samples, and as a result it fails to single out the earliest inter-war corporation estates—Ayresome, Whinny Banks and, particularly, Brambles Farm. This grouping, which is detected by the group average method (region 6 on Fig. 6B), is

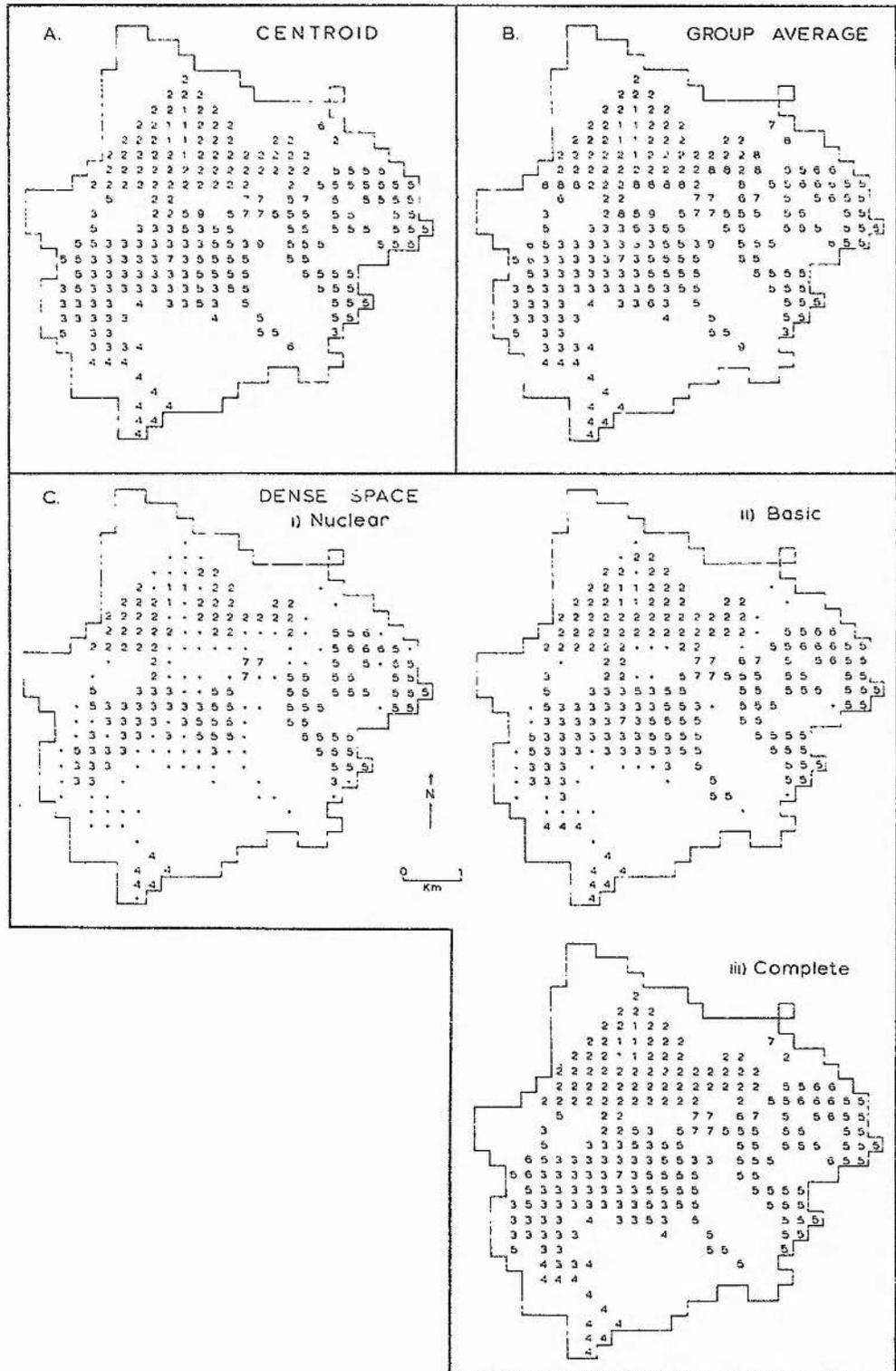


FIGURE 6—Division of Middlesbrough into urban regions: (A) centroid method, six major component scores, last nine groups; (B) group average method, six major component scores, last nine groups; (C) dense-space method, 120 spheres, critical density = 3, at nuclear, basic and complete levels

MULTI-FACTOR UNIFORM REGIONS

generally distinct from its post-war counterpart, especially in the proportion of terraced dwellings (Table Xb). The group average method also distinguishes a separate region (8) along

TABLE X
Mean Variate Values for Urban Regions of Middlesbrough derived from Centroid, Group Average and Dense-Space Methods
(a) Centroid

| Variable | Urban Region | | | | | | | | |
|-------------------------|--------------|-------|-------|--------|-------|--------|--------|--------|--------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Age | 40.5 | 58.4 | 98.8 | 118.9 | 114.2 | 48* | 97.8 | 104.9* | 86.0 |
| Private housing | 100* | 97.9* | 94.3* | 94.6 | 3.0* | 100* | 98.8* | 100* | 87.4* |
| Terraced housing | 100* | 98.3* | 71.2 | 20.7 | 82.4 | 0* | 13.3 | 84.9* | 65.5* |
| Open space | 3.3* | 26.1 | 45.0 | 169.9 | 73.4 | 50* | 60.8 | 209.9* | 159.9* |
| Population density | 1030.0 | 972.9 | 450.3 | 329.5* | 606.9 | 955.0* | 558.1* | 504* | 547.0* |
| Access to shops | 281.6 | 88.3 | 7.6* | 1.1* | 8.0* | 9* | 9.3* | 0* | 6.0* |
| Distance to town centre | 0.18* | 0.68 | 1.78 | 2.73 | 1.91 | 1.38* | 1.22* | 2.59* | 1.20 |
| Distance to iron works | 0.64* | 0.80 | 1.97 | 2.97 | 1.37 | 0.11* | 1.23 | 2.33* | 1.43* |
| Total samples | 6 | 65 | 53 | 13 | 84 | 1 | 6 | 1 | 2 |

* Denotes values with a variance ≤ 0.25 .

(b) Group Average

| Variable | Urban Region | | | | | | | | |
|-------------------------|--------------|--------|-------|--------|-------|-------|-------|-------|--------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Age | 40.5 | 52.4 | 98.8 | 118.9 | 115.6 | 105.5 | 90.7 | 80.2 | 92.3 |
| Private housing | 100* | 97.7* | 94.3* | 94.6 | 3.1* | 2.3* | 99.0* | 98.6* | 91.6* |
| Terraced housing | 100* | 99.5* | 71.2 | 20.7 | 88.8 | 43.7 | 11.4 | 94.0 | 72.0 |
| Open space | 3.3* | 12.3 | 45.0 | 169.9 | 74.3 | 68.3 | 59.2 | 76.4 | 176.6 |
| Population density | 1030.0 | 1012.9 | 450.3 | 329.5* | 602.0 | 630.0 | 614.8 | 826.9 | 532.6* |
| Access to shops | 281.6 | 98.4 | 7.6* | 1.1* | 7.7* | 9.7* | 9.2* | 51.5 | 4.0* |
| Distance to town centre | 0.18* | 0.59 | 1.78 | 2.73 | 1.92 | 1.86 | 1.24* | 1.01 | 1.66 |
| Distance to iron works | 0.64* | 0.79 | 1.79 | 2.97 | 1.43 | 1.01 | 1.07 | 0.88 | 1.73 |
| Total samples | 6 | 51 | 53 | 13 | 72 | 12 | 7 | 14 | 3 |

* Denotes values with a variance ≤ 0.25

the south and south-eastern perimeter of the town as it was in 1918. This seems more justified on the northern perimeter of Albert Park than elsewhere; in this grouping of only fourteen samples it is interesting to note that only a single variable has a variance of less than 0.25. This

may be compared with the northern region (2), from which it is differentiated, where there are two diagnostic variables in a grouping with almost four times the number of samples.

Classification by the dense-space method, using 120 spheres with a critical density level of three, produces a seven-fold division (Fig. 6c) which combines the best of the two earlier groupings. Single sample regions are avoided, but the earliest inter-war corporation estates are distinguished. The northern, pre-1914 section of the town (2) differs from that of the centroid method by a single sample, which has been taken from the town centre (1). Similarly the two

TABLE X (cont.)
(c) Dense Space

| Variable | Urban Region | | | | | | |
|-------------------------|--------------|-------|-------|--------|-------|--------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Age | 37.2 | 58.3 | 97.3 | 119.4 | 115.4 | 105.3 | 90.7 |
| Private housing | 100* | 97.9* | 94.2* | 94.9 | 5.4 | 2.0* | 99.0* |
| Terraced housing | 100* | 98.4* | 70.9 | 22.8 | 87.9 | 41.7 | 11.4 |
| Open space | 3.0* | 25.8 | 46.8 | 164.9 | 76.4 | 69.0 | 59.2 |
| Population density | 1068.4* | 970.8 | 455.1 | 328.5* | 603.0 | 610.5* | 614.8 |
| Access to shops | 289.9 | 90.6 | 7.8* | 1.1* | 7.7* | 9.2* | 9.2* |
| Distance to town centre | 0.15* | 0.68 | 1.74 | 2.71 | 1.93 | 1.92 | 1.24* |
| Distance to iron works | 0.60* | 0.80 | 1.94 | 2.95 | 1.45 | 0.88 | 1.07 |
| Total samples | 5 | 66 | 53 | 14 | 76 | 10 | 7 |

* Denotes values with a variance ≤ 0.25

south-western regions of private development differ from the previous classifications by only one sample. The one sample, added to the region at the south-west extremity (4), is now more suitably grouped.

Apart from a consideration of diagnostic variables, which may be done for all three methods (Table X), the three logical stages in the dense-space classification permit a fuller appreciation of the development and nature of each region. Three-fifths of the samples for Middlesbrough are classified at the initial, nuclei stage and all but thirty-seven samples at the basic stage, with the unassigned samples displaying a border or peripheral distribution (Fig. 6c). The latter, it will be noted, represent the 'problem areas' such as the perimeter of Albert Park, Cargo Fleet and the south-west, all of which are assigned at the final stage for purposes of complete classification. The classification sequence is shown graphically on Figure 7 where the two principal component scores for each sample are plotted at the nuclei and complete stages. At the initial phase the only regions not obviously distinctive are the two of corporation development (5 and 6). Of the seven regions these are the two most alike in character (Table Xc), although it has been argued that a division is preferable and, if the additional dimensions of the other components are considered at the complete stage, they may be envisaged on a different plane. Even in two dimensions the distribution at the complete stage clearly separates three of the regions—the northern section, in the middle of which is the town centre, and the extreme south-western area.

Conclusion

The conclusions to be drawn from the empirical example may be briefly stated. The dense-space method has produced the most satisfactory or 'natural' classification, while its tripartite

procedure enables a fuller appreciation of the constitution of each region to be made. Moreover, the computer time taken for the survey, four and a half hours, was but a quarter of that necessary for the analysis of the same data by centroid and group-average techniques. Lastly,

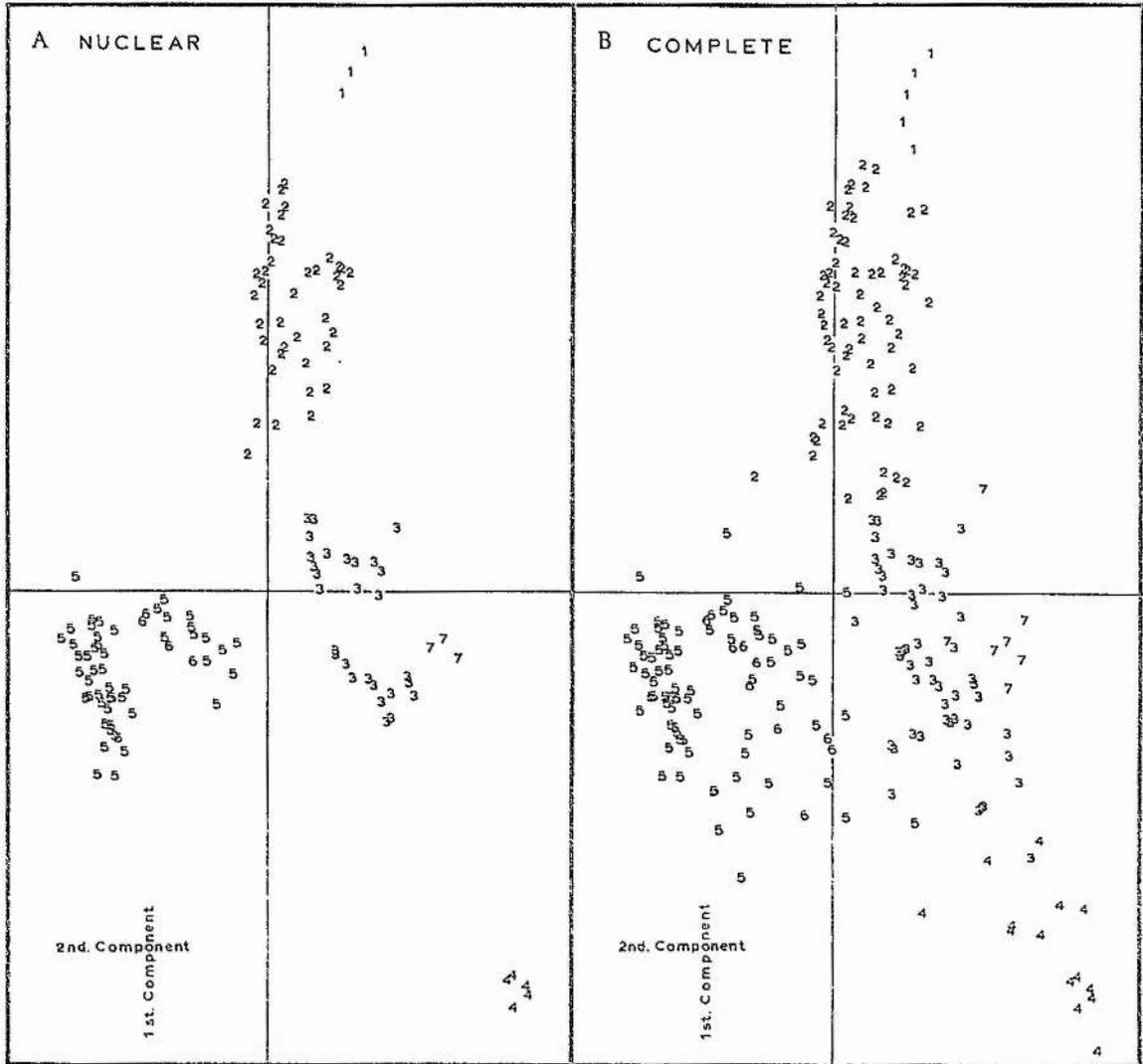


FIGURE 7—Scatter diagram of two principal component scores of Middlesbrough urban samples, dense-space method of 120 spheres, density = 3, at nuclear and complete classification levels. Numbers are identical to regions in Figure 6c.

the size of the Middlesbrough survey—231 samples—is nearing the maximum possible with the two latter techniques, given the present computer and programs, whereas the dense-space method can accommodate up to 2000 samples.

APPENDICES

1. Definitions

$$U = \begin{bmatrix} U_{11} & \dots & U_{1m} \\ \vdots & & \vdots \\ U_{n1} & & U_{nm} \end{bmatrix}$$

is the matrix of observation, where U_{ij} is the value of the j th variable for the i th sample.

\bar{U}_j = the mean of the j th variable sample distribution.

σ^2_j = the variance of the j th variable sample distribution.

2. Standardization

The matrix of standard scores on the original observations

$$U^* = \begin{bmatrix} U^*_{11} & \dots & U^*_{1m} \\ \vdots & & \vdots \\ U^*_{n1} & & U^*_{nm} \end{bmatrix}$$

is obtained using the transformation

$$U^*_{ij} = \frac{U_{ij} - \bar{U}_j}{\sigma_j}$$

3. Measure of similarity

When standard scores are used, the squared distance d^2_{ik} between two points U_i and U_k ,

$$d^2_{ik} = \sum_{j=1}^m (U^*_{ij} - U^*_{kj})^2$$

is adopted as coefficient of similarity. This is approximated by

$$d^2_{ik} = \sum_{j=1}^f (V_{ij} - V_{kj})^2$$

when the first f principal components are chosen to define the sample space, V_{ij} being the j th factor score for the i th individual.

4. Cluster diagnosis

For a sub-set of points with standardized variances $\{S^2_j\}$, the j th variable is diagnostic if

$$S^2_j \leq S^*$$

where S^* is some chosen significant variance level (usually $S^* < 1$). When a variable distribution can be assumed to approximate some theoretical distribution, statistical tests such as Fisher's ratio can be employed.

The mean μ_{oj} and variance σ^2_{oj} of variable j for a sub-set distribution in terms of the original values (before standardization) are given by

$$\begin{aligned} \mu_{oj} &= \sigma_j \times \mu_{sj} + \mu_j \\ \sigma^2_{oj} &= S^2_j \times \sigma^2_j \end{aligned}$$

where μ_{sj} , S^2_j are the mean and variance for the sub-set standardized values of variable j .

5. Dense Space Algorithm

(a) The procedure for constructing a grid of spheres:

(1) Choose the sphere radius r and first sphere centre (for example, sample point 1 or the origin).

(2) Find the distance d^2_1 of every point from the first sphere. Select the point which has maximum d^2_1 as centre for the second sphere.

(3) Compute distances d^2_2 of points from the second sphere and the distances d^2_{\min} of each point from its nearest sphere.

$$d^2_{\min}^{(1)} = \min(d^2_1(i), d^2_2(i))$$

(4) Select the point with maximum d^2_{\min} as centre for the third sphere, compute point distances d^2_3 from this sphere, update d^2_{\min} to contain the distances of each point from its nearest sphere and repeat the procedure for a fourth, fifth sphere and so on.

(5) When the maximum value d^2_{\min} is less than r^2 , the grid construction of n_s spheres encloses all the points, and the grid formation is terminated.

(b) The procedure for forming cluster nuclei:

(1) A density level k is selected, and spheres with a density which is not less than k are denoted 'dense spheres'. That is, the i th sphere is rejected if the number of values in d^2_i which are smaller than r^2 is less than k . An array DENMIN(i) of distances from the nearest dense sphere SP(i) is computed for each point (this can be compiled during the grid formation rendering the storage of all the d^2_i unnecessary).

(2) Suppose n_D dense spheres are obtained, then they are coded from 1 to n_D in an array ILINK, zero being allocated to the non-dense spheres. The squared distances d^2_{ik} between every possible pair of dense spheres (corresponding to non-zero ILINK) are computed and when, for spheres i and j , $d^2_{ij} \leq 4r^2$, then the spheres intersect. If ILINK(i) differs from ILINK(j), and $p' = \max(\text{ILINK}(i), \text{ILINK}(j))$, $p = \min(\text{ILINK}(i), \text{ILINK}(j))$, then linkage is achieved if every value of p in ILINK is changed to p' .

(3) ILINK now contains the cluster codings of each dense sphere, or zero for non-dense spheres. A classification array ICLAS is set to zero for each sample, and nuclei classifications are obtained by considering values of DENMIN which do not exceed r^2 . If, for sample i , DENMIN(i) $\leq r^2$ and $j = \text{SP}(i)$ (that is, sample point i lies inside its nearest dense sphere j), then ICLAS(i) = ILINK(j) will set the sample cluster codings in ICLAS for nuclei classifications.

Basic classification is now achieved by considering only those samples which correspond to zero ICLAS, and repeating the procedure with an extension of the distance restriction to DENMIN(i) $\leq 4r^2$, and complete classification is achieved in the same way by removing the distance restriction altogether.

6. The fusion methods

If W_i and W_k are the weights of the two synthetic points U_i and U_k which are most similar (corresponding to the smallest value of the similarity matrix D at an intermediate stage), then the distances $d^2_{(ik)r}$ of any other point U_r from the new synthetic point $U_{(ik)}$, obtained on fusion of U_i with U_k can be computed from the formulae:

(a) centroid:

$$d^2_{(ik)r} = \frac{W_i}{W_{(ik)}} d^2_{ri} + \frac{W_k}{W_{(ik)}} d^2_{rk} - \frac{W_i}{W_{(ik)}} \times \frac{W_k}{W_{(ik)}} d^2_{ik}$$

(b) group average:

$$d^2_{(ik)r} = \frac{1}{W_{(ik)}} \times (W_i d^2_{ri} + W_k d^2_{rk})$$

where $W_{(ik)} = W_i + W_k$

The fusion is effectively completed if the matrix D is modified by

$$d_{ir}^2 = d_{(ir)r}^2, \quad r = 1, n \text{ for } W_r > 0$$

$$W_i = W_i + W_k$$

and $W_k = 0$

The synthetic point $U_{(ik)}$ therefore replaces U_i , and U_k becomes redundant. A subsequent search for the next most similar pair of points should only be conducted on all elements d_{ik}^2 of D corresponding to non-zero W_i and W_k .

Sample classification is achieved by initially numbering each sample from 1 to n in an array ICLAS. When points U_i and U_k are fused to form a new synthetic individual which replaces U_i , values of ICLAS equal to ICLAS(k) should be changed to ICLAS(i). Hence at an intermediate fusion stage, there are exactly W_i elements of ICLAS equal to i , corresponding to the group of sample points which constitute the synthetic point U_i (provided that $W_i > 0$).

ACKNOWLEDGEMENTS

The authors acknowledge the co-operation of the Borough Engineer, Middlesbrough, in permitting the inspection of 1:2500 plans and thank Dr. A. J. Cole of the University of St. Andrews for permission to use his sub-routine CLERG for the computation of eigenvalues and eigenvectors. The programmes were written in Fortran II D and the data processed by the I.B.M. 1620 Model II, with disk backing store, of the University of St. Andrews. Grateful acknowledgement is also made to the Carnegie Trust for the Universities of Scotland for a grant towards the cost of illustrations.

NOTES

¹ B. J. L. BERRY, 'A method for deriving multi-factor uniform regions', *Przegląd geogr.* 33 (1961), 263-79

² For a comprehensive review of possible coefficients, see R. R. SOKAL and P. H. A. SNEATH, *Principles of numerical taxonomy* (1963); G. N. LANCE and W. T. WILLIAMS, 'A general theory of classificatory sorting strategies, 1. hierarchical systems', *Comput. J.* 9 (1967), 373-80

³ For a discussion of principal components analysis, see W. W. COOLEY and P. R. LOHNES, *Multivariate procedures for the behavioural sciences* (1964); M. G. KENDALL, *A course in multivariate analysis* (1957); D. F. MORRISON, *Multivariate statistical methods* (1967)

⁴ H. F. KAISER, 'Comments on communalities and the number of factors', paper read at the symposium 'Applications of computers to psychological problems' at the Annual Meeting of the American Psychological Association, 1959

⁵ D. F. MORRISON, *op. cit.*

⁶ B. J. L. BERRY, 'The mathematics of economic regionalization', *Proc. 4th General Meeting, Comm. Methods Econ. Regionalization, Int. Geogr. Un.*, 1965, 77-106

⁷ D. M. RAY and B. J. L. BERRY, 'Multivariate socio-economic regionalization: a pilot study in central Canada', in T. RYMES and S. OSTRY (eds.), *Regional statistical studies* (1965), 1-48

⁸ For example, G. N. LANCE and W. T. WILLIAMS, *op. cit.* 373-80; W. T. WILLIAMS, J. M. LAMBERT and G. N. LANCE, 'Multivariate methods in plant ecology, 5, similarity analyses and information analysis', *J. Ecol.* 54 (1966), 427-45

⁹ R. M. M. CRAWFORD and D. WISHART, 'A rapid multivariate method for the detection and classification of groups of ecologically related species', *J. Ecol.* 55 (1967), 505-24; *idem.*, 'A rapid classification and ordination method and its application in vegetation mapping', *J. Ecol.* 56 (1968) (in press)

¹⁰ J. H. WARD, 'Hierarchical grouping to optimize an objective function', *J. Am. statist. Ass.* 58 (1963), 236-44

¹¹ G. H. BALL, 'Data analysis in the social sciences: what about the details?', *Publ. Stanford Res. Inst., Calif.* (1965); R. C. JANCEY, 'Multidimensional group analysis', *Aust. J. Bot.* 14 (1966), 127; J. B. MACQUEEN, 'Some methods for classification and analysis of multivariate observations', *Western Management Sci. Inst., Univ. Calif., Working Pap. No. 96* (1966)

¹² M. J. SHEPHERD, 'Automatic classification methods in numerical taxonomy', Unpubl. M.Sc. thesis, Univ. of Manchester Inst. of Science and Technology (1966)

¹³ D. M. RAY and B. J. L. BERRY, *op. cit.*

¹⁴ R. R. SOKAL and C. D. MICHENER, 'A statistical method for evaluating systematic relationships', *Kans. Univ. Sci. Bull.* 38 (1958), 1409

¹⁵ B. L. J. BERRY, op. cit. (1961), 263-79

¹⁶ W. T. WILLIAMS, J. M. LAMBERT and G. N. LANCE, op. cit.

¹⁷ For example, C. A. M. KING, *An introduction to factor analysis, with a geomorphological example from Northern England* (1966); B. T. ROBSON, 'Multivariate analysis of urban areas', in *The social structure of cities* (Inst. Br. Geogr. Urban Study Group, 1966)

¹⁸ The problem of distance variables is difficult, but the inclusion of the second one, nearness to an iron or steel plant, was considered apposite in a chorological study of a town whose growth has been largely dependent on this one industry.

RÉSUMÉ—Méthodes pour la dérivation de régions uniformes et à facteurs multiples. Ce document présente une nouvelle méthode pour l'obtention de régions à facteurs multiples et uniformes. Ces méthodes de fusion, telles que le «centroid» et le «group average» qui sont basées sur des contraintes de variance minimum, pourraient bien générer des classifications artificielles pendant que la procédure du groupement par degrés et la facilité du dendrogramme qui lui correspond, ne sont pas efficaces lorsqu'il s'agit de grandes populations. La nouvelle méthode, nommée méthode du «dense space» recherche premièrement les sphères denses qui indiquent la présence d'importantes régions uniformes ou d'espace dense, et ensuite établit par dérivation des régions distinctes en unissant toutes les sphères denses qui s'intersectent. Trois niveaux de classification sont suggérés; «nuclear» «basic» et «complete». On décrit le noyau de chaque région distincte par un groupe de sphères denses qui s'entrecoupent—chaque sphère ayant $\frac{1}{2}r$ de rayon, et ayant pour propriété de couper aucune sphère dense d'une autre région distincte quelle qu'elle soit. Le groupe de points à l'intérieur des sphères denses est nommé «nuclei points» et constitue un groupement. En ce qui concerne les points situés à l'extérieur des sphères denses mais non à une distance plus grande que r du centre d'une sphère dense, on les inclue dans la classification qui concerne la sphère au niveau «basic». Les points qu'on ne peut classer au niveau «basic» sont relativement éloignés, et leur classification au niveau «complete» dans les régions qui contiennent les sphères denses les plus proches d'eux ne devrait être employée que lorsqu'une classification de chaque point est requise.

La méthode du «dense space» permet des classements plus «naturels» et exige moins de calcul et moins de mémoire de l'ordinateur. Ses avantages sur d'autres méthodes sont montrés par une révision des données du recensement (1954) des États-Unis (employés pour la première fois par B. J. L. Berry) et en se référant à une enquête urbaine sur Middlesbrough.

FIG. 1—Une répartition type hypothétique, montrant les effets d'allongement provoqués sur des groupements bien définis par l'introduction d'une variable sans rapport

FIG. 2—Erreurs de classement engendrées par la méthode du «dense space» au niveau «nucleus». Le pointillé délimite l'espace sphérique d'un groupement au niveau «basic».

FIG. 3—Dendrogramme les données du recensement (1954) des États-Unis: (A) la méthode du «centroid» (B) la méthode du «group average»

FIG. 4—Quelques traits de la structure urbaine de Middlesbrough qui montrent des endroits cités dans le texte. Notes: 1—Constructions d'avant 1918; 2—Logement municipal de l'entre-deux-guerres; 3—Logement municipal d'après-guerre; 4—Bâtiments privés de l'entre-deux-guerres; 5—Bâtiments privés d'après-guerre; 6—Principales artères commerçantes; 7—Forges et aciéries.

FIG. 5—Division de Middlesbrough en régions urbaines par l'analyse des composantes principales, 1961: (A) 1^{ère} composante, la division quartile (B) 1^{ère} et 2nde composantes

FIG. 6—Division de Middlesbrough en régions urbaines: (A) par la méthode du «centroid», les six premiers facteurs, les neuf derniers groupes; (B) par la méthode du «group average» les six premiers facteurs, les neuf derniers groupes; (C) par la méthode du «dense space», 120 sphères, densité critique = 3, aux niveaux «nuclear», «basic», et «complete»

FIG. 7—Diagramme de dispersion des deux premiers facteurs des données de Middlesbrough, méthode du «dense space», 120 sphères, densité = 3, aux niveaux de classement «nuclear» et «complete». Les chiffres sont identiques à ceux des régions représentées en Figure 6(c).

ZUSAMMENFASSUNG—Methoden für die Ableitung einheitlicher Regionen vieler Faktoren. Diese Abhandlung stellt eine neue Methode für die Ableitung einheitlicher Regionen vieler Faktoren auf. Solche Verschmelzungsmethoden, wie 'centroid' und 'group average', die auf den Einschränkungen kleinstmöglicher Varianz beruhen, könnten wohl unnatürliche Klassifizierungen zur Folge haben, während die Stufenweisegruppierungstechnik (un das Dendrogramm dazugehörig) leistungsfähig ist, wenn es sich um grosse Bevölkerungen handelt. Die neue Methode welche die 'dense space' Methode benannt ist, sucht angangs nach dichten Kugeln, welche die Gegenwart bedeutender einheitlichen Regionen oder dichten Raums anzeigen, und danach leitet sie verschiedene Regionen mittels der Verbindung aller dichten Kugeln ab, die sich schneiden. Drei Klassifizierungsniveaus sind vorgeschlagen: 'nuclear', 'basic' und 'complete'. Der Kern jeder verschiedenen Region ist durch eine Gruppe dichter Kugeln, die sich schneiden, beschrieben. Jede Kugel hat den Radius $\frac{1}{2}r$ und die Eigenschaft, dass sie keine dichte Kugel von irgendanderer verschiedenen Region schneidet. Die Gruppe Punkte innerhalb

der dichten Kugeln sind 'nuclei points' benannt, und sie bildet einen Klumpen. Diese Punkte die sich ausserhalb der dichten Kugeln befinden, sondern nicht mehr als r vom Mittelpunkt einer dichten Kugel entfernt liegen, sind in der Klassifizierung der Kugel auf dem 'basic' Niveau enthalten. Die Punkte, die auf dem 'basic' Niveau nicht klassifizierbar sind, sind verhältnismässig entfernt, und ihre Klassifizierung auf dem 'complete' Niveau in die Region, welche die nächste dichte Kugel des Punkts enthält, sollte nur benutzt sein, wenn es notwendig ist, eine Klassifizierung jedes Punkts zu haben.

Die 'dense space' Methode erzeugt natürlichere Klassifizierungen und verlangt weniger Berechnung und weniger Computergedächtnis. Ihre Vorteile vor anderen Methoden sind durch die Anwendung der zum ersten Male von B. J. L. Berry gebrauchten Volkszählungsdaten der Vereinigten Staaten von Amerika und mit Bezugnahme auf eine städtische Prüfung von Middlesbrough, bewiesen.

ABB. 1—Ein hypothetischer Stichprobenraum, der die Verlängerungswirkungen auf wohl bestimmte Klumpen zeigt, die durch die Einführung einer unverwandten Variable verursacht sind

ABB. 2—Klassifizierungsirrtümer durch die 'dense space' Methode auf dem 'nucleus' Niveau verursacht. Die punktierte Linie grenzt sphärischen Raum eines Klumpens auf dem 'basic' Niveau ab.

ABB. 3—Dendrogramm der Volkszählungsdaten der Vereinigten Staaten von Nordamerika für die Methoden (A) 'centroid', (B) 'group average'

ABB. 4—Einige Hauptpunkte der städtischen Struktur von Middlesbrough, die Orte zeigen, die im Texte erwähnt sind. Noten: 1—Gebäude vor dem Jahre 1918; 2—Stadtbehausung zwischen den Kriegen; 3—Nachkriegsstadtbehausung; 4—Privathäuser zwischen den Kriegen; 5—Nachkriegsprivathäuser; 6—Hauptgeschäftsviertel; 7—Eisenwerk und Stahlwerk

ABB. 5—Einteilung von Middlesbrough in städtische Regionen durch die Analyse von Hauptkomponenten, 1961: (A) erste Komponente, Viertelswerteinteilung, (B) erste und zweite Komponenten

ABB. 6—Einteilung von Middlesbrough in städtische Regionen: (A) die 'centroid' Methode, sechs erste Faktoren, neun letzte Gruppen; (B) die 'group average' Methode, sechs erste Faktoren, neun letzte Gruppen; (C) die 'dense space' Methode, 120 Kugeln, kritische Dichte = 3, auf den 'nuclear', 'basic' und 'complete' Niveaus

ABB. 7—Streubild der zwei ersten Faktoren der Middlesbroughdaten, die 'dense space' Methode, 120 Kugeln, dichte = 3, auf den 'nuclear' und 'complete' Klassifizierungsniveaus. Die Ziffern sind identisch zu denjenigen der Regionen in der Abbildung 6(c).

Appendix Ic: Classification of diseases.

Author's transcript of a paper read at the Proceedings of the Scottish Society of Experimental Medicine in Glasgow on 25th January, 1969.

Abstract appears in the Scottish Medical Journal (1969), v. 14, p. 96.

The Use of Cluster Analysis in the Classification of Diseases

by David Wishart

The Computing Laboratory, University of St. Andrews.

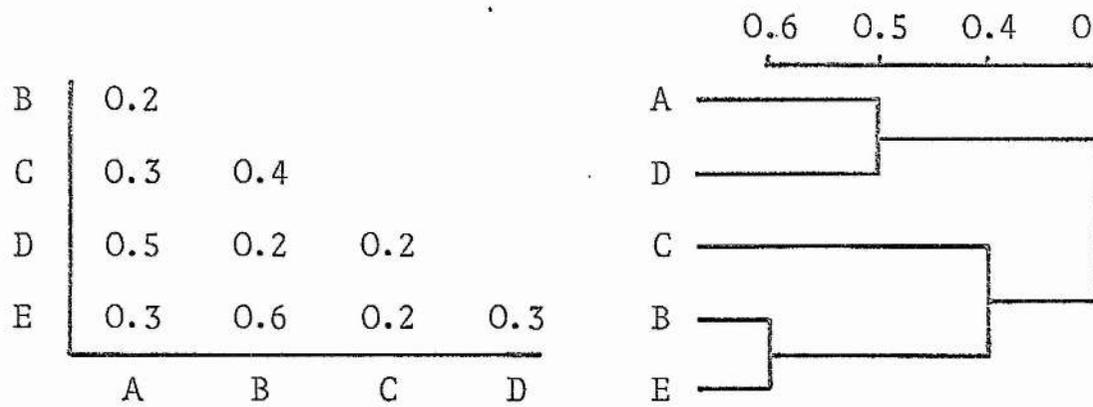
Cluster analysis can be regarded as the modern branch of numerical classification, and owes its existence to the availability of electronic computers that can economically perform the enormous computations required for detailed analyses of the inter-relationships between organisms. The general objective of a cluster analysis is to partition a population of individuals into 'meaningful' classes. During the last decade, thirty or more methods have been proposed to tackle the different interpretations of 'meaningful', and it now seems that no general theory is likely to evolve. An important point to remember is that these methods vary not only in technique, but also in results: hence one cannot state "cluster analysis failed (or succeeded)" without qualifying that statement by the name of the method which was used. For the classification of diseases, the population under study would be a group of patients who appear to suffer from closely related diseases, and the objective of the analysis is to cluster the patients into distinct disease or syndrome classes.

We begin by eliciting characters or symptoms from each patient, and then define the similarity between two patients A and B by a similarity coefficient. In figure 1, the similarity between patients A and B is obtained by dividing the number of 'matched'

| SYMPTOMS | PATIENTS | |
|------------------------|----------|---|
| | A | B |
| HOARSENESS | ✓ | — |
| CHOKING | ✓ | ✓ |
| PAIN IN GOITRE | — | ✓ |
| COUGH OR STRIDOR | ✓ | ✓ |
| INCREASE IN GLAND SIZE | — | — |

$$\begin{aligned} \text{SIMILARITY (A,B)} &= \frac{\text{NO. OF CHARACTERS IN COMMON}}{\text{TOTAL NO. OF CHARACTERS}} \\ &= \frac{3}{5} = 0.6 \end{aligned}$$

Figure 1



SIMILARITY MATRIX

NEAREST NEIGHBOUR
DENDROGRAM

Figure 2

symptoms by the total number of symptoms, hence the result $3/5$ or 0.6. In figure 2, a similarity matrix has been constructed for a population of 5 patients A-E by considering all possible distinct pairs of individuals. The first step towards forming classes of patients is to consider each patient as a single-member cluster. The clusters are now joined in successive fusion steps by combining, at each step, those two distinct clusters which are most 'similar'. This process is common to seven of the methods^{1,2} discussed here, their differences lie in the definition of the similarity between two clusters: for the 'nearest-neighbour' (single linkage) method, this similarity criterion is defined as the highest similarity coefficient between two patients belonging to different clusters; 'group average' (average linkage) uses the highest average of the similarity coefficients between patients of different clusters; Ward's error sum method² minimises the overall cluster variance; and so on. It is not essential to appreciate the significance of these inter-cluster similarity criteria - indeed some are purely intuitive concepts - however, it is important to realise that they cause different fusion sequences. The complete fusion process is conveniently represented by a dendrogram which shows the fusion steps as a hierarchy from N (population size) clusters down to 1. In the dendrogram for nearest neighbour, shown in figure 2, the steps are as follows:

1. Fuse B and E at similarity 0.6
2. Fuse A and D at similarity 0.5
3. Fuse C and (BE) at similarity 0.4 - the highest coefficient for patients (C,B)
4. Fuse (AD) and (BCE) at similarity 0.3

In order to assess the potential of these methods towards the recognition of distinct diseases, Dr. J. A. Boyle et al.³ kindly provided a data set for 67 patients, whose diagnoses of the non-toxic thyroid diseases simple goitre, Hashimoto's disease and cancer were

known. These diagnostic details and the identities of the 30 measured symptoms were deliberately undisclosed so that an objective study could be made without any knowledge of the correct disease groups - in fact, the data set consisted solely of a 67 x 30 array of coded digits. Computer programs⁴ for an IBM 1620 II were used to form groups of these patients by 8 different methods of cluster analysis, and the corresponding dendrograms were drawn by a separate program using an on-line graph plotter. The dendrogram in figure 3 for nearest neighbour clustering shows that the method completely failed to recognise any disease classes. This was due to the 'chaining effect'⁵, characteristic of this method when used with large data sets, which causes individuals to successively fuse into one cluster universe. The 'centroid' and 'median' methods also chained to the extent that their results were worthless. By contrast, the group average dendrogram in figure 4 shows good clustering for which the three disease classes are well represented in the final four groups: in fact, only 9 patients were misclassified by this method. In figure 5, the results of all 8 methods are compared by identifying the disease classes where possible and calculating a success rating as the percentage of patients correctly placed. The most successful method is Mode analysis,^{5,6} a derivative of nearest neighbour which uses a slightly different clustering process and does not provide a useful dendrogram. With this data, Mode analysis produced only two groupings: at the 3-cluster and 2-cluster levels. The 3-cluster level identified the clinical diagnostic groups with four patients misclassified, while the 2-cluster level distinguished cancer patients from the rest. In figure 6, the misclassifications and the success ratings are given for each disease class at the 3-cluster level obtained by Mode analysis.

The most important conclusion that can be drawn from this exercise is that it is insufficient, at present, to adopt only one method of cluster analysis for the classification of diseases. The

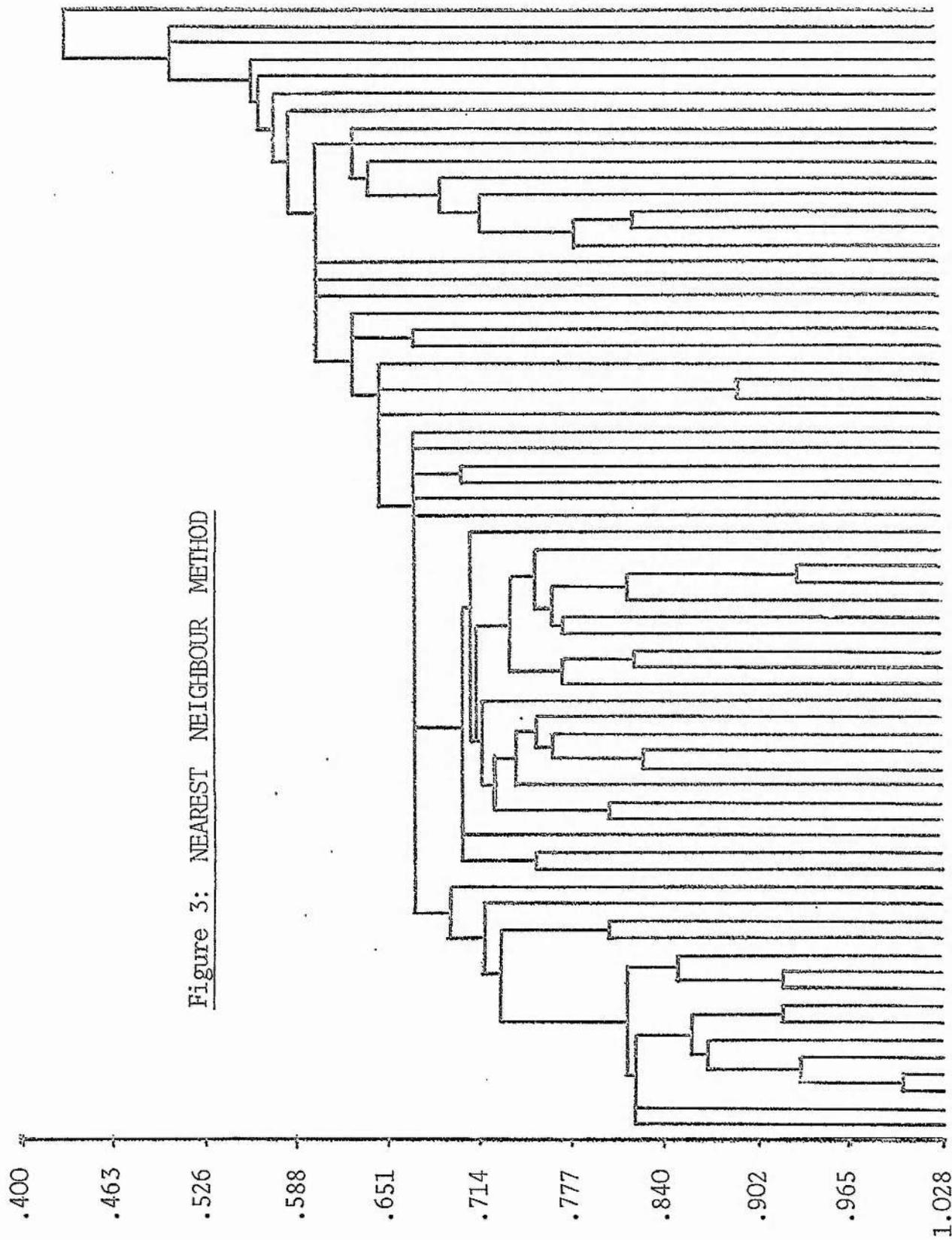


Figure 3: NEAREST NEIGHBOUR METHOD

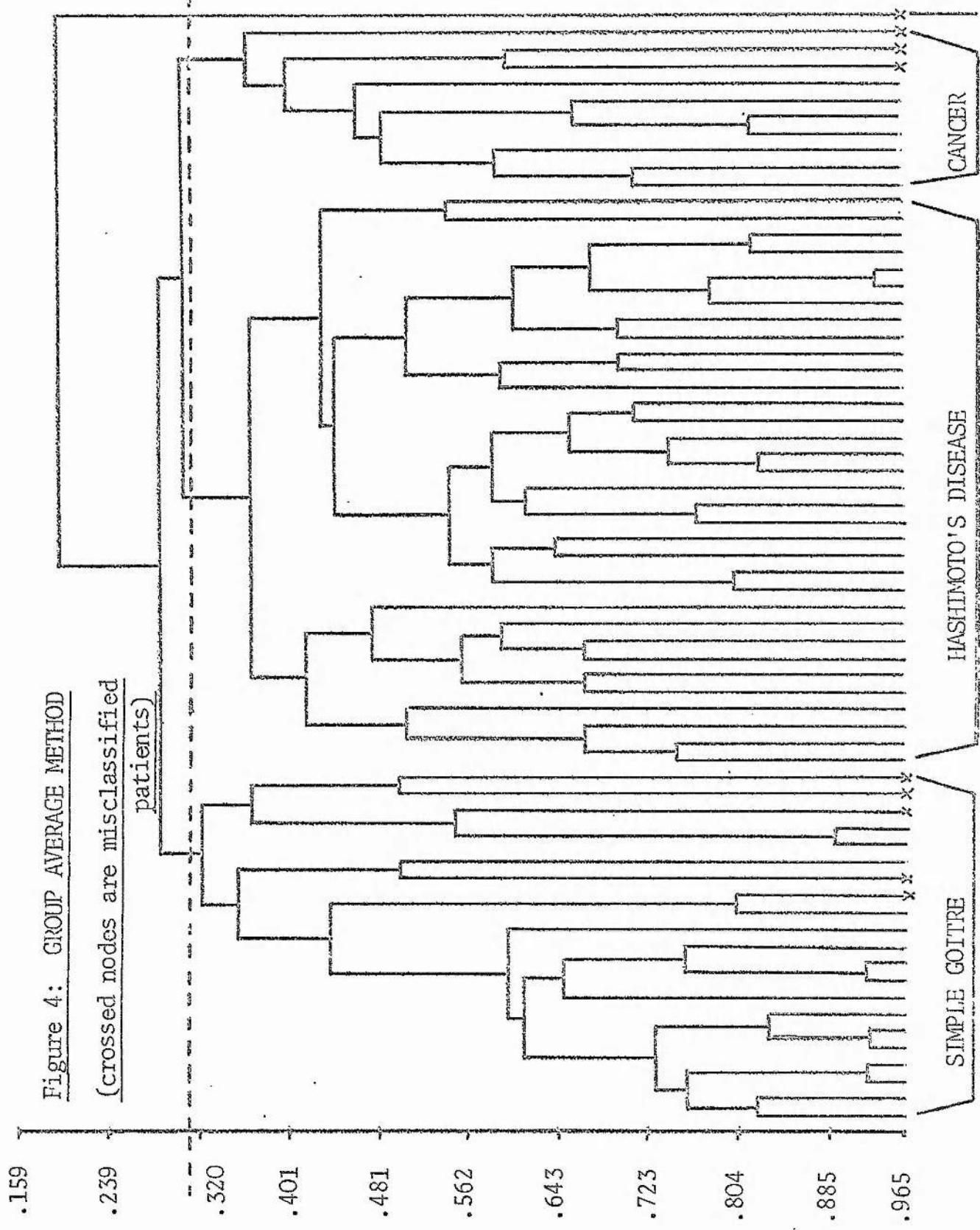


Figure 4: GROUP AVERAGE METHOD
(crossed nodes are misclassified patients)

| METHOD | SUCCESS RATING |
|--------------------|----------------|
| MODE ANALYSIS | 94% |
| NEAREST NEIGHBOUR | FAILED |
| FURTHEST NEIGHBOUR | 84% |
| GROUP AVERAGE | 87% |
| CENTROID | FAILED |
| MEDIAN | FAILED |
| WARD'S METHOD | 63% |
| FLEXIBLE METHOD | 64% |

SUCCESS RATING =

NO. OF PATIENTS IN CORRECT GROUP x 100/67

Figure 5

| CLUSTER | SIZE | CORRECT | MISC. | MISCLASSIFIED AS | | | % CORRECT |
|-------------|------|---------|-------|------------------|---|---|-----------|
| | | | | S | H | C | |
| SAMPLE | 19 | 18 | 1 | - | 1 | 0 | 94.7 |
| HASHIMOTO'S | 37 | 36 | 1 | 1 | - | 0 | 97.3 |
| CANCER | 11 | 9 | 2 | 1 | 1 | - | 81.8 |

SUMMARY OF MODE ANALYSIS RESULT

Figure 6

phrase "cluster analysis was used ..." which so often appears in applied papers is inadequate unless the actual method is specified and has previously been shown to succeed in allied studies under comparison with other methods. Furthermore, it cannot be claimed that these methods provide conclusive evidence of new disease classes or syndromes. They should merely be used as tools to aid the researcher to obtain classes of patients that must subsequently be tested for their significant independence by conventional means. Figure 7 shows the sort of procedure which must be used to extend our experience of these techniques before they can become useful. The emphasis in this flow chart lies with testing, retesting and improvement, and not as a direct process from the unsolved classification problem to the discovery of new diseases.

References

1. Lance, G. N., Williams, W. T. (1967), *Comp. J.*, 9, p. 373.
2. Ward, J. H. (1963), *J. Amer. Stat. Ass.*, 58, p. 236.
3. Boyle, J. A., Greig, W. R., Franklin, D. A., Harden, R. McG., Buchanan, W. W., McGirr, E. M., (1966), *Quart. J. Med.*, 35, p. 140.
4. Wishart, D. (1969), *Kansas Computer Contribution*, Kansas, (in press).
5. Wishart, D. (1968), Numerical Taxonomy (Proceedings of the University of St. Andrews Colloquium), Academic Press, (in press).
6. Wishart, D. (1969), *Nature*, 221(5175), p. 97.

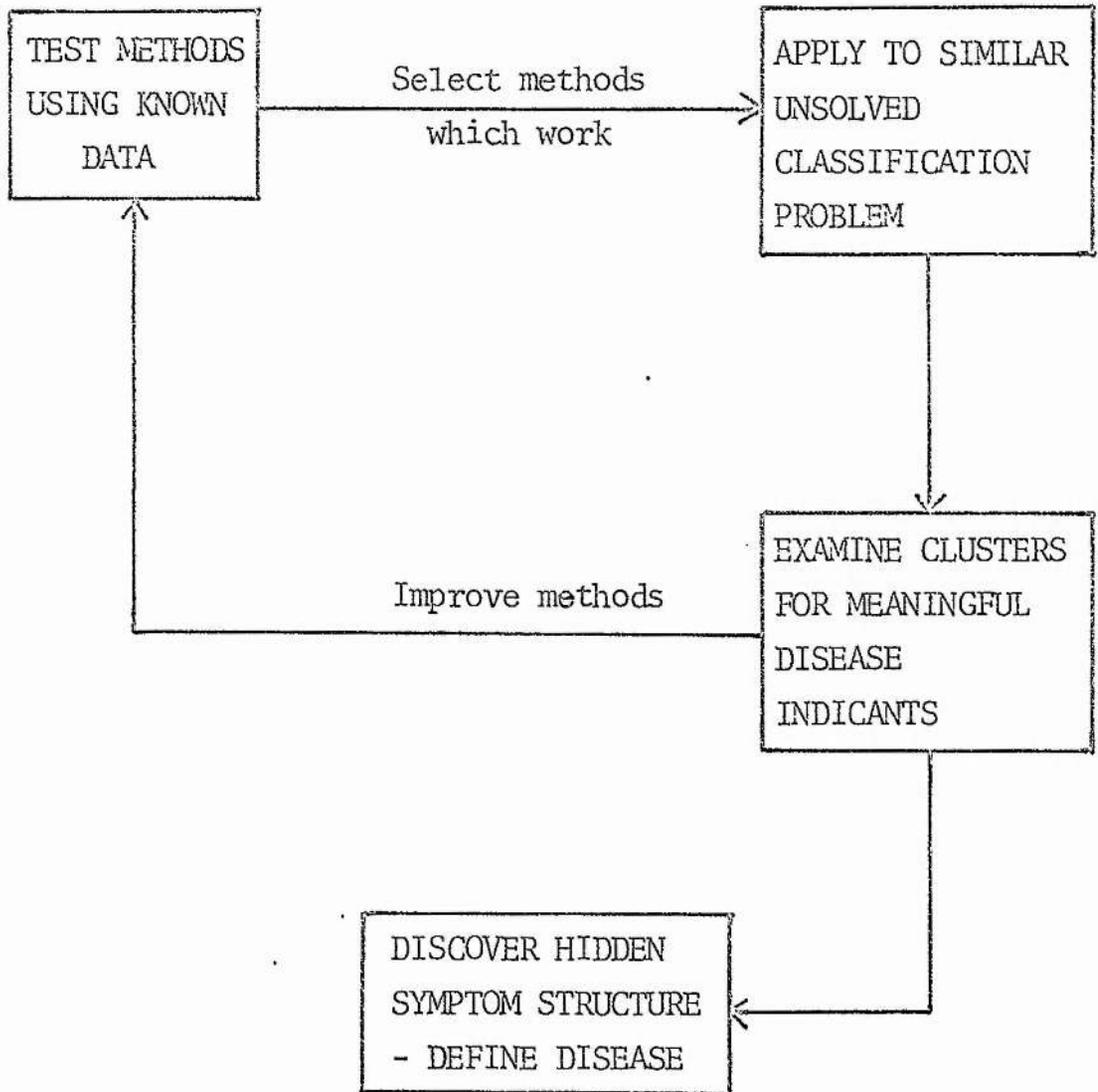


Figure 7

Appendix Id: Author's transcript of the section entitled 'COMPARISON OF CLUSTERING METHODS' which appears in: Wishart, D. (1969), 'FORTRAN II PROGRAMS FOR 8 METHODS OF CLUSTER ANALYSIS (CLUSTAN I)', Kansas Computer Contribution No. 38, Kansas, U.S.A.

Data were collected from an areal geology map, given in the Hollidaysburg-Huntingdon (Pennsylvania) Folio (Butts, 1945), which shows the distribution at land surface of 48 igneous, sedimentary and metamorphic rock units in a region of about 430 square miles. Figure 1 indicates the location of major units, and provides a rough representation of the region's topography. The map was divided gridwise into 176 4 - cm square units, and the extent of each formation was estimated visually, from the mapped coloured sections, as a score out of 10 for each unit (these data are listed immediately after the program). A key to the formation codes is given in Table 11, together with the overall percentage extent of these units as estimated.

The object of the exercise was to classify the units into groups using each of the 8 clustering methods, reconstruct a mapping of the groups to obtain a simplified representation of the areal geology, and provide a visual means of comparing the methods. Standardization of the variable values (rock unit percentages) seemed unwise in this instance, because this would tend to promote the small percentages of the thin formations to relatively high values. Consequently, the distance components of such variables would be out of proportion to their importance. It was

Table 11. - Key to formation codes used in Table 12, and measured from areal geology map (Butts, 1945)

| Rock Code | Symbol | Formation | Total Percentage Cover |
|-----------|--------|--|------------------------|
| 1 | QAL | ALLUVIUM | 5.9 |
| 2 | CA | ALLEGHENY FORMATION | 0.1 |
| 3 | CPV | PUTTSVILLE FORMATION | 0.8 |
| 4 | CMC | MAUCH CHUNK FORMATION - TOP LAYER | 5.5 |
| 5 | CTC | MAUCH CHUNK FORMATION - BOTTOM LAYER | 0.1 |
| 6 | CB | POCONO FORMATION - TOP LAYER | 6.7 |
| 7 | CPO | POCONO FORMATION - BOTTOM LAYER | 0.8 |
| 8 | DHA | HAMPSHIRE FORMATION | 4.1 |
| 9 | DCC | CHEMUNG FORMATION - CONGLOMERATE LENTILS | 0.1 |
| 10 | DSX | CHEMUNG FORMATION - SAXTON CONGLOMERATE MEMBER | 0.3 |
| 11 | DA | CHEMUNG FORMATION - ALLEGRIPPIS SANDSTONE MEMBER | 0.1 |
| 12 | DP | CHEMUNG FORMATION - PINEY RIDGE SANDSTONE MEMBER | 0.1 |
| 13 | DCH | CHEMUNG FORMATION | 9.4 |
| 14 | DB | BRALLIER SHALE | 6.6 |
| 15 | DHR | HARRELL SHALE - UPPER LAYER | 0.4 |
| 16 | DH | HAMILTON FORMATION | 3.7 |
| 17 | DM | MARCELLUS SHALE | 1.1 |
| 18 | DO | ONONDAGA FORMATION | 0.5 |
| 19 | DR | RIDGELEY SANDSTONE | 1.4 |
| 20 | DS | SHRIVER LIMESTONE | 1.2 |
| 21 | DHB | HELDERBERG LIMESTONE | 0.9 |
| 22 | STW | TONOLOWAY LIMESTONE | 3.7 |
| 23 | SWC | WILLS CREEK SHALE | 3.1 |
| 24 | SB | BLOOMSBURG REDBEDS | 1.1 |
| 25 | SMK | MCKENZIE FORMATION | 1.8 |
| 26 | SK | CLINTON FORMATION - LAYER NEAR TOP | 0.2 |
| 27 | SC | CLINTON FORMATION | 5.7 |
| 28 | SCS | CLINTON FORMATION - BOTTOM LAYER | 0.1 |
| 29 | ST | TUSCARORA QUARTZITE | 3.4 |
| 30 | OJ | JUNIATA FORMATION | 2.8 |
| 31 | OO | OSWEGO SANDSTONE | 2.3 |
| 32 | ORV | REEDSVILLE SHALE | 3.5 |
| 33 | OT | TRENTON LIMESTONE | 1.1 |
| 34 | OR | RODMAN LIMESTONE | 0.2 |
| 35 | OL | LOWVILLE LIMESTONE | 0.7 |
| 36 | OC | CARLIM LIMESTONE | 0.6 |
| 37 | OB | BELLEFONTE DOLOMITE | 6.1 |
| 38 | OA | AXEMANN LIMESTONE | 0.4 |
| 39 | ON | NITTANY DOLOMITE | 4.7 |
| 40 | OLA | LARKE DOLOMITE | 1.0 |
| 41 | OM | MINES DOLOMITE | 1.0 |
| 42 | EO | GATESBURG FORMATION - MIDDLE LAYER | 0.1 |
| 43 | EG | GATESBURG FORMATION | 3.9 |
| 44 | ES | GATESBURG FORMATION - BOTTOM LAYER | 0.6 |
| 45 | EW | WARRIOR LIMESTONE | 0.6 |
| 46 | EPH | PLEASANT HILL LIMESTONE | 0.1 |
| 47 | DEK | HARRELL SHALE - LOWER PART | 0.2 |
| 48 | EWB | WAYNESBORO FORMATION | 0.1 |

decided, therefore, to use a similarity matrix of Euclidean distances (code 1) computed from the raw data values, and the units were classified using $KL = 3$ for mode analysis, and $BETA = -.25$ for the flexible option of HIERAR.

The results of the methods nearest neighbour, median, group average and centroid exhibited the 'chaining effect'; that is, the fusion hierarchy tended to clump individual units successively into one universal group (for a discussion of chaining, see Williams, Lambert and Lance, 1966; or Wishart, 1968b). For this reason, these methods are omitted in the analysis of performance. From the remaining analyses, the 10 cluster level of fusion was chosen for the comparison of the hierarchical procedures using farthest neighbour, Ward's error sum and the Lance-Williams flexible method. Mode analysis produced 16 different groupings, of which the 5th contained the maximum of 12 clusters, and is considered here because this widest separation of the measurement units may be treated as the most general classification possible. The dominant characteristics of each cluster, together with their sizes, are set out in Table 12, and in Figures 2-5 the cluster distributions are mapped on to the original measurement grid using distinctive shading to demark each region. It is apparent immediately, from a comparison of Figures 4 and 5 with reference to Table 12, that the groupings of Ward's method and the Lance-Williams' flexible method are practically identical. In fact,

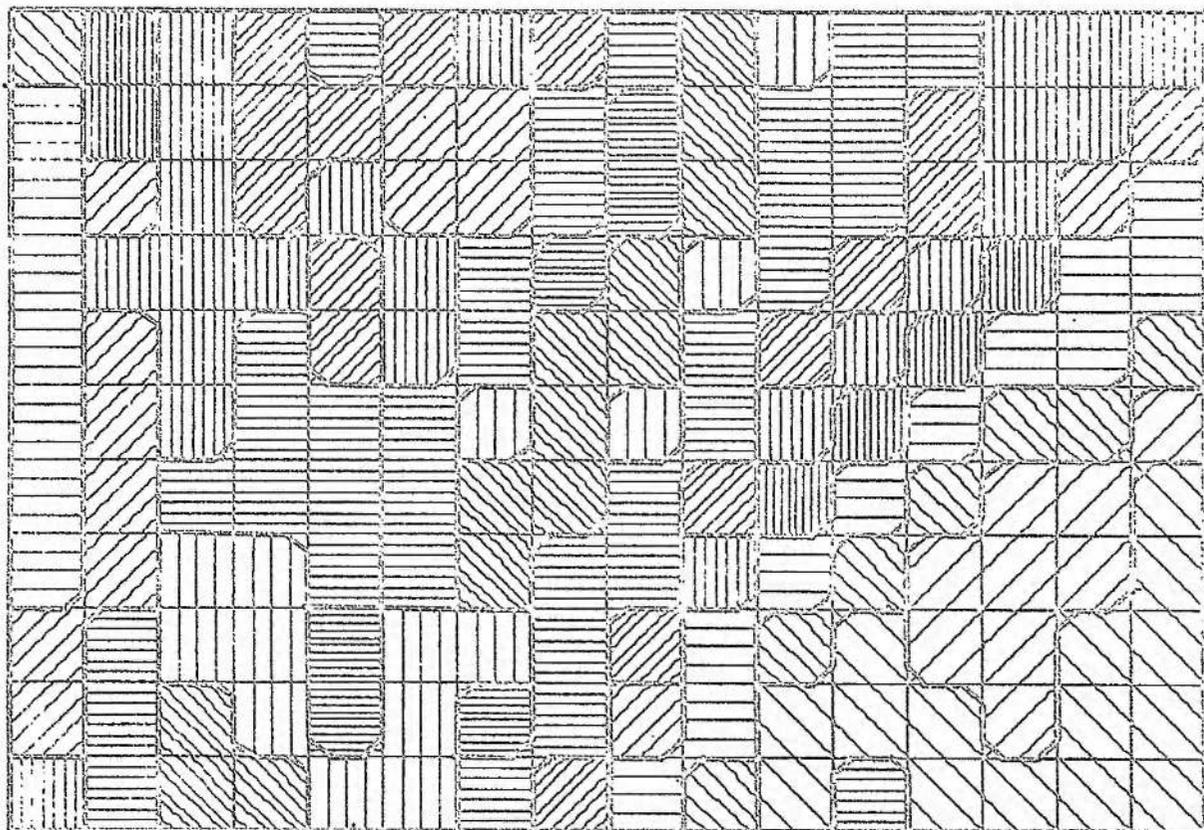


Figure 2 - Classification into 12 clusters by mode analysis.

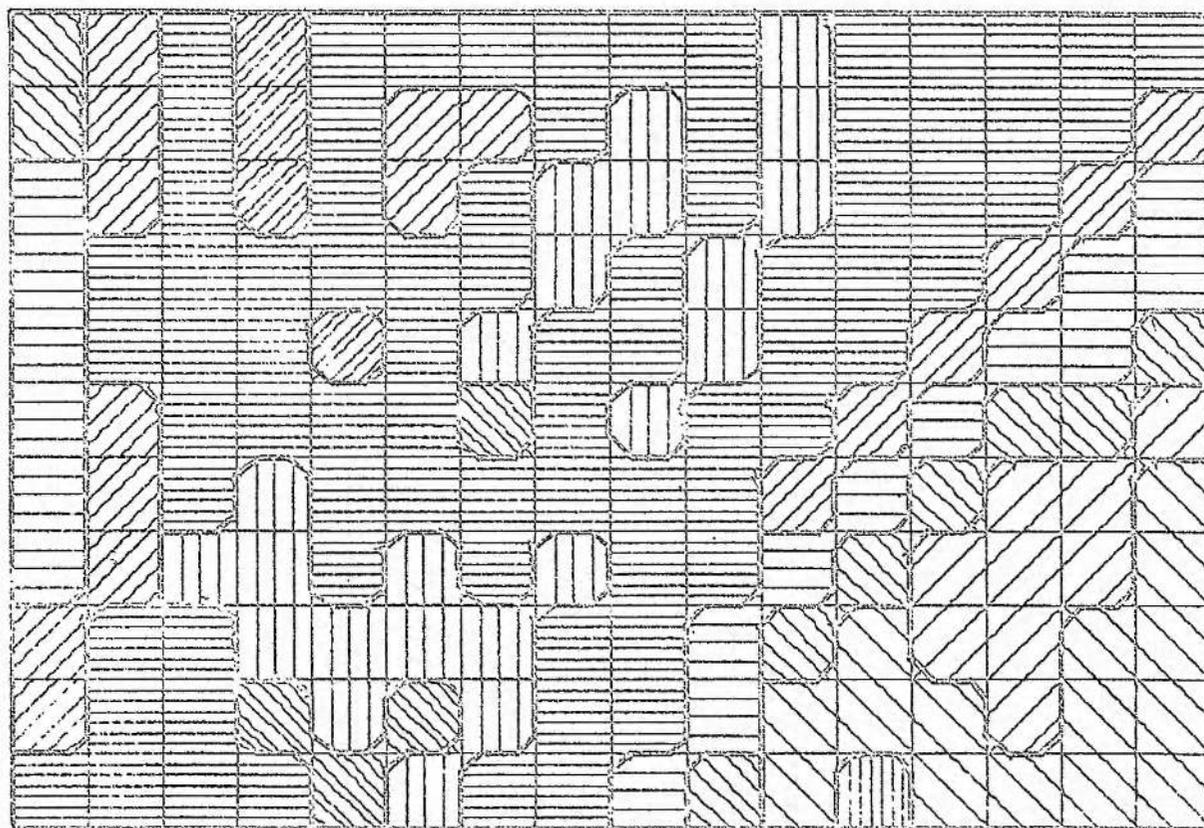


Figure 3 - Classification into 10 clusters by farthest neighbour.

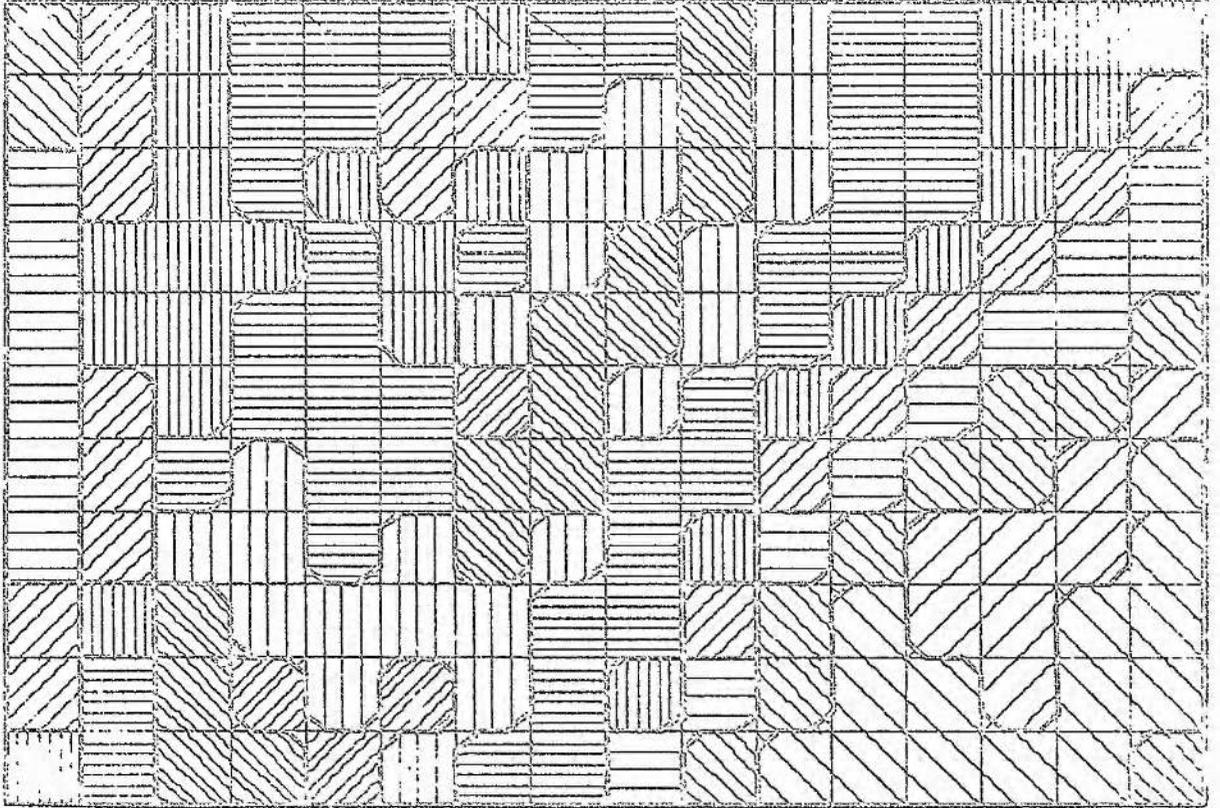


Figure 4 - Classification into 10 clusters by Ward's error sum method.

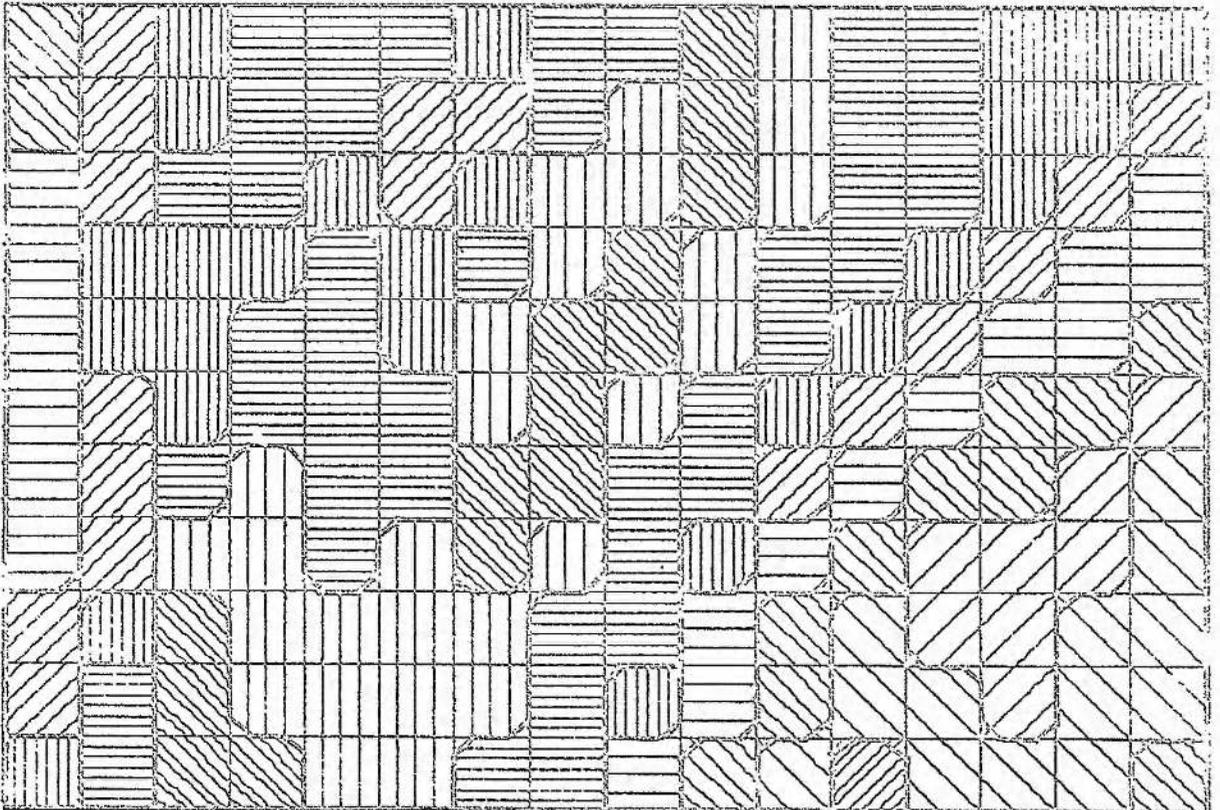


Figure 5 - Classification into 10 clusters by Jancec-Williams' flexible method.

| CLUSTER METHOD | CLUSTER CODE | CLUSTER SIZE | MAIN CONSTITUENT ROCKS (NUMBERS ARE ROCK CODES - SEE TABLE 11 - WITH PERCENTAGE COVER IN BRACKETS. ROCKS WITH LESS THAN 4% COVER ARE OMITTED) | | | | | | | | | | PERCENT OTHER | | | | |
|----------------|--------------|--------------|---|----------|----------|----------|----------|----------|----------|----------|--|--|---------------|--|--|--|------|
| MODE | 1 | 15 | 4(60.2) | 6(20.7) | 1(5.1) | 8(4.2) | | | | | | | | | | | 9.8 |
| MODE | 2 | 9 | 6(86.4) | 4(4.6) | 8(4.6) | | | | | | | | | | | | 4.4 |
| MODE | 3 | 18 | 13(72.7) | 8(9.2) | 14(8.7) | | | | | | | | | | | | 9.4 |
| MODE | 4 | 14 | 39(36.7) | 37(27.4) | 43(5.1) | 40(4.9) | 32(4.5) | | | | | | | | | | 21.4 |
| MODE | 5 | 8 | 8(56.4) | 6(12.0) | 13(11.8) | 7(10.0) | 1(7.8) | | | | | | | | | | 2.0 |
| MODE | 6 | 14 | 14(50.5) | 16(14.4) | 1(12.1) | | | | | | | | | | | | 23.2 |
| MODE | 7 | 35 | 32(15.0) | 30(12.8) | 29(12.5) | 31(10.9) | 27(9.9) | 37(8.1) | 33(4.0) | | | | | | | | 26.8 |
| MODE | 8 | 23 | 22(18.4) | 1(15.9) | 16(13.1) | 23(10.2) | 19(8.7) | 20(6.0) | 17(5.9) | 21(4.5) | | | | | | | 17.5 |
| MODE | 9 | 13 | 43(43.8) | 39(18.2) | 41(8.4) | 44(7.5) | 40(7.0) | 45(5.4) | | | | | | | | | 9.7 |
| MODE | 10 | 15 | 27(37.1) | 25(12.5) | 23(11.3) | 29(8.7) | 22(7.7) | 24(5.5) | | | | | | | | | 17.4 |
| MODE | 11 | 6 | 37(61.3) | 43(5.0) | 36(4.5) | 1(4.3) | 45(4.2) | | | | | | | | | | 20.7 |
| MODE | 12 | 6 | 14(37.2) | 13(32.5) | 16(13.0) | 1(6.7) | | | | | | | | | | | 10.6 |
| FN | 1 | 15 | SAME AS MODE 1 | | | | | | | | | | 9.8 | | | | |
| FN | 2 | 9 | SAME AS MODE 2 | | | | | | | | | | 4.4 | | | | |
| FN | 3 | 17 | 13(74.6) | 14(9.2) | 8(6.2) | 1(4.1) | | | | | | | | | | | 5.9 |
| FN | 4 | 23 | 37(39.6) | 39(13.2) | 32(10.6) | 33(5.0) | | | | | | | | | | | 31.6 |
| FN | 5 | 9 | 8(56.8) | 13(14.9) | 6(10.7) | 7(9.0) | 1(6.9) | | | | | | | | | | 1.7 |
| FN | 6 | 17 | 14(50.1) | 16(15.5) | 13(12.6) | 1(10.7) | | | | | | | | | | | 13.1 |
| FN | 7 | 77 | 33(9.4) | 22(8.0) | 43(7.9) | 29(6.8) | 25(6.7) | 1(6.6) | 30(5.8) | 16(5.3) | | | | | | | 43.5 |
| FN | 8 | 1 | 3(77.0) | 4(25.0) | | | | | | | | | | | | | 0.0 |
| FN | 9 | 4 | 39(63.8) | 37(17.0) | 38(5.5) | | | | | | | | | | | | 15.9 |
| FN | 10 | 4 | 27(65.0) | 29(11.8) | 1(6.0) | 23(4.0) | 25(4.0) | | | | | | | | | | 9.2 |
| W | 1 | 14 | 4(62.2) | 6(18.9) | 3(9.6) | 1(4.4) | | | | | | | | | | | 4.9 |
| W | 2 | 8 | 6(90.6) | 4(5.1) | | | | | | | | | | | | | 4.6 |
| W | 3 | 16 | 13(76.1) | 14(8.7) | 8(6.6) | 1(4.2) | | | | | | | | | | | 4.4 |
| W | 4 | 23 | SAME AS FN 4 | | | | | | | | | | 31.6 | | | | |
| W | 5 | 12 | 8(49.3) | 6(16.2) | 13(12.2) | 7(9.4) | 1(6.5) | 4(4.8) | | | | | | | | | 1.6 |
| W | 6 | 18 | 14(48.3) | 13(14.7) | 16(13.4) | 1(10.2) | | | | | | | | | | | 13.4 |
| W | 7 | 40 | 27(22.1) | 29(13.6) | 30(10.7) | 32(8.5) | 31(8.2) | 23(6.8) | 25(6.2) | 22(4.4) | | | | | | | 19.5 |
| W | 8 | 27 | 22(16.7) | 1(15.1) | 16(13.2) | 23(9.7) | 19(8.1) | 14(5.6) | 20(5.5) | | | | | | | | 26.1 |
| W | 9 | 14 | 43(42.5) | 39(17.9) | 41(8.4) | 40(7.2) | 44(6.9) | 45(5.0) | | | | | | | | | 12.1 |
| W | 10 | 4 | SAME AS FN 9 | | | | | | | | | | 13.9 | | | | |
| LW | 1 | 13 | 4(65.2) | 6(20.4) | 1(4.7) | 3(4.4) | | | | | | | | | | | 8.6 |
| LW | 2 | 8 | SAME AS W 2 | | | | | | | | | | 4.6 | | | | |
| LW | 3 | 17 | SAME AS FN 3 | | | | | | | | | | 5.9 | | | | |
| LW | 4 | 27 | 37(36.2) | 39(20.7) | 32(9.5) | 33(4.4) | | | | | | | | | | | 29.1 |
| LW | 5 | 12 | SAME AS W 5 | | | | | | | | | | 1.6 | | | | |
| LW | 6 | 17 | SAME AS FN 6 | | | | | | | | | | 13.1 | | | | |
| LW | 7 | 41 | 27(22.3) | 29(13.3) | 30(10.4) | 32(8.3) | 31(8.0) | 23(6.8) | 25(6.2) | 22(4.4) | | | | | | | 20.3 |
| LW | 8 | 26 | 22(17.0) | 1(14.2) | 16(13.7) | 23(9.8) | 19(8.4) | 14(5.8) | 17(5.8) | 20(5.6) | | | | | | | 19.1 |
| LW | 9 | 14 | SAME AS W 9 | | | | | | | | | | 12.1 | | | | |
| LW | 10 | 1 | SAME AS FN 8 | | | | | | | | | | 0.0 | | | | |

Table 12 - Rock distributions for each cluster derived by mode analysis (Mode 1-12), farthest neighbour (FN 1-10), Ward's error sum method (W 1-10) and Lance-Williams flexible (LW 1-10) for areal geology map data.

Type of shading



Key to shaded regions of Figure 2-5.

Cluster Number:

12 11 10 9 8 7 6 5 4 3 2 1

these two methods differ in their allocation of only 7 of the 176 measurement units. The distinction, by Ward's method, between cluster 10 and cluster 4 does not seem particularly useful despite being consistent with the results of farthest neighbour. For this reason, the flexible method seems marginally preferable. Farthest neighbour tended to oversimplify the geology of cluster 7, and this criticism is substantiated by the relatively low percentage extent attributed to the dominant units in this cluster, as shown in Table 12. By contrast, mode analysis produced rather too complex a general structure - we can expect clustered units to be reasonably contiguous by virtue of the nature of stratification and sedimentation, but although Figure 2 repeats the basic patterns common to all the maps, the number of regional discontinuities is markedly higher (this might, however, be due partly to the extra two clusters present in the MODE grouping).

For the purpose of finding a classification tool which will produce simplified patterns of areal geology from sampled data, the methods discussed here can be provisionally rated in the following preferential order:

1. Lance-Williams flexible method
2. Ward's method
3. Mode analysis
4. Farthest neighbour

The conclusion that can be drawn from this experiment could

probably have been stated at the outset, namely, that in isolating regions of uniform geology the classification method should search for groups of units which possess overall uniformity or lack of variation. Excluding mode analysis, the other three methods are of the 'minimum-variance' type and have been found to succeed in similar applications, noticeably ecology. The relative success of mode analysis, which is normally used to detect the presence (or absence) of 'natural' classes and can be considered out of context here, can be attributed to the contrasting topographic features of the area which are associated with distinctive geologic formations.

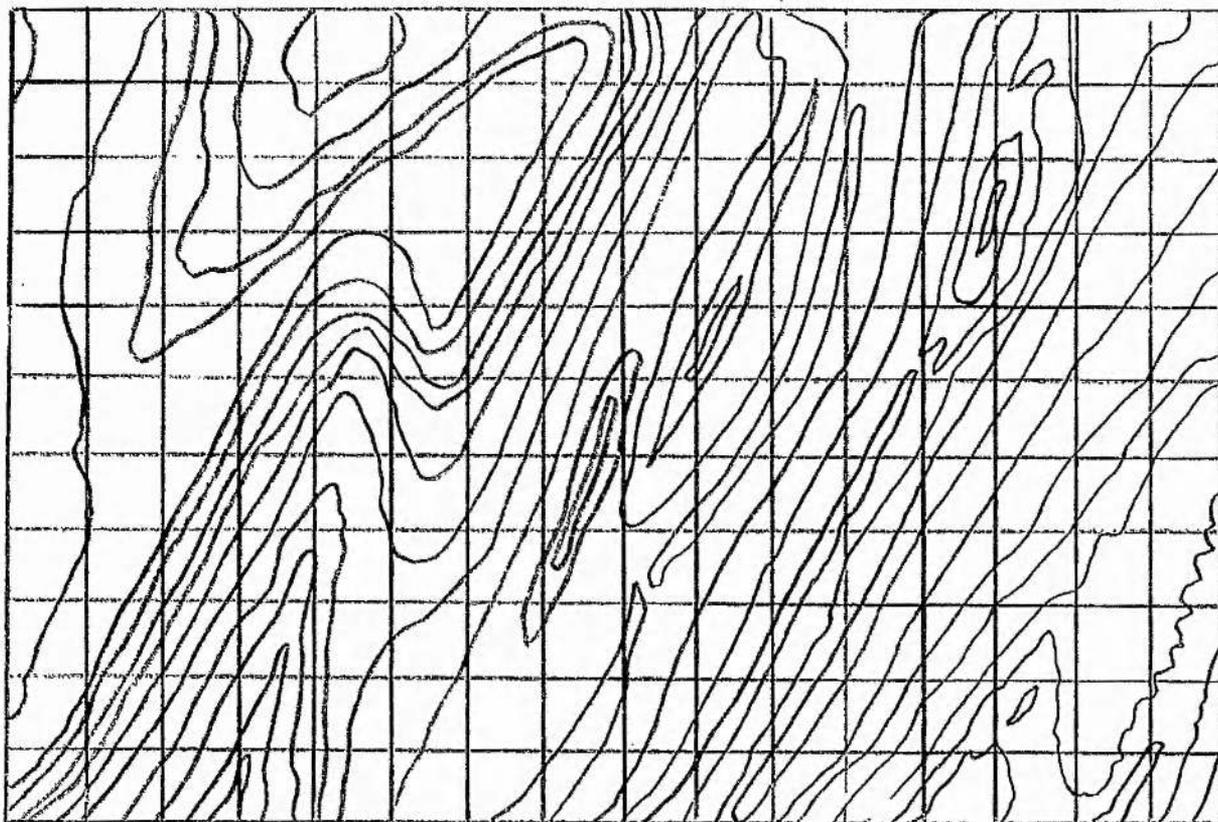


Figure 1 - Line drawing showing separation of strata on areal geology map (Butts, 1945)

Appendix Ie: Platonic prose rhythm and chronology. The following paper had, at the time of the completion of this thesis, been submitted to Computer Studies in the Humanities and Verbal Behaviour. It is reproduced here in its entirety.

A multivariate analysis of Platonic prose rhythm

by

David Wishart,
University of St. Andrews
Fife, Scotland.

Stephen V. Leach
International Computing Services Ltd.,
Kidsgrove,
Staffordshire, England.

Summary

The article suggests that Platonic prose rhythm is characterised numerically by the occurrence of 5-syllable sequences throughout passages of text. Such data, collected from 33 passages representing 10 Platonic texts, are subjected to five different multivariate procedures. It is suggested firstly, that the 10 texts can be classified 'naturally' into four groups; secondly, the variation in rhythm can be represented by a graph having two orthogonal meaningful components; and thirdly, Platonic chronology may be inferred from the ordering of the texts by two of the methods, confirming the results of previous analyses of clausulae.

It is notable that our results place Phaedrus as a relatively early work, which directly conflicts with modern theory, and it is further suggested that the 'Lysias speech' from the Phaedrus could have been a genuine 'epideiktic paignion' by Lysias, rather than Platonic parody.

Introduction

The rhythm of classical Greek verse depends on the inherent quantities of the syllables (whether they are short or long), as opposed to English verse rhythm where the stress given to a syllable is determined by its context. In both Roman and Greek literature prose-rhythm was a recognised branch of rhetoric. Cicero says that sentence endings are the most important, although the rest of the sentence is not to be neglected. Young Roman prospective legal pleaders were great imitators of Ciceronian clausulation, and instructions on prose rhythm are to be found in various classical authors. Analysis has already been carried out on Platonic clausulation (Billig, 1920; Cox and Brandwood, 1959; for an article on prose rhythm see the Oxford Classical Dictionary, 1949). The aim of this paper is to see what can be deduced from an analysis of the prose rhythm of Platonic works, looking at the rhythm throughout the text rather than at the ends of sentences.

The remarkable feature of these analyses is that they use data collected from a very basic sampling frame (the patterns of groups of syllables) which receives no subjective interference.

We find that five very different multivariate procedures produce meaningful results with these data, showing considerable agreement in the areas of classification and chronologic interpretation.

Data

We describe as a sentence any unit of words, ending with a colon, questionmark or period, and ignore commas. Each sentence is coded as a string of short and long syllables, and then all groups of five consecutive syllables are examined. There are $2^5 = 32$ possible groupings of five syllables, which are shown in table 2 starting with five shorts and ending with five longs. For a sentence of N syllables ($N \geq 5$), we start with the syllable groups 1-5, 2-6, 3-7, and so on, until all the $N-4$ groups have been considered. Every sentence throughout a sample of prose is treated in this way, and we accumulate the frequencies of occurrence for the 32 possible syllable groups. Each group frequency is then converted to a percentage of occurrence throughout the sample, on division by the total number of syllable groups observed.

For the purposes of our analysis, we have selected the 33 samples from ten texts which are given in table 1. One group comprises the Symposium and the Phaedrus. The Symposium is generally regarded as earlier than the other works we have examined; the position of the Phaedrus has in recent times been put as late (De Vries, 1969), on the grounds of its philosophical arguments and many features of language and style. From the Symposium we have

| <u>Sample Code</u> | <u>Reference</u> | <u>Number of Syllable Groups</u> |
|--------------------|--|----------------------------------|
| TIM1 | <i>Timaeus</i> 21e1 - 25d6 | 2536 |
| TIM2 | <i>Timaeus</i> 25d7 - 31b2 | 2773 |
| TIM3 | <i>Timaeus</i> 31b3 - 38b5 | 3326 |
| TIM4 | <i>Timaeus</i> 38b6 - 42e4 | 2989 |
| TIM5 | <i>Timaeus</i> 42e5 - 47e2 | 3307 |
| TIM6 | <i>Timaeus</i> 47e3 - 53c3 | 3578 |
| TIM7 | <i>Timaeus</i> 53c4 - 58c4 | 3254 |
| TIM8 | <i>Timaeus</i> 58c5 - 64a1 | 3667 |
| TIM9 | <i>Timaeus</i> 64a2 - 69a5 | 3288 |
| SOPH | <i>Sophistes</i> 242c8-244b3, 251a7-251e1, 253d5-260b1, Xen-E1 only; 260c11-261d2 | 2930 |
| PHIL | <i>Philebus</i> 14d4-15c2, (Soc. only); 15d3-16b7; 16c7-17a5; 19a1-20c1 (no 1 line interruptions); 46d6-47d3 (Soc); 58b10-58d8; 63d1-64a5 | 3355 |
| CRIT1 | <i>Critias</i> 106a1-111d8 | 3243 |
| CRIT2 | <i>Critias</i> 111e1-116c2 | 3125 |
| CRIT3 | <i>Critias</i> 116c3-121c5 | 3478 |
| LANS1 | <i>Laws IX</i> 876a9-879b5 | 2220 |
| LANS2 | <i>Laws IX</i> 879b6-882c3 | 2079 |
| LANS3 | <i>Laws V</i> 726a1-731d5 | 3099 |
| LANS4 | <i>Laws V</i> 731d6-736c4 | 3130 |
| LANS5 | <i>Laws V</i> 736c5-741a5 | 3030 |
| EP7 | <i>Seventh Epistle</i> 326b3-330c8 | 2761 |
| REP1 | <i>Republic II</i> 365a4-367e5 | 1880 |
| REP2 | <i>Republic X</i> 614b2-617d5 | 2312 |
| REP3 | <i>Republic X</i> 617d6-621d2 | 2374 |
| POL | <i>Politicus</i> 270c12-274e4 | 2661 |
| PHA1 | <i>Phaedrus</i> 244a3-248e3 | 3039 |
| PHA2 | <i>Phaedrus</i> 248e4-253c6 | 2979 |
| PHA3 | <i>Phaedrus</i> 253c7-257b6 | 2512 |
| PHA4 | <i>Phaedrus</i> 230e7-234c5 ('Lysias speech') | 2169 |
| PHA5 | <i>Phaedrus</i> 237b6-238c4; 236d8-241d1 | 2566 |
| SYMP1 | <i>Symposium</i> 189d5-193d6 (<i>Aristophanes</i>) | 2631 |
| SYMP2 | <i>Symposium</i> 180c4-185c2 (<i>Pausanias</i>) | 2990 |
| SYMP3 | <i>Symposium</i> 185e6-188e5 (<i>Eryximachus</i>) | 1397 |
| SYMP4 | <i>Symposium</i> 208c1-212a8 (<i>Diotima</i>) | 2371 |

Table 1. Origins and sizes of the 33 passages. The left column contains codes which are used to identify the passages in the text, and also in figures 1 and 3.

taken four speeches dealing with one central theme in different ways or 'styles'; from the Phaedrus we have, firstly, the 'myth' comprising three passages, and secondly, two other complementary speeches.

From the Cox and Brandwood 'late group' we have nine samples from the Timaeus, five from the Laws, three from the Republic (two from the 'Myth of Er' and one from Book II), three samples that comprise the Critias, and also one sample each from the Philebus, Politicus, and Sophistes; the last sample is from the Seventh Letter which, if genuine, would also belong to this group. All of these works apart from the Philebus, the Sophistes, and to a certain extent the Politicus, have long passages, as opposed to dialectic argument, which lend themselves readily to analysis.

We have treated these data in two separate ways. Firstly, all 33 samples have been considered individually, so that the relationships between samples from the same source can be examined. If it can be shown that, in general, such samples are very similar, then we can conclude that our sampling frame is adequate.

Secondly, we have concatenated samples from the same texts to yield a reduced population of 10 samples, each representing a separate work. Concatenation is achieved by summing the syllable group frequencies for the collection of samples taken from the same text, and then deriving percentage occurrences as before. Table 2 shows these 32 syllable group percentages for the ten texts, after concatenation.

| <u>SYLLABLE</u> <u>GROUP</u> | <u>TIM</u> | <u>CRIT</u> | <u>LAWS</u> | <u>REP</u> | <u>PHA</u> | <u>SYMP</u> | <u>SOPH</u> | <u>PHIL</u> | <u>EP7</u> | <u>POL</u> | <u>PCI</u> | <u>PCII</u> |
|---------------------------------|------------|-------------|-------------|------------|------------|-------------|-------------|-------------|------------|------------|------------|-------------|
| 1 UUUUU | 2.09 | 1.93 | 1.37 | 0.85 | 0.52 | 1.07 | 2.22 | 1.28 | 1.09 | 1.71 | 0.180 | 0.171 |
| 2 -UUUU | 2.77 | 2.79 | 2.10 | 1.64 | 1.11 | 1.68 | 2.80 | 2.32 | 2.50 | 3.01 | 0.221 | 0.137 |
| 3 U-UUU | 3.11 | 2.83 | 2.52 | 1.98 | 1.81 | 2.03 | 3.34 | 2.59 | 2.86 | 2.69 | 0.198 | 0.178 |
| 4 UU-UU | 3.13 | 3.00 | 2.00 | 2.36 | 2.18 | 2.50 | 3.07 | 2.09 | 2.43 | 2.41 | 0.048 | 0.265 |
| 5 UUU-U | 2.95 | 2.95 | 2.35 | 1.81 | 1.61 | 2.05 | 2.70 | 2.41 | 3.30 | 2.87 | 0.202 | 0.140 |
| 6 UUUU- | 2.75 | 2.83 | 2.02 | 1.60 | 1.08 | 1.73 | 2.87 | 2.30 | 2.39 | 2.90 | 0.215 | 0.144 |
| 7 --UUU | 3.45 | 3.38 | 4.00 | 2.24 | 1.74 | 2.08 | 3.28 | 3.82 | 3.88 | 4.09 | 0.243 | -0.006 |
| 8 -U-UU | 2.99 | 2.57 | 2.23 | 2.60 | 3.33 | 2.64 | 3.24 | 2.09 | 2.72 | 2.45 | -0.133 | 0.241 |
| 9 -UU-U | 3.00 | 2.88 | 1.79 | 2.98 | 3.80 | 3.16 | 2.80 | 1.88 | 2.75 | 1.99 | -0.208 | 0.154 |
| 10 -UUU- | 3.71 | 3.40 | 4.46 | 2.53 | 2.49 | 2.51 | 3.55 | 4.02 | 4.27 | 3.81 | 0.222 | -0.023 |
| 11 J---UU | 3.50 | 3.10 | 2.40 | 2.67 | 2.71 | 2.59 | 2.97 | 2.68 | 2.90 | 3.46 | 0.091 | 0.241 |
| 12 U-U-U | 2.57 | 1.84 | 1.81 | 2.79 | 3.80 | 2.71 | 2.32 | 1.91 | 2.54 | 1.47 | -0.231 | 0.070 |
| 13 U-UU- | 2.97 | 2.74 | 1.70 | 2.89 | 3.63 | 3.08 | 2.90 | 1.61 | 2.28 | 2.10 | -0.198 | 0.173 |
| 14 UU--U | 3.55 | 2.85 | 2.66 | 2.58 | 2.81 | 2.57 | 3.17 | 2.83 | 2.83 | 3.36 | 0.107 | 0.236 |
| 15 UU-U- | 2.85 | 2.77 | 2.16 | 2.50 | 3.22 | 2.63 | 2.53 | 2.24 | 3.55 | 2.45 | -0.091 | 0.163 |
| 16 UUU-- | 3.50 | 3.24 | 4.00 | 2.35 | 1.95 | 2.25 | 3.65 | 3.90 | 3.40 | 4.09 | 0.242 | -0.001 |
| 17 ----UU | 3.27 | 3.69 | 4.10 | 3.29 | 3.24 | 3.32 | 3.11 | 4.05 | 3.91 | 3.50 | 0.141 | -0.196 |
| 18 --U-U | 3.20 | 2.75 | 2.79 | 3.69 | 3.94 | 3.39 | 3.48 | 3.01 | 2.72 | 2.94 | -0.218 | 0.053 |
| 19 --UU- | 3.23 | 3.38 | 2.47 | 3.66 | 4.27 | 3.81 | 2.73 | 2.80 | 2.97 | 2.94 | -0.229 | 0.046 |
| 20 -U--U | 3.55 | 3.32 | 3.19 | 3.58 | 3.71 | 3.12 | 3.55 | 3.07 | 3.33 | 3.50 | -0.112 | 0.205 |
| 21 -U-U- | 2.82 | 2.03 | 2.35 | 3.88 | 4.52 | 3.39 | 2.66 | 2.83 | 2.54 | 1.99 | -0.241 | -0.020 |
| 22 -UU-- | 3.24 | 3.27 | 2.41 | 3.47 | 4.13 | 3.76 | 2.90 | 2.56 | 2.54 | 3.08 | -0.214 | 0.076 |
| 23 UU---- | 3.20 | 3.59 | 3.74 | 3.21 | 3.22 | 3.38 | 3.38 | 3.55 | 2.97 | 3.77 | 0.128 | -0.131 |
| 24 U-U-- | 3.05 | 2.94 | 2.74 | 3.50 | 3.98 | 3.21 | 3.00 | 3.22 | 3.51 | 2.90 | -0.209 | 0.018 |
| 25 U--U- | 3.59 | 3.04 | 3.50 | 3.64 | 3.84 | 3.10 | 3.69 | 3.31 | 3.37 | 3.39 | -0.102 | 0.091 |
| 26 U---U | 3.28 | 3.48 | 2.94 | 3.70 | 4.20 | 3.52 | 3.11 | 3.19 | 3.11 | 2.97 | -0.237 | 0.024 |
| 27 U----- | 3.00 | 3.71 | 5.02 | 4.14 | 3.44 | 4.31 | 3.31 | 4.65 | 3.77 | 4.19 | 0.062 | -0.302 |
| 28 -U--- | 3.08 | 3.57 | 4.15 | 4.64 | 4.45 | 4.46 | 3.04 | 4.44 | 3.84 | 3.55 | -0.143 | -0.250 |
| 29 --U-- | 3.57 | 3.98 | 4.51 | 4.71 | 4.27 | 4.34 | 3.58 | 4.17 | 3.69 | 4.09 | -0.106 | -0.256 |
| 30 ----U- | 3.18 | 3.62 | 4.05 | 4.63 | 4.45 | 4.55 | 3.24 | 3.96 | 3.19 | 3.74 | -0.171 | -0.207 |
| 31 -----U | 3.05 | 3.76 | 5.15 | 4.20 | 3.46 | 4.28 | 3.31 | 4.74 | 3.91 | 4.26 | 0.070 | -0.300 |
| 32 ----- | 2.75 | 4.68 | 7.32 | 5.68 | 3.07 | 6.50 | 4.47 | 6.47 | 4.93 | 4.37 | 0.060 | -0.278 |

Table 2. Percentage occurrences of the 32 5-syllable groups for the 10 book data. These figures were obtained from concatenation of the original 33 extracts into their parent sources. The two columns at the right are the eigenvectors associated with components I and II of the principal components analysis using these data.

Measures of resemblance.

For the comparison of samples of prose on the basis of the incidence of 5-syllable groups, we must first devise some measure of the resemblance between two samples, or two groups of samples. There are many suitable 'similarity coefficients' to choose from (Sokal and Sneath, 1963; Ball, 1966), but for this study, we have adopted the 'euclidean distance' coefficient because it is consistent with the analysis of within-group variance (Ward, 1963; Wishart, 1969b). The distance d_{AB}^2 between any two samples A and B is obtained by summing the squares of the differences between all 32 pairs of syllable group percentages for A and B; that is, from the formula

$$d_{AB}^2 = \frac{1}{32} \sum_{j=1}^{32} (p_{Aj} - p_{Bj})^2$$

where p_{Aj} and p_{Bj} are the percentages of occurrence for syllable group j in samples A and B, respectively. Furthermore, the relationship between any two groups of samples can also be represented with this statistic, using the averages of the percentage occurrences for each syllable group. In geometric terms, each sample is represented by a point in a sample space which consists of 32 dimensions, and d_{AB}^2 is equivalent to the squared distance between points A and B, or in the case of groups of points, the squared distance between the group centres.

In some studies it is appropriate that the data should be 'standardised' before distances are computed (Wishart, 1969b); that is, the percentages would be transformed so that each syllable group has unit standard deviation. In this instance, however, it is clear that where one syllable group exhibits greater variation than another, the difference in variation is a factor of rhythm and should therefore be taken into account. Consequently, the distances were calculated using the original percentage scores, and table 3 shows the triangular matrix of all distances between pairs of samples using the concatenated data for the ten texts.

Cluster analysis

Our objective in using cluster analysis was to form clusters of samples in such a way that each cluster represents a homogeneous stylistic block. This serves firstly to classify the works of Plato into groups whose constituent passages display uniform rhythm while a marked difference between mean cluster rhythms delineates changes in style. Secondly, if the first fusions by hierarchical cluster analysis (Ward's method) result in the grouping of the original 33 samples into their parent texts, this will confirm that the books have consistent rhythm and were, therefore, probably written continuously. In this event, the concatenation of the original 33 passages into ten books is justified. Alternatively, if any one passage from a book is not grouped with the others, this will indicate either heterogeneity within the book owing to variation in genre, or chronological breaks in continuity, or raise

questions regarding the origin of the passage in question.

Three methods of cluster analysis were used with both the 33 passages and also the ten books resulting from concatenation. Of these three methods, two are designed to optimise the within-group error sum of squares at different levels of clustering. Suppose that we obtain a number of clusters of the samples, evaluate the average syllable group percentages for each cluster and compute the squared euclidean distances from each sample to the mean of its parent group. The error sum of squares is defined as the sum of all these distances, and measures the degree of 'compactness' of the clusters. For any population of samples, there will be one or more grouping of the samples into a given number of clusters for which the error sum of squares is an absolute minimum - the first two methods used here are designed to find or approximate this optimum solution.

Ward's method (Ward, 1963; Wishart, 1969b) begins by considering each sample as a single-member cluster. In the first fusion step we group those two samples whose union minimises the resulting error sum of squares. Thereafter, at each cycle of the method we fuse two clusters whose union results in the minimum increase in the error sum of squares, and the method stops when two final clusters are fused. This process can be represented diagrammatically by means of a 'dendrogram'. Initially, each sample is located at one node on the base of the dendrogram; the first fusion is indicated

by a connecting 'stem' which joins the two nodes representing the first two samples which are grouped, and a new node for the resulting cluster is drawn from the centrepoint of this connecting stem. At subsequent fusion cycles, this procedure is repeated using single sample nodes or cluster nodes where appropriate. In order that we may recognise those fusions which correspond to a marked rise in the error sum of squares, the connecting stem is always drawn according to scale proportional to the increase in the error sum of squares caused by fusion.

The second method is an iterative relocation procedure (Jancey, 1966; Forgey, 1965) which starts with a classification of the population into clusters which may be random, or a part-optimum solution obtained from some other method. During one cycle of the method, each sample is compared in turn with the clusters, and transferred from its parent cluster to some other if the move results in a profitable decrease in the error sum of squares. The population is repeatedly scanned in this way until no relocations occur during a complete cycle. When this stage has been reached, the optimum error sum of squares solution for the given number of clusters may have been obtained.

The third method that we have used (proposed by Wishart, 1968, 1969a), is designed to search for 'natural' or 'distinct' clusters. Groups of points in the sample space are separated if there is a distinct discontinuity of the density of points between the groups. No constraint is imposed on cluster variance - hence,

the resulting clusters are not required to exhibit internal compactness. The classification criterion is derived from a probabilistic model for which classes correspond to disjoint density surfaces within the 32-dimensional sample space; in a sense, the classes must be 'distinct' or non-overlapping to be distinguished, but they can have any shape. One characteristic of the method is that it does not derive a complete hierarchy of nested groupings like Ward's method. Instead, a reduced number of groupings is obtained, of which the first usually contains the largest number of clusters, and the others conform to higher levels of classification. It has been suggested (Wishart, 1969a), that that grouping which contains the largest number of clusters corresponds to the widest 'natural' classification of the data that is possible. Theoretically, the method is radically different from the previous two, because it permits clusters to have any shape or variance in the sample space.

Ward's method

Figure 1 shows the dendrogram for Ward's method used with the original 33 passages. It is immediately apparent that, in general, the passages are initially grouped according to their sources. We observe that at the 5 cluster cut-off point the groups roughly represent Timaeus, Critias, Laws, Symposium and Phaedrus; Sophistes and Republic Book II are located with the Timaeus Politicus and the Seventh Letter are located with Critias; Philebus

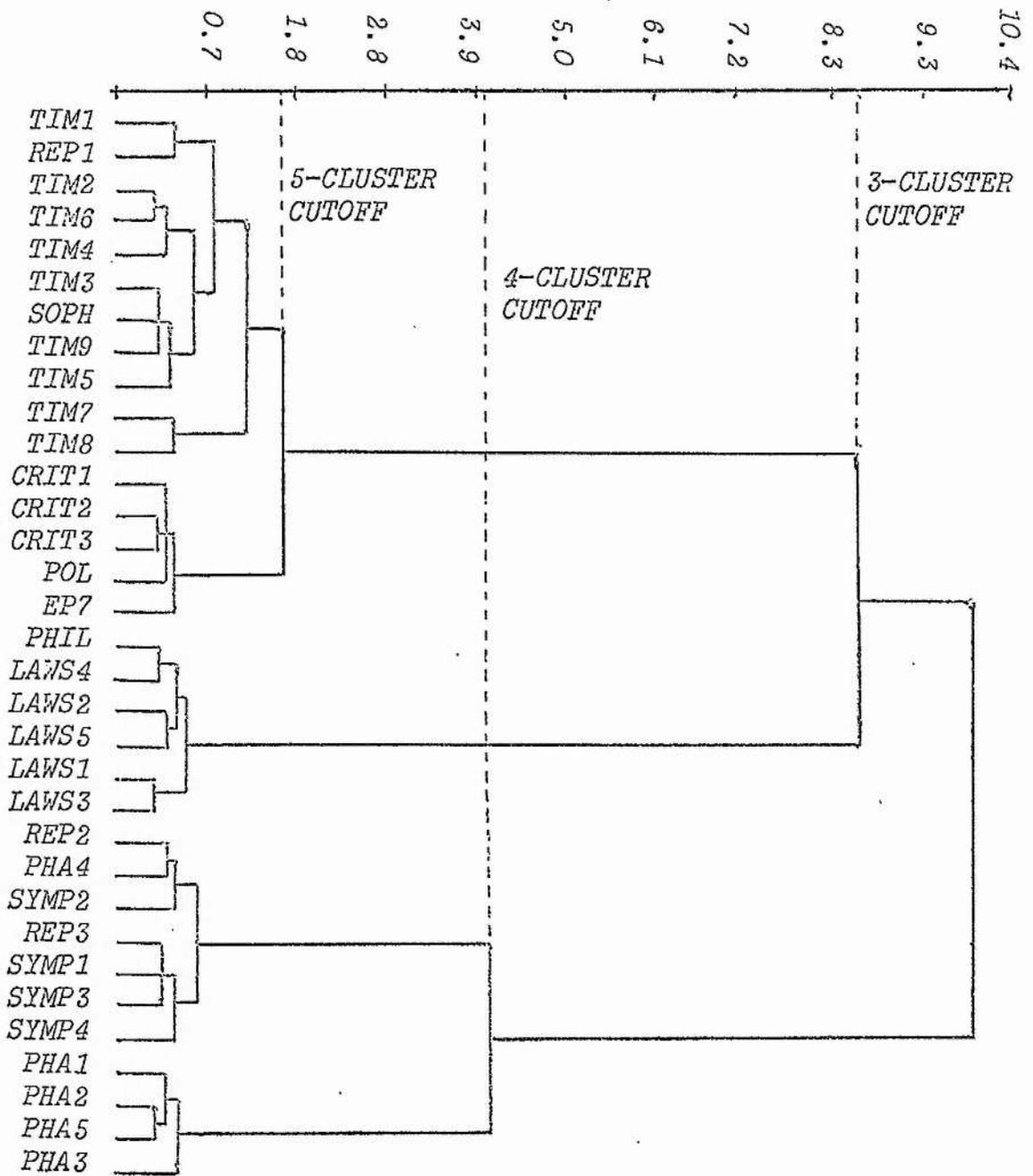


Figure 1. Dendrogram for the hierarchical classification by Ward's method using the 33 extracts in Table 1. The 4-cluster cutoff probably indicates the best grouping of the passages.

is located with the Laws; and the fourth passage from Phaedrus is classified with the Symposium together with the two samples from Book X of the Republic. At the 4 cluster level Critias, Politicus and the Seventh Letter are grouped with the Timaeus-Sophistes class. It is interesting to note that the error sum of squares shows marked rises from the 4 to 3 cluster, and 3 to 2 cluster levels. This would suggest, according to Ward's guide to the use of his method, that the levels which best minimise the within-group variation are those comprising 4 and 3 clusters.

Of interest is the grouping together of the four passages from the Symposium; although these are four speakers dealing with a common theme (the praise of Eros) in different styles, the result indicates uniform prose rhythm despite Plato's conscious effort to vary style. The close grouping of passages from the Laws suggests homogeneity within this work, as also shown by the Cox-Brandwood analysis of clausulae. Similarly, the Critias, the Timaeus and four passages from the Phaedrus all cohere nicely into their parent sources. That the Sophistes sample is located with the Timaeus group suggests that these works may have been written at about the same time.

The first unexpected result of the analysis is the classification of Republic Book II with the Timaeus. There are three possible explanations, all of which are rather unsatisfactory:

(a) The passage is too short, containing only 1880 syllable

groups. However, one would have thought that this was a sufficiently large sample.

(b) It is of a different genre to the other passages from the Republic, being from a speech by Adeimantus, whereas the other two are extracts from a myth.

(c) Book II may have been written some considerable time before Book X, and the difference in rhythm is associated with Plato's chronological development - further evidence to be submitted in the next section.

(d) Lesky (1966) dates the completion of the Republic to 347 B.C., after the Symposium and Phaedo, and before the Theaetetus, on the grounds of content. Certainly there has been for a long time a school of thought which treats the 1st. Book of the Republic as earlier, and to some extent separate, from the rest of the work on the grounds of its content, and no proof positive has yet been offered on the dating of the Republic as a whole or the time it took to write.

The second unexpected result is the placement of the fourth passage from the Phaedrus. This speech may well have been written by Plato, yet he does portray Phaedrus as saying that the speech is written by Lysias. On the other hand, deliberate parody may affect the results, although we optimistically think our data and methods of analysis probe deeper. We would have further reason to support the second quasi-explanation for Republic Book II above,

if this passage had been classified with the fifth Phaedrus sample, being the answering speech. However, the fifth Phaedrus sample is grouped with the first three passages, suggesting that the 'Lysias speech' may have been a genuine 'epideiktic paignion' by Lysias (evidence to be supported later).

In figure 2, the results are shown for Ward's method using the 10 samples resulting from concatenation of the original 33 passages. It is interesting to note that the order of books at the base of the dendrogram corresponds exactly to the inferred chronological order for those works which were examined by Cox and Brandwood. The Seventh Letter (not considered by Cox and Brandwood) has been placed between Politicus and Philebus, which is more satisfactory from the historical point of view than the position allocated by Levison et al (1968), using the Cox-Brandwood method. Another point of interest is that the fusion sequence is basically similar to that obtained for the 33 passage analysis. One difference is the large increase in the error sum of squares beyond the 4 cluster level. This confirms that the 4 cluster level is probably the best stylistic classification.

Iterative Relocation

Two starting classifications were used in each analysis of the within-group error sum of squares by the method of iterative relocation. One was the associated level derived by Ward's method, and therefore represents a part-optimum initial solution; the other

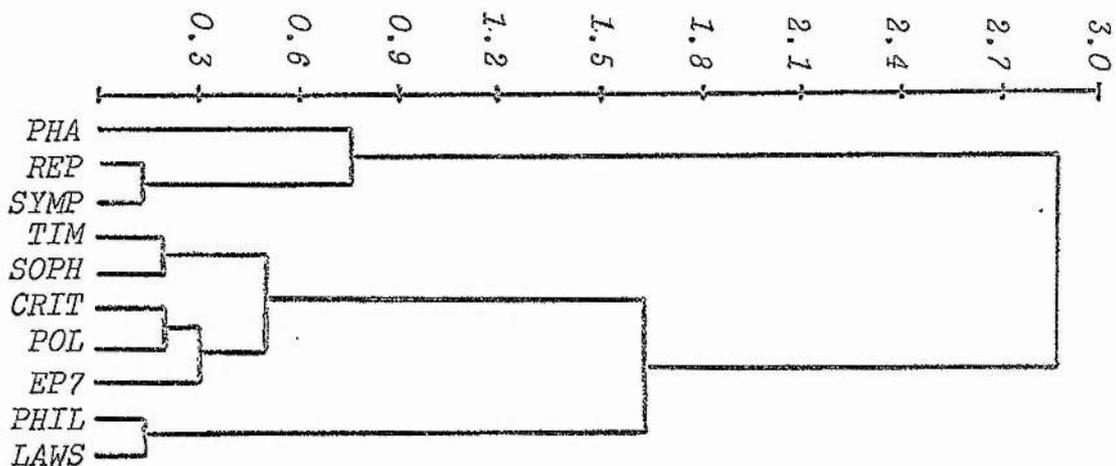


Figure 2. Dendrogram for the hierarchical classification by Ward's method of the 10 books given in table 2.

was as bad an initial grouping as we could devise. For example, to optimise the grouping into 4 clusters we chose firstly, the 4-cluster result obtained with Ward's method, and secondly, four clusters of equal size each containing roughly one-fourth share of the members of the previous groups. In every analysis, both initial classifications produced the same final result, although of course the bad grouping required more relocations than the part-optimum solution before reaching stability.

At the 5 cluster level, the only changes in the result for Ward's method with the 33 samples were that Republic Book II was transferred to the Phaedrus group, and the first Timaeus sample was relocated with the Critias-Politicus-Seventh Letter cluster. At the 4 cluster level, the only transference was that of Republic Book II. This further supports explanation (c) of the previous section that Republic Book II is an early writing, since our evidence suggests that the Phaedrus is the earliest of the Platonic works that we have considered.

The re-allocation of the Timaeus sample is probably because this passage is introductory to the main part of the Timaeus, and therefore differs slightly from the rhythm of the rest of the work. Also, the Timaeus group is fused with the Critias at the 4 cluster level, showing that the relocation is not a dramatic one.

With the concatenated 10-book data, the results for iterative relocation and Ward's method coincide at both the 5 and 4 cluster levels indicating that the absolute minimum within-group error sum

of squares has probably been achieved.

The probabilistic method

The computer program for Wishart's method produced only three classifications of the 33 passage data: at the 4, 3 and 2 cluster levels respectively. Following Wishart's recommendation that the grouping with the greatest number of clusters can be regarded as the widest level of classification which is 'natural' or 'distinct', we chose the 4 cluster level for the comparison of results. In fact, these clusters are identical with the result obtained with Ward's method. It is interesting to note that Wishart's method failed to divide the large Timaeus-Critias-Sophistes group into two subsets corresponding to the 5 cluster level of Ward's method. This result illustrates the difference between the methods: Ward's method divides the group into two sections at the 5 cluster level so that the within-group variation is minimised; Wishart's method fails to find subclusters because the group exhibits a uniformly dense structure (see figure 3 and the next section). It could be argued that any partition of this group to form two 'tight' subclusters would be quite arbitrary, and these considerations further substantiate the inference that the Platonic works we have examined can be grouped 'naturally' into 4 rhythmic classes.

Principal Components Analysis

We use principal components analysis in this study to find that 2-dimensional scatter diagram which best describes the

relationships between our samples. As previously explained, each sample can be represented by a point in a sample space of 32 dimensions (one for each 5-syllable group percentage), and the distances between points correspond to the similarities between samples. However, because the population of points is in 32 dimensions, we cannot easily visualise (let alone describe) the general characteristics or trends of the distribution. Principal components analysis is designed to find a smaller frame of reference (reduced set of dimensions) which accurately describes the swarm of points. In this instance, we are looking for that 2-dimensional frame of reference (graph) which best approximates the relationships in 32-space between our 33 samples. In effect, the method finds the plane through the 32 dimensions which is the best least-squares fit to the distribution: the points are then projected on to this plane, and its goodness-of-fit is assessed by the percentage of the total variance which is accounted for by the plane. The higher this percentage, the closer the resulting graph represents the distribution in 32-space.

Figures 3 and 4 show the graphs produced by principal components analysis using the 5-syllable group percentage data for both populations. Figure 3 accounts for 72.6% of the overall variation within the 33 passages; figure 4 accounts for a massive 80.2% of the variation within the data for the 10 books. In Table 2, the eigenvectors are shown for Principal Components I and II of the 10

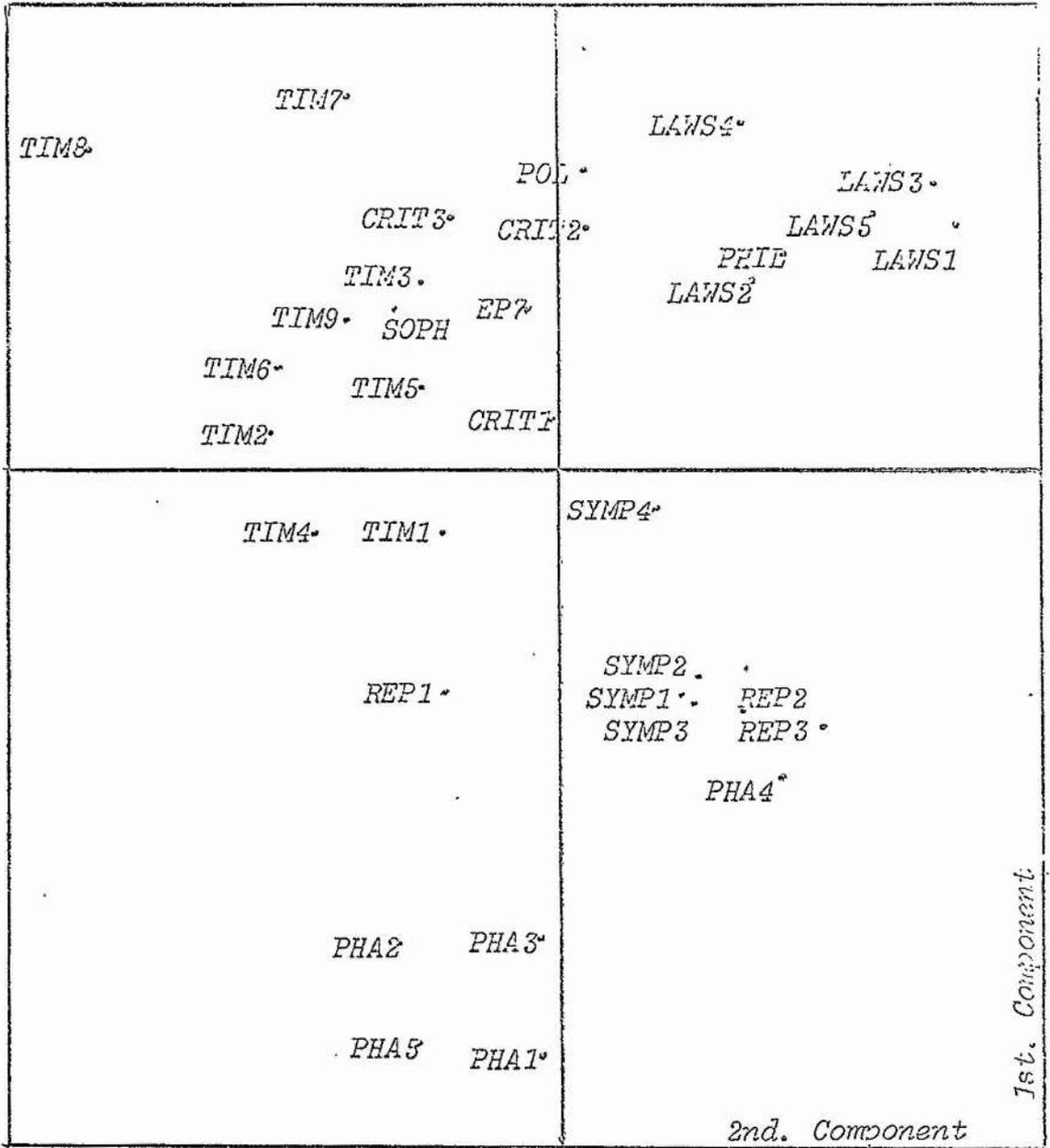


Figure 3. Scatter diagram for the 33 passage data obtained by plotting principal component I against principal component II. 72.6% of the overall variation is contained in this diagram.

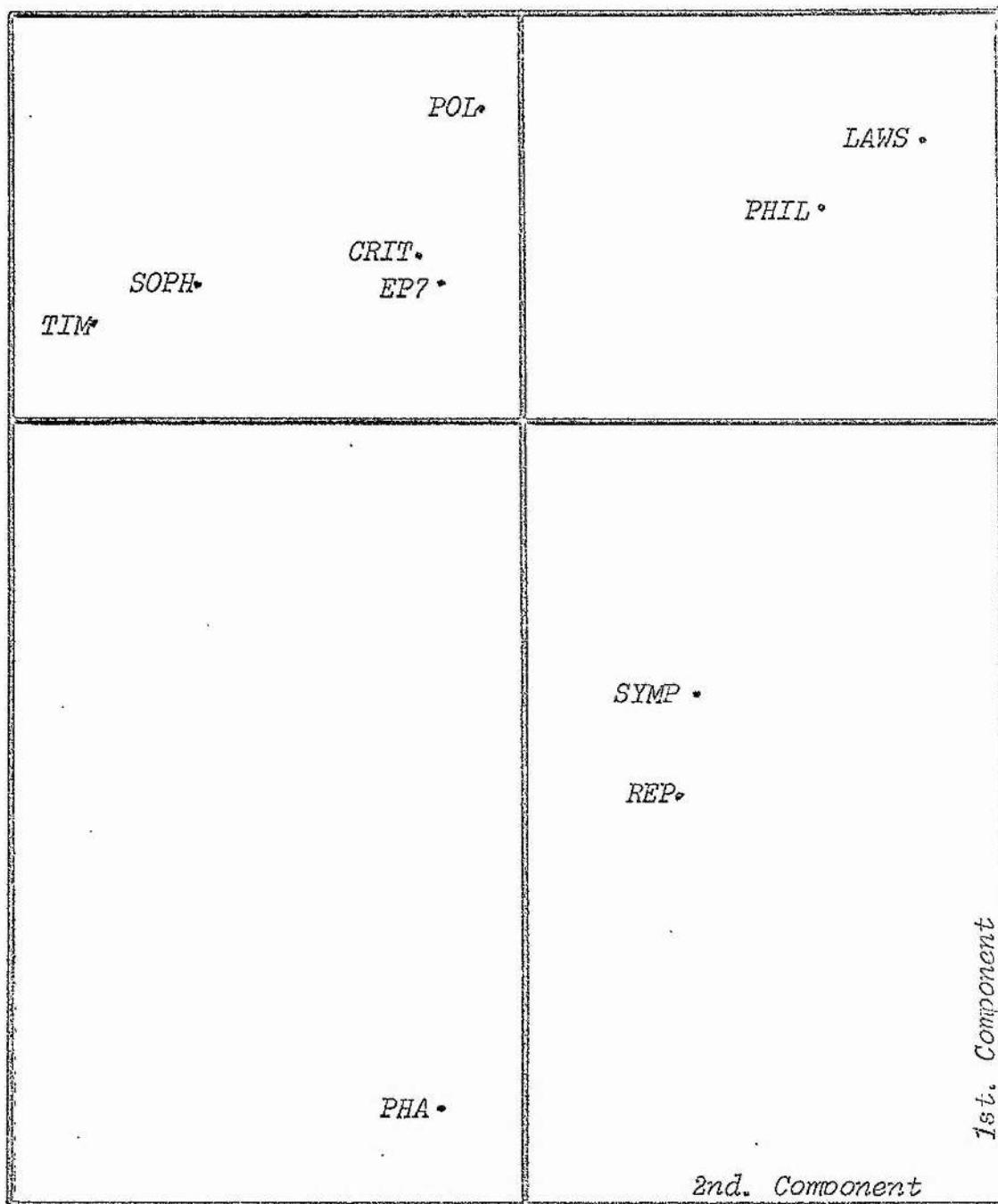


Figure 4. Scatter diagram of component I versus component II for the ten book data obtained by concatenating the 33 original samples. 80.2% of the variation is explained by this diagram.

sample data. It is immediately apparent that the distribution of points, excepting those of the Phaedrus, are roughly in chronological order along Component I. Both diagrams therefore suggest that the Phaedrus is an early writing, although it should be noted that the point PHA on figure 4 represents all five passages, including the 'Lysias speech' (PHA₄). It could be argued that the 'Lysias speech' should not have been concatenated with the others on the grounds that there is some doubt as to its actual origin. However, figure 3 shows the relative position of PHA₄ to be 'later' than the other passages, so that its exclusion would make the Phaedrus appear even earlier on figure 4. Component II, on the other hand, is associated with the ratio of the numbers of long and short syllables within a sample. This is most clearly demonstrated in Table 2, where the eigenvector for Component II is seen to have large values corresponding to the short and long syllable groups. These large values are essentially correlations between Component II and the long and short syllable groups. It is evident from Table 2 that the short syllable groups are positively correlated, while the long syllable groups are negatively correlated with the scores on Component II. However, it should be noted that the data must first be 'standardised' before principal components analysis is evaluated, and for the reasons previously given for not standardising, we expect that a certain amount of the rhythm variation will have been lost. However, the results are still more or less those

that we expected. Of interest is the very broad variation on Component II for those passages written in Plato's last period. From the graph it appears that there was not such a wide rhythm variation in the earlier period of Plato's writing, which suggests his growing awareness of rhythmical patterns towards the end of his life.

Figure 3 clearly shows the separation of PHA4 from the other passages from the Phaedrus, which are very closely clustered. This supports the previous suggestion that the apparent 'misclassification' of PHA4 by cluster analysis might be because the 'Lysias speech' was written by Lysias, rather than Plato. The unexpected clustering result for REPl is also explained on the diagram by its distance from the Republic-Symposium group.

The principal components diagram can also be used to represent graphically the results of cluster analysis. In figure 5, the 5-cluster solution of Ward's method is shown by plotting cluster codes, instead of identification labels, for each of the 33 passages. Each cluster is also represented by a circle whose radius is proportional to the joint variance of the cluster distribution; small circles therefore represent compact clusters. Figure 5 also indicates the two reassignments that occurred when iterative relocation was used with the 5-cluster solution of Ward's method: REPl, assigned to cluster 1 by Ward's method, is moved to cluster 5; TIM1, also assigned to cluster 1 by Ward's

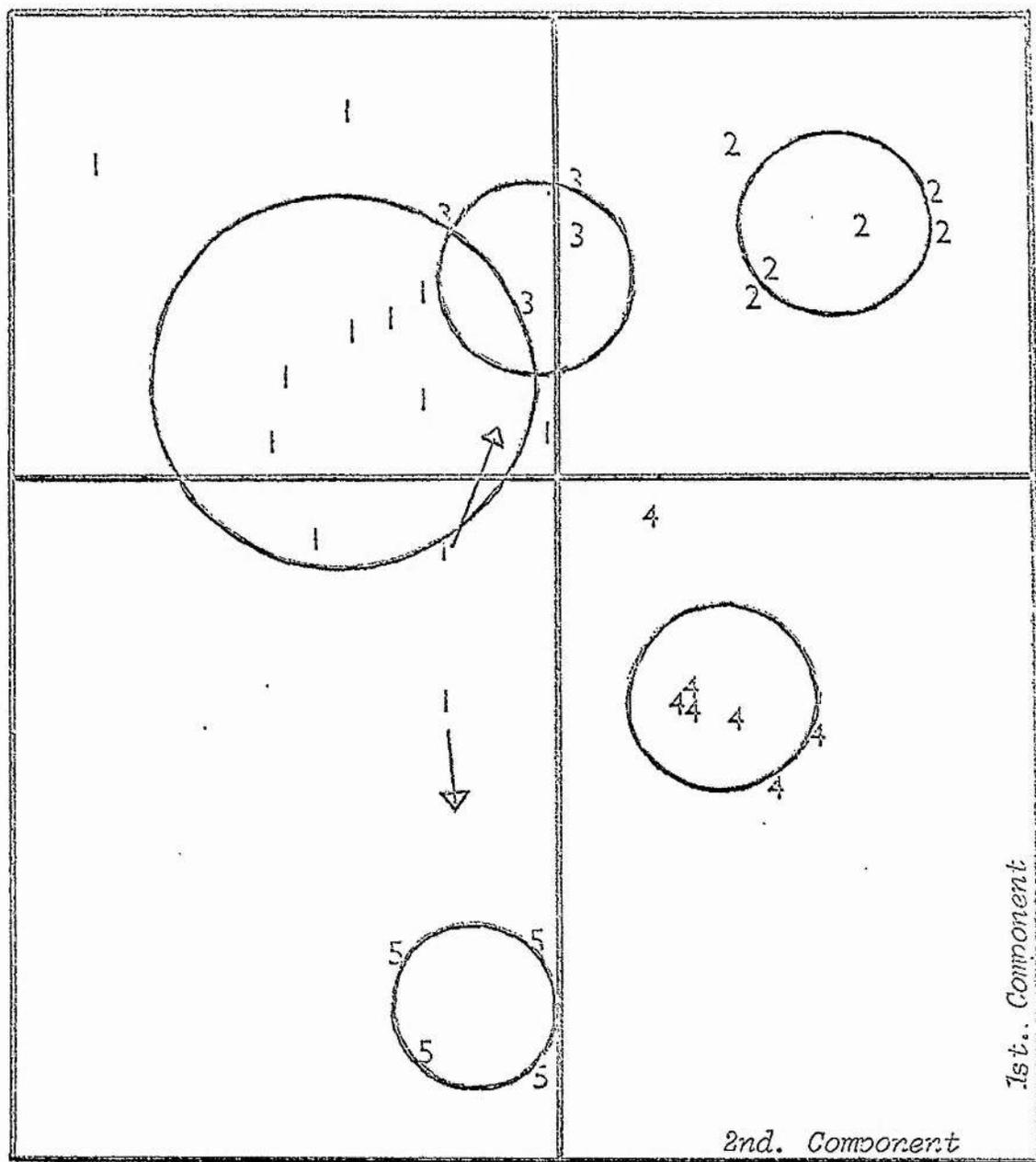


Figure 5. Classification of the 33 passages by Ward's method into 5 clusters, superimposed on the principal components diagram of figure 3. Arrows denote the reassignments of REPI with cluster 5 and TIM1 with cluster 3 by iterative relocation.

method, is moved to cluster 3.

Figure 6 shows the results for both Ward's and Wishart's methods at the 4-cluster level: the solutions are identical. Also indicated on figure 6 is the removal of REPl from cluster 1 and its reassignment with cluster 4 by iterative relocation. The relocation of REPl is the only difference between the three methods, and figure 3 clearly shows the discontinuities between the clusters at the 4-cluster level, suggesting that this grouping constitutes a natural classification of the data. Indeed, it is clear that the resulting clusters are not only separated, but also reasonably compact - hence the concurrence of the cluster analysis methods of Ward and Wishart.

Multidimensional scaling

The fifth method of analysis used here is a technique developed by J.B. Kruskal called multidimensional scaling (Kruskal, 1964; Kruskal and Hart, 1966). This method begins with the distance matrix (Table 3) which, in this instance, contains distances measured in 32 dimensional space (one dimension for each 5-syllable group). The objective of multidimensional scaling is to rearrange the points in a smaller number of dimensions so that the new distances 'bear a sensible relationship to the original distances'. For our purposes, we require a descriptive solution, and have therefore chosen to reduce the distribution to 2 dimensions and 1 dimension so that the results can be displayed

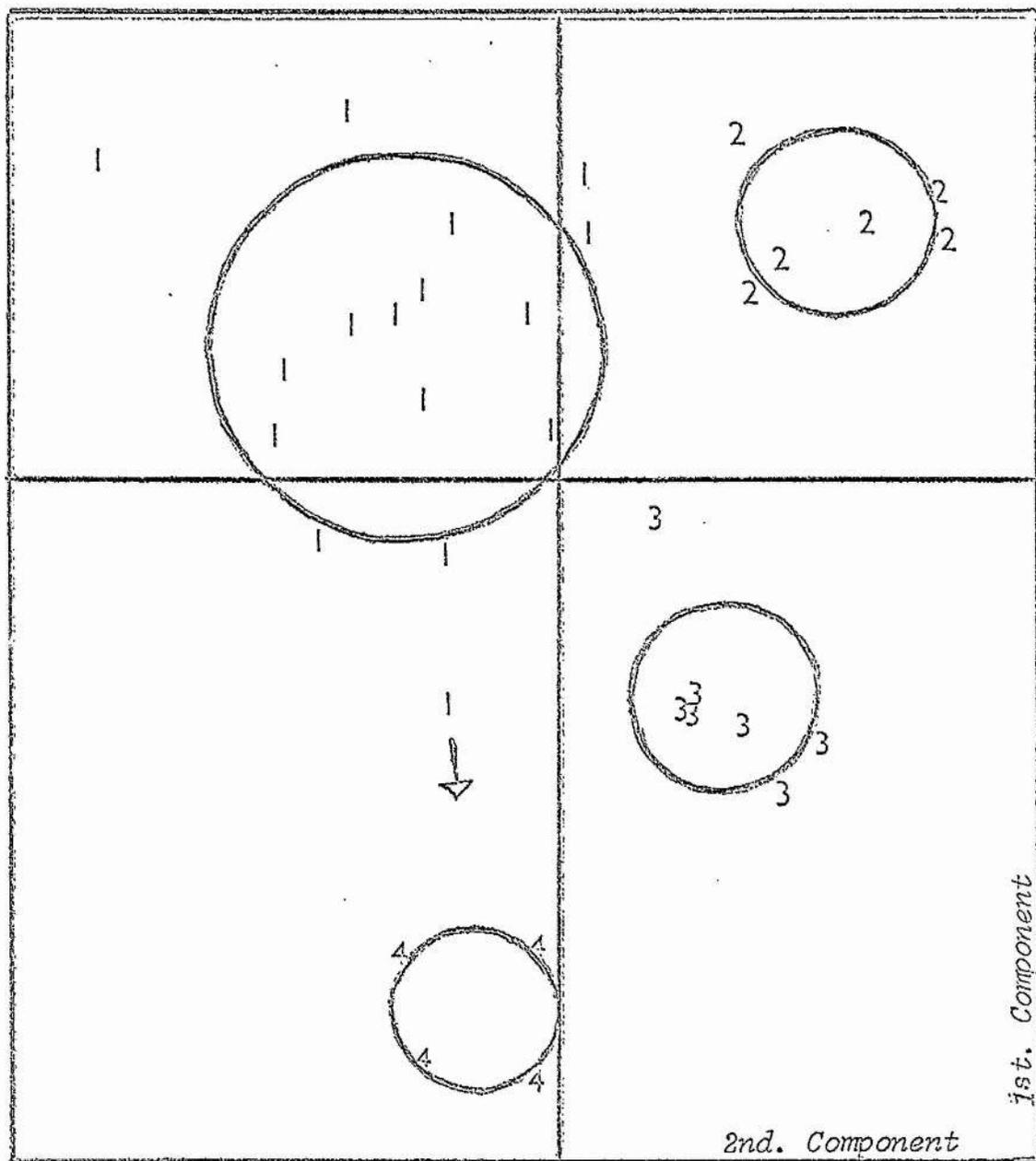


Figure 6. Classification of the 33 passages into 4 clusters by the methods of Ward and Wishart. The arrow indicates the reassignment of REP1 with cluster 4 by iterative relocation.

graphically. Figure 7 shows the distribution of points resulting from multidimensional scaling of the distance matrix (Table 3) for the 10 book data. This orientation was achieved after 33 iterations of the method, when the stress value (see Kruskal) had been reduced to 0.038. It is clear that the same basic pattern of clustering, cluster separation and ordering is repeated on this diagram.

Figure 8 shows the distribution of points along a line which was obtained by multidimensional scaling after 16 iterations, when the stress factor had been reduced to 0.384. This value for stress shows a marked rise over the 2 dimensional case, suggesting that considerable variation in rhythm has been sacrificed to project the points into 1 dimension. However, the ordering so obtained does correspond to the results for the clausulation method of Cox and Brandwood, and Levison et al (with the Symposium and Phaedrus omitted).

Conclusions

1. The chronological order for what we have taken to be the last group of Plato's writings, from the Republic to the Laws, is found to be the same for the dendrogram using Ward's method, the reduction by Kruskal's multidimensional scaling to 1 dimension and the clausulation method of Cox and Brandwood. The analyses reported here not only confirm the conclusion of Cox and Brandwood, but achieve the result by means of a more thorough investigation of the prose. During his lifetime, Plato certainly developed and

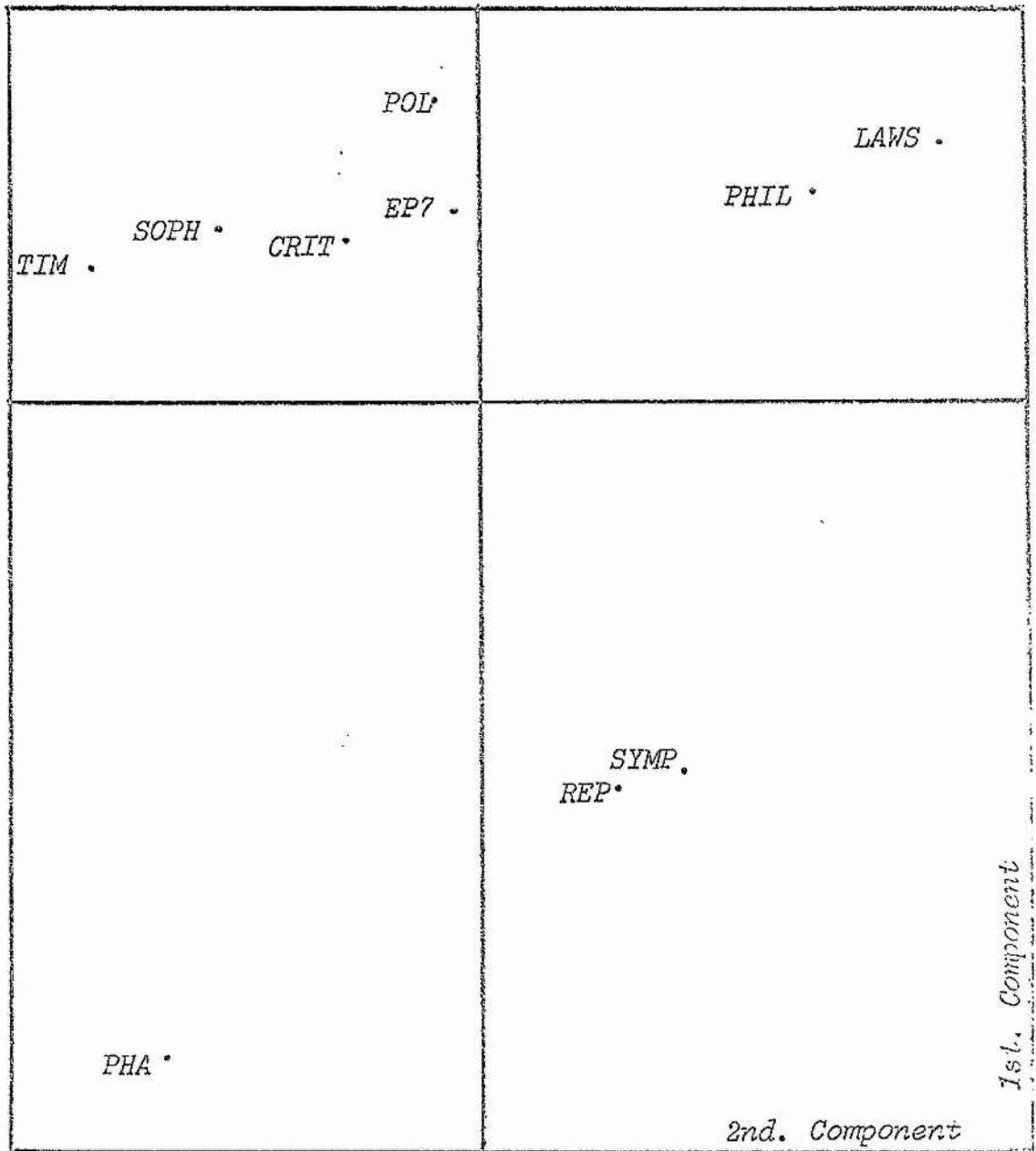


Figure 7. Two dimensional graph of the distribution of points resulting from multidimensional scaling of the unstandardised distance matrix (Table 3) for the 10 book data.

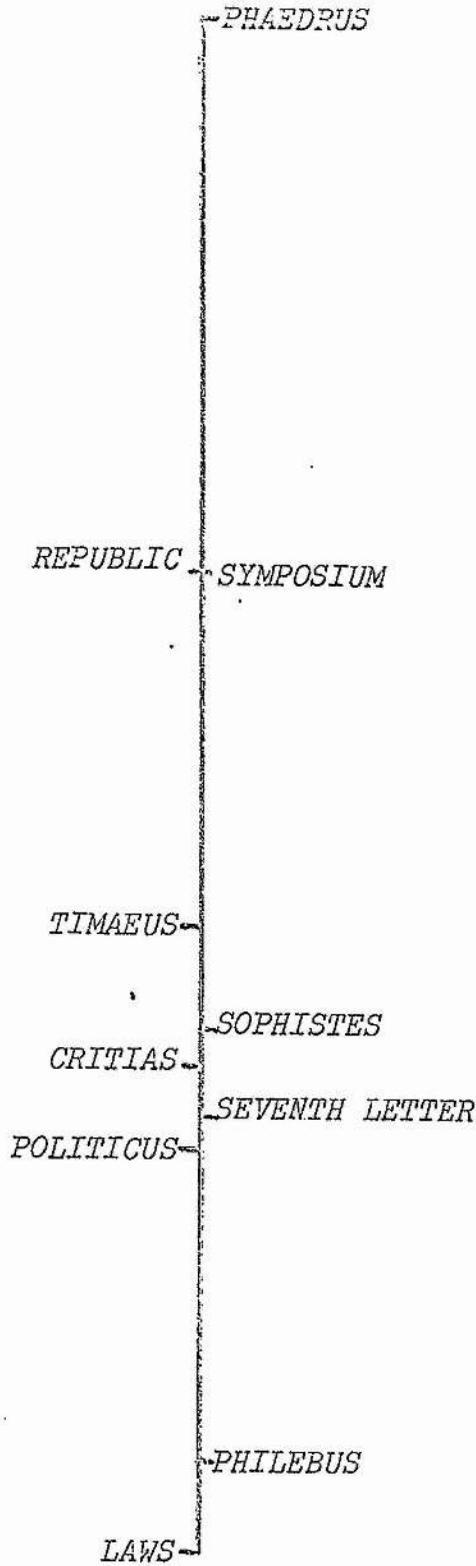


Figure 8. Projection of the 10 books into one dimension by multidimensional scaling using the distance matrix of table 3. The scale is believed to correspond closely to Platonic chronology.

altered preferences for clausulae, as one would have expected since clausulae played such an important role in classical prose-writing (cf Oxford Classical Dictionary 'Prose Rhythm' pp 738-40, para 11). However, the difficulty of maintaining predetermined ideas of rhythm throughout the prose means that the data constitute a broader representation of his rhythmic variation. We infer the following chronological order for the writings of Plato which we have considered:

PHA [REP & SYMP] TIM SOPH CRIT (EP7 POL) PHIL LAWS

The only deviation within the three methods is the ordering of the Seventh letter and Politicus. The Republic and Symposium appear together rather than in any particular order.

2. We find no stylistic evidence to support the present theory that the Phaedrus is a homogeneous late work. On all our diagrams it completely opposes the Laws, which we take to be the latest work considered, and since the other samples fall into a chronological order between the two, we are obliged to place the Phaedrus early. This is particularly evident in figure 3, where four of the Phaedrus samples (excluding the 'Lysias speech') fall neatly into one extreme group.

3. There is some evidence to suggest that the 'Lysias speech' is not closely related to the other passages from the Phaedrus, and it is possible that this passage may have been written by Lysias rather than Plato. While it might be objected that the

deviation occurs because of parody, the grouping together of the Symposium speeches would seem to be against such an argument.

4. The 33 passages or 10 books can be classified into the following four groups:

GROUP 1: Phaedrus.

GROUP 2: Republic (excluding Book II) and Symposium.

GROUP 3: Timaeus, Sophistes, Critias, Politicus and the Seventh Letter.

GROUP 4: Philebus and Laws.

5. The rhythmic differences in Plato's writings can be displayed graphically showing a major axis of variation which is attributed to his 'development' and corresponds to our present idea of the chronological order of his works, together with a secondary axis of variation in the rhythm of his later books that we associate with his experimental period. Furthermore, the results of principal components analysis and multidimensional scaling indicate that almost all the rhythmic variation can be expressed by this graph.

6. While we have attempted to combine five very different methods of analysis to check our inferences against the previous work of Cox and Brandwood, and Levison et al, there remains the possibility of weaknesses in the observation data or the sampling frame. By investigating the rhythmic patterns throughout the prose, rather than at the ends of sentences, we have tried to guard against the possibility of introducing errors or unwanted factors of variation. However, until such time as further

evidence can be produced to substantiate or contradict these findings, the results must be regarded as provisional. It is hoped to proceed with further work along similar lines.

Computer programs

The computer program, which was used to translate the original passages of text into long/short syllable coding was written in Fortran IV by S.V. Leach. This program also computes the percentage occurrences for the 32 5-syllable groups.

Programs which draw the scatter diagrams and dendrograms on a graph plotter, and evaluate the cluster analysis methods and principal components analysis are incorporated in a published suite of programs entitled CLUSTAN I (Wishart, 1969c), which has been written in Fortran II for the IBM 1620 and Fortran IV for the IBM 360 series.

The Fortran IV program for multidimensional scaling was provided by J.B. Kruskal of the Bell Telephone Laboratory, Murray Hill, New Jersey, USA.

The computations for this work were carried out at the University of St. Andrews Computing Laboratory using the IBM 1620 and IBM 360/Model 44 computers. Programs are available on application to the relevant authors.

Acknowledgements

We are greatly indebted to Professor K.J. Dover of the Department of Greek at the University of St. Andrews for suggesting the topic, and for his stimulating advice throughout. We also acknowledge our helpful discussions with Professor A.J. Cole, and the assistance of other members of the University of St. Andrews computing laboratory.

References

- Ball, G.H. (1966), 'A comparison of some cluster seeking techniques
Stanford Research Institute, California.
- Billig, L. (1920), 'Clausulae and Platonic chronology', J. Philol.,
v. 35, p. 225.
- Booth, A.D., Brandwood, L., and Cleave, J.P. (1958), Mechanical
Resolution of Linguistic Problems, Butterworth
Scientific Publications, London.
- Cox, D.R., and Brandwood, L. (1959), 'On a discriminatory problem
connected with the works of Plato', J.R. Statist.
Soc. B, v. 21, p. 195-200.
- Forgey, E. (1965), Biometrics, v. 21, p. 768 (abs).
- Jancey, R.C. (1966), 'Multidimensional group analysis', Aust. J.
Bot., v. 14, p. 127.
- Kruskal, J.B. (1964), 'Multidimensional scaling by optimising
goodness of fit to a nonmetric hypothesis',
Psychometrika, v.29, p. 1-27.

- Kruskal, J.B., and Hart, R.E. (1966) 'A geometric interpretation of diagnostic data from a digital machine', *The Bell Syst. Tech. J.*, v. 45, p. 1299-1338.
- Lesky, A. (1966), A History of Greek Literature, Methuen, London.
- Levison, M., Morton, A.Q., and Winispear, A.D. (1968), 'The Seventh Letter of Plato', *Mind*, v. 77, no. 307, p. 309-325.
- Oxford Classical Dictionary (1949), Oxford Univ. Press, p. 738-740.
- Sokal, R.R., and Sneath, P.H.A. (1963), Principles of Numerical Taxonomy, Freeman, London.
- De Vries, G.J. (1969), A commentary on the Phaedrus of Plato, Hakkert, Amsterdam.
- Ward, J.H. (1963), 'Hierarchical grouping to optimise an objective function', *J. Amer. Statist. Assoc.*, v. 58, p. 236.
- Wishart, D. (1968), 'Mode analysis: a generalisation of nearest neighbour which reduces chaining effects', *Proc. St. Andrews Colloq. In Numerical Taxonomy*; also in - Numerical Taxonomy (1969), Academic Press, London, p. 282.
- Wishart, D. (1969a) 'A numerical classification method for deriving natural classes', *Nature*, v. 221, p. 97.
- Wishart, D. (1969b), 'An algorithm for hierarchical classifications' *Biometrics*, v. 22, p. 165.
- Wishart, D. (1969c), 'Fortran II programs for 8 methods of cluster analysis (CLUSTAN I)', *Kansas Geol. Comp. Contr. No. 38*, Kansas, USA,

Appendix II. Fortran IV paging version of subroutine DISKIO using partial incore file-simulation (array A). This version is currently being distributed with all copies of CLUSTAN IA. Refer also to Section 10.2.

```

90401 SUBROUTINE DISKIC (ISEL,LSEC,IBIN,LB,XAR,LN)
90501 *****
90601 DESCRIPTION
90701 *****
90801 C THE DATA FILE STORED ON DISK IS DIVIDED INTO A SERIES OF DATA PAGES
90901 C WHICH ARE NUMBERED SEQUENTIALLY STARTING AT 1. EACH PAGE CCNTAINS
91001 C LKCRD REAL OR INTEGER WCRDS AND IS LOGICALLY SUBDIVIDED BY THE
91101 C SUBROUTINE INTC LREC 10-WORD RECCRDS.
91201 C THE FIRST NBLCKP PAGES ARE READ INTO ARRAY A AS REQUIRED, AND RESIDE
91301 C IN CORE THROUGHOUT THE DURATION OF A PROGRAM EXECUTION. A PAGE I IS
91401 C ONLY READ WHEN FIRST REQUIRED, AND THEREAFTER ALL TRANSFER CALLS CF
91501 C DISKIC WHICH REFERENCE THAT PAGE ACCESS CORE ONLY. ASSOCIATED WITH
91601 C PAGE I IS A READ/WRITE STATUS KEY IWRD(I) WHICH TAKES THE VALUES
91701 C 0, 1 OR 2 ACCORDING TO WHETHER PAGE I HAS -
91801 C (0) NOT BEEN READ INTO A,
91901 C (1) BEEN READ INTO A AND RECEIVED ONLY READ TRANSFERS (ISEL=4,6), OR
92001 C (2) BEEN READ INTO A AND RECEIVED ONE OR MORE WRITE TRANSFERS (ISEL=
92101 C 3,5).
92201 C THESE PROGRAMS WHICH WRITE TO DISK ALL FINISH WITH THE ENDCING CALL
92301 C OF DISKIC ISEL=2, AT WHICH ALL THOSE PAGES ASSOCIATED WITH IWRD(I)=2
92401 C ARE WRITTEN TO DISK. THIS EFFECTS ALL THE FILE UPDATES THAT HAVE
92501 C OCCURRED DURING THE EXECUTION.
92601 C IF THE DATA FILE EXTENDS BEYOND NBLCKP PAGES CF DISK, THEN ARRAY
92701 C DISK IS USED TO STORE ONE OF THE PAGES NUMBERED GREATER THAN NBLCKP
92801 C IN CORE AT ANY ONE TIME. IPAGE IS THE NUMBER OF THE PAGE IN ARRAY
92901 C DISK, AND IWR IS THE READ/WRITE STATUS CF IPAGE WHICH TAKES THE
93001 C VALUES 1 OR 2, AS BEFORE. IF A TRANSFER CALL OF DISKIC REFERENCES
93101 C PAGE IPAGE ONLY, THEN THE TRANSFER IS PERFORMED ON ARRAY DISK AND NO
93201 C DIRECT ACCESS OPERATION OCCURS. IF A TRANSFER OPERATIGN REFERENCES A
93301 C PAGE JPAGE (OTHER THAN IPAGE) WHICH EXCEEDS NBLCKP, THE SUBROUTINE
93401 C INITIATES A PAGE SWCP OPERATION. PAGE IPAGE IS WRITTEN TO DISK IF
93501 C IWR=2, THEN PAGE JPAGE IS READ INTO ARRAY DISK AND DISKIO SETS IPAGE=
93601 C JPAGE AND IWR=1. THE SUBROUTINE SETS IWR=2 FOR ANY WRITE TRANSFER TO

```

1
10
11

```

C SPECIAL CASE OF A COLD-START INITIALISATION OF CONSTANTS AND ARRAY
C IWRD. 94CC1
C THE CALL ISEL=1 IS USED AT THE START OF ALL OTHER PROGRAMS TO 94101
C INITIALISE DISKIO AND READ THE FILE PARAMETERS N-LNEXT AND TEXT FROM 94201
C PAGE 1. 94301
C THE CALL ISEL=2 IS USED TO WRITE THE FILE PARAMETERS TO PAGE 1, AND 94401
C ALSO WRITE UPDATED PAGES CONTAINED IN A TO DISK (AS PREVIOUSLY 94501
C DESCRIBED). 94601
C ***** 94701
C METHOD OF USE 948C1
C ***** 94901
C ***** 95C01
C SUBROUTINE DISKIO IS USED EXACTLY AS THE FORTRAN II VERSION (CARDS 95101
C 58C-113C) OF CLUSTAN I (SEE MANUAL, REFERENCE 52). 952C1
C THESE OPERATIONS ARE AS FOLLOWS - 95301
C ISEL=1. READ DISK FILE PARAMETERS N-LNEXT AND ARRAY TEXT. 95401
C ISEL=2. WRITE DISK FILE PARAMETERS N-LNEXT. 95501
C ISEL=3. WRITE INTEGER ARRAY (IBIN(I),I=1,LB) STARTING AT RECORD LSEC. 95601
C ISEL=4. READ INTEGER ARRAY (IBIN(I),I=1,LB) STARTING AT RECORD LSEC. 95701
C ISEL=5. WRITE REAL ARRAY (XAR(I),I=1,LN) STARTING AT RECORD LSEC. 95801
C ISEL=6. READ REAL ARRAY (XAR(I),I=1,LN) STARTING AT RECORD LSEC. 95901
C ISEL=7. NEW COLD-START INITIALISATION OF DISKIO FOR PROGRAMS WHICH 96C01
C DO NOT REFERENCE AN EXISTING DATA FILE, BUT CREATE A NEW ONE. 96101
C FOR OPERATIONS ISEL=3 TO 6, THE RECORD POINTER LSEC IS ADJUSTED BY 962C1
C DISKIO SO THAT IT POINTS TO THE NEXT IC WORD-RECORD WHICH FOLLOWS 963C1
C THAT RECORD WHICH CONTAINS THE LAST ELEMENT OF THE TRANSFERRED ARRAY. 96401
C FOR EXAMPLE, THE CALL ISEL=6 USED TO READ (XAR(I),I=1,35) WITH 96501
C LSEC=41, WILL RETURN LSEC=45 SINCE THE ARRAY OCCUPIES THE 4 RECORDS 96601
C 41-44. 96701
C PROGRAMS WHICH WRITE TO DISK MUST ALL END WITH THE FILE-CLOSING CALL 968C1
C ISEL=2. 969C1
C ***** 97CC1
C MODIFICATIONS 97101
C ***** 972C1

```



```

1 DEFINE FILE 2(6000,90,U,IV)
2 DIMENSION A(5400),IWRD(60)
3 DIMENSION DISK(90),LDISK(1),JBIN(1),XAR(1),L(1)
  EQUIVALENCE (DISK,LDISK),(A,L)
C BRANCH ON OPERATION SELECTOR
  GO TO (5,35,60,60,65,65,5),ISEL
C DECLARE PAGE LENGTH - DIMENSIONS OF DISK, LDISK
5 LWORD=90
C DECLARE NUMBER OF INCRE PAGES - DIMENSION OF A/LWORD
6 NBLCK=CC
C SET PROGRAM CONSTANTS AND INITIALISE IWR, IPAGE AND IWRD
  LRFC=LWORD/10
  L1=LREC-1
  L2=1-LWORD
  L3=NBLCK+1
  DO 10 I=1,NBLCK
10 IWRD(I)=C
  IWRD(1)=I
  IPAGE=L3
  IWR=1
  IF (ISEL-1)25,25,15
C PRINT INCORE SIMULATION MESSAGE WHEN ISEL=7
15 J=L3+LREC
  WRITE (6,20) J
20 FORMAT (5H MAXIMUM NUMBER OF RECORDS FOR INCORE PROCESSING =,I6)
  RETURN
C READ PAGES 1 AND (NBLCK+1)
25 READ (2,1) (A(I),I=1,LWORD)
  READ (2,13) DISK
C TRANSFER FILE PARAMETERS (ISEL=1)
  N=L(1)
  FB=L(2)
  PR=L(3)
  RPE=L(4)

```

104801
 104901
 105001
 105101
 105201
 105301
 105401
 105501
 105601
 105701
 105801
 105901
 106001
 106101
 106201
 106301
 106401
 106501
 106601
 106701
 106801
 106901
 107001
 107101
 107201
 107301
 107401
 107501
 107601
 107701
 107801
 107901
 108001
 108101

IMASK=L(7)
 IDATA=L(8)
 ICCCF=L(9)
 ITYPE=L(10)
 KMAX=L(11)
 LNDATA=L(12)
 LBDATA=L(13)
 LPEANS=L(14)
 LVARS=L(15)
 LCCRS=L(16)
 LEIGS=L(17)
 LEIGVS=L(18)
 LSCORS=L(19)
 LENGVS=L(20)
 LFREQS=L(21)
 LNPASK=L(22)
 LRMASK=L(23)
 LMAT=L(24)
 LKLJST=L(25)
 LNEXT=L(26)
 J=1
 DO 30 I=51,70
 TEXT(J)=A(I)
 30 J=J+1
 RETURN
 C JSFL=2, WRITE FILE PARAMETERS ON PAGE 1.
 35 L(1)=N
 L(2)=MD
 L(3)=MN
 L(4)=NPCI
 L(5)=NPC
 L(6)=I STAND
 L(7)=IMASK
 L(8)=IDATA

1084C1
 1085C1
 1086C1
 108701
 1088C1
 1089C1
 1090C1
 109101
 1092C1
 1093C1
 109401
 109501
 1096C1
 109701
 109801
 1099C1
 1100C1
 110101
 110201
 110301
 1104C1
 1105C1
 110601
 110701
 1108C1
 1109C1
 1110C1
 111101
 1112C1
 111301
 1114C1
 1115C1
 1116C1
 111701

```

L(11)=KMAX
L(12)=LNCATA
L(13)=LBCATA
L(14)=LWFEANS
L(15)=LVARS
L(16)=LCCRS
L(17)=LEIGS
L(18)=LEIGVS
L(19)=LSCURS
L(20)=LENGS
L(21)=LFREQS
L(22)=LNPASK
L(23)=LBPASK
L(24)=LPMAT
L(25)=LKLIST
L(26)=LNEXT
  IWRD(1)=2
C WRITE INCCRE PAGES TO DISK IF IWRD(1)=2
  DC 45 I=1,NBLCK
  IF (IWRD(1)-2)45,40,40
40 IA=I*LWORD+L2
  IB=IA-L2
  WRITE (2,I) (A(J),J=IA,IB)
45 CONTINUE
C WRITE PAGE IPAGE TO DISK IF IWR=2.
  GC TC (55,5G),I&R
50 WRITE (2,IPAGE) DISK
55 RETURN
C SET IB=LENGTH OF ARRAY TO BE TRANSFERRED.
60 IB=LB
  GO TO 70
65 IB=LK
70 IGPT=ISEL-2
  IA=1

```

```

LEP=(LSEC*(IP-1)/10+LI)/LREC
C SET JWR=1 CR 2 ACCORDING TO WHETHER THIS TRANSFER IS READ OR WRITE
  JWR=1
  GO TO (75,80,75,60),ICPT
75 JWR=2
C TEST LSP FOR INCRE PAGE TRANSFER
80 IF (LSP-NBLCK)85,85,165
C SECTION FOR INCRE TRANSFER
C SET STARTING ADDRESS OF A (LSAD) AND ENDING ADDRESS OF A(LEAD) FOR
C THIS TRANSFER -- INCRE. RESET ENDING PAGE NUMBER LEND AND ENDING
C ADDRESS LEAD IF THIS TRANSFER OVERFLOWS A.
85 LEND=LEP
  LSAD=LSEC*10-9
  LEAD=LSAD+10-1
  IF (LEND-NBLCK)95,95,9C
90 LEND=NBLCK
  LEAD=NBLCK*LWCRD
C INCREMENT RECCR POINTER LSEC.
95 LSEC=(LEAD+19)/10
C READ PAGES IF REQUIRED AND UPDATE INRD ACCORDING TO READ/WRITE
C INDICATOR JWR FOR THIS TRANSFER.
  DO 110 I=LSP,LEND
  IF (IWRD(I)-1)100,105,110
100 JA=I*LWCRD+L2
  JB=JA-L2
  READ (2,I) (A(J),J=JA,JB)
105 IWRD(I)=JWR
110 CONTINUE
C NOW PERFORM APPROPRIATE TRANSFER OPERATION ON A,L.
  GO TO (115,125,135,145),ICPT
115 DO 120 I=LSAD,LEAD
  L(I)=IBIK(IA)
120 IA=IA+1

```

```

112001
112101
112201
112301
112401
112501
112601
112701
112801
112901
113001
113101
113201
113301
113401
113501
113601
113701
113801
113901
114001
114101
114201
114301
114401
114501
114601
114701
114801
114901
115001
115101
115201

```

```

130 IA=IA+1
GC TC 155
135 DC 140 I=LSAD, LEAD
A(I)=XAR(IA)
140 IA=IA+1
GC TC 155
145 DC 150 I=LSAD, LEAD
XAR(IA)=A(I)
150 IA=IA+1
* C TEST WHETHER CURRENT TRANSFER OVERFLOWS A.
155 IF (LCP-NBLCK)/275, 275, 160
C CURRENT TRANSFER OVERFLOWS A, SC RESET STARTING PAGE LSP AND PROCEED
C TC DIRECT ACCESS TRANSFER
160 LSP=L3
C SECTION FOR DIRECT ACCESS PAGING OPERATION -- IS STARTING PAGE LSP=
C IPAGE, THE CURRENT INCCRE PAGE.
165 IF (LSP-IPAGE)/175, 170, 175
C PAGE SkOP OPERATION - WRITE IPAGE TO DISK IF IWR=2
170 GC TO (150, 155), IWR
175 GC TC (185, 180), IWR
180 WRITE (2, IPAGE) DISK
C READ PAGE LSP INTO CCRE
185 READ (2, LSP) DISK
C RESET READ/WRITE STATUS CF INCCRE PAGE.
190 IWR=JWR
C NGW SCAN FROM FIRST PAGE TO LAST PAGE FOR THIS TRANSFER
195 DO 270 IFAGE=LSP, LEP
C COMPUTE STARTING/ENDING ADDRESS LSAD/LEAD IN DISK FOR CURRENT PAGE
LSAD=(LSEC-LREC*IPAGE+LI)*IC+1
LEAD=LSAD+IB-IA
C MODIFY ENDING ADDRESS LEAD IF PAGE OVERFLOWS
IF (LEAD-LWORD)/205, 205, 200
200 LEAD=LWORD

```

```

117601
117701
117801
117901
118001
118101
118201
118301
118401
118501
118601
118701
118801
118901

```

```

119201
119301
119401
119501
119601
119701
119801
119901
120001
120101
120201
120301
120401
120501
120601
120701
120801
120901
121001
121101
121201
121301
121401
121501
121601
121701
121801
121901
122001

LLISK(I)=LDIR(IA)
215 IA=IA+1
    GO TO 250
220 DC 225 I=LSAD,LFAD
    IBIR(IA)=LDIR(I)
225 IA=IA+1
    GO TO 250
230 DC 235 I=LSAD,LFAD
    LISK(I)=XAR(IA)
235 IA=IA+1
    GO TO 250
240 DC 245 I=LSAD,LFAD
    XAR(IA)=LISK(I)
245 IA=IA+1
C INCREMENT RECORD POINTER LSEC
250 LSEC=LSEC+(LFAD-LSAD+10)/10
C IF LAST PAGE HAS BEEN READ, THEN EXIT WITH IPAGE = CURRENT INCCRE
C PAGE.
    IF (IPAGE-LPP)255,275,275
C TRANSFER NOT YET CUMPLISHED, SO INITIATE PAGE SHCP - WRITE PAGE IPAGE
C IF IWR=2,
255 GO TO (265,260),IWR
260 WRITE (2,IPAGE) LISK
C RESET IWR=JWR IF THEY DIFFER, AND READ PAGE (IPAGE+1).
265 IWR=JWR
    READ (2,IPAGE+1) DISK
270 CONTINUE
275 RETURN
    END

```

Appendix III: CLUSTAN distribution list.

The following is a list of all the organisations which have purchased complete copies of CLUSTAN IA, either on cards or on magnetic tape, from St. Andrews. The list excludes sales of the package by Computer Contributions, Kansas, and the distribution of individual programs.

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|--|---|
| Arthur Anderson and Co., St. Alphage House, 2 Fore Street, London, E.C.2. | Dr. F. Goronzy | ----- 3/9/68 ¹ |
| Department Geologique Central ELF.RE, 7 Rue Nelaton, Paris IVE, France | Dr. H. F. Leroy | ----- 3/9/68 |
| Lehigh University Bethlehem, PA 18015, USA | Dr. J. M. Parks, Director, Marine Science Center | CDC 6400 3/9/68 |
| University of Bradford, Bradford 7, England | Dr. R. J. Ord-Smith Director, The Computing Laboratory | ICL 1909 3/9/68 |
| University of Oxford, Oxford, England | Mrs. L. Hayes, Computing Laboratory, 19 Parks Road, Oxford | KDF9 23/3/69 |
| University of New York, New York, USA | Dr. K. M. Warwick, Consultant, 353 East 83rd Street, New York 10028 | CDC 6600 25/5/69 ² |

¹ First distribution of CLUSTAN I (Fortran II edition)

² First distribution of CLUSTAN I (Fortran IV edition)

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|---|---|---|
| Edinburgh Regional Computing Centre, The King's Buildings, Mayfield Road, Edinburgh EH9 3J2, Scotland | Mrs. J. Hornby | IBM 360/50 6/6/69 |
| Unilver Ltd., Colworth House, Sharnbrook, Bedford, England | Mr. I. W. Tully, Statistics Section, Unilever Research Laboratory | IBM 360/50 11/6/69 |
| University of Chicago, Chicago, Illinois 60037, USA | Dr. A. Herzog, Computation Center, 7094-7040 Operations - C B - 21, 5640 Ellis Avenue | IBM 360/50 IBM 360/40 17/6/69 |
| South Dakota State University, Brookings, South Dakota 57006, USA | Dr. M. D. Rumbaugh, College of Agriculture and biological sciences | IBM 360/- 14/7/69 |
| London School of Economics Houghton Street, Aldwych, London, W.C.2. | Mr. P. Wakeford | CDC 6600 15/7/69 |
| Research and Intelligence Unit, Greater London Council, County Hall, London, S.E.1. | Mrs. F. Kelly | ICL 4/50 30/7/69 |
| University of Guelph, Guelph, Ontario, Canada | Dr. A. A. Sheth, Section Head, Systems and Pro- gramming | IBM 360/50 25/8/69 |

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|---|---|
| Canada Department of Agriculture, Sir John Carling Building, Central Experimental Farm, Ottawa 3, Ontario, Canada | H. F. Beingessner, Chief, Data processing service, Room E-203 | IBM 360/65 25/8/69 |
| University of Texas at Austin, College of Education, Austin, Texas 78712, USA | Dr. H. F. Dingman, Department of Educational Psychology | CDC 6600 25/8/69 |
| United States Department of Agriculture, Forest Service, Northern Forest Fire Laboratory, Drawer 7, Missoula, Montana 59801, USA | Mr. R. E. Green | CDC - 25/8/69 |
| Yale University, New Haven, Connecticut 06520, USA | Mrs. B. Amato, Department of Statistics, Box 2179, Yale Station | ----- 25/8/69 |
| Kansas University, Lawrence, Kansas 66044, USA | Dr. D. F. Merriam, Chief, State Geological Survey | GE 635 25/8/69 |
| Cambridge University, Cambridge, England | Mr. N. J. Butler, Institute of Theoretical Astronomy, Madingley Rise, Madingley Road, Cambridge | IBM 360/44 19/11/69 ¹ |

¹First distribution of CLUSTAN IA

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|---|---|
| Brigham Young University, Provo, Utah 84601, USA | Dr. C. D. Jorgensen, Department of Zoology and Entomology, Brigham Young Uni- versity | IBM 360/ - 20/11/69 |
| Università di Milano, 20122 Milano - Via Francesco Sforza N.35, Italy | Dr. P. Faglioni, Clinica della Malattie Nervose e Mentali, Università di Milano | IBM 7094 24/11/69 |
| University of Washington, Seattle, Washington 98105, USA | Dr. L. Fisher, Department of Mathematics | ----- 27/11/69 |
| Nuffield College, University of Oxford, Oxford, England | Dr. C. Payne | Atlas 4/12/69 |
| University of Stockholm, P.O. Box 23144, Stockholm, 23, Sweden | Dr. I. Mattsson, Institute of Statistics | ----- 6/12/69 |
| Scientific Control Systems Ltd., Sanderson House, 49-57 Berners Street, London, W.1. | Mr. D. Falck | Univac 1100 9/12/69 |
| Harvard University, 55 Shattuck Street, Boston, Massachusetts 02115, USA | Dr. W. J. Carr, Department of Health Services Adminis- tration | ----- 9/12/69 |
| University of Newcastle upon Tyne, Newcastle upon Tyne 1, England | Head of Department, Geography | IBM 360/67 13/12/69 |

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|--|---|
| Institut für Kristallographie und Petrographie, Sonneggstrasse 5, 8006 Zürich, Switzerland | Prof. Dr. C. Burri | CDC 6600 17/12/69 |
| Clinical Research Centre, Medical Research Council, Division of Computing and Statistics, 171-174 Tottenham Court Road, London, W.1. | Mr. M. J. R. Healy | ICL 1909 29/12/69 |
| University of York, England | Dr. N. B. Usher, Department of Biology | ----- 30/12/69 |
| University of Western Ontario, London, Canada | Prof. L. Orloci, Department of Botany | ----- 13/1/70 |
| University College, Dublin, Ireland | Prof. P. Clinch, Department of Botany | IBM 360/50 14/1/70 |
| CEMREL, Inc., 10646 St. Charles Rock Road, St. Ann, Missouri - 63074, USA | Dr. T. J. Johnson | ----- 16/1/70 |
| Central College, Pella, Iowa 50219, USA | Dr. C. M. Humphrey | ----- 16/1/70 |
| Unilever Research Laboratory, The Frythe, Welwyn, Hertfordshire, England | Mrs. M. McCormick | IBM 360/50 19/1/70 |

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|--|---|
| University of Toronto, Toronto 5, Canada | Mr. A. J. Olbrecht, Department of Epidemiology and Biometrics | IBM 360/65 20/1/70 |
| Oklahoma State University, Stillwater, Oklahoma, USA | Prof. C. M. Dollar, Department of History | IBM 360/50 30/1/70 |
| University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637, USA | Dr. P. M. Lankford, Department of Geography | IBM 360/65 30/1/70 |
| Lunds Universitet, Östra Vallgatan 14, 223 61 Lund Sweden | Dr. J. Nilsson, Avd. för Ekologisk Botanik | ----- 5/2/70 |
| University of Miami, Coral Gables, Florida 33124, USA | Dr. R. G. Banks, Arts and Sciences, Room 323 Ashe Building, Main Campus | ----- 5/2/70 |
| University of East Anglia, Norwich, NOR 88C, England | Dr. J. Barkham, School of Environmental Sciences, University Village | ICL 1905 E 7/2/70 |
| Katholieke Universiteit, Nijmegen, Holland | Dr. C. J. M. Aarts, Faculteit der Wiskunde en Natuurwetenschappen, Toernooiveld, Driehuizerweg 200 | ----- 9/2/70 |
| Cornell University Ithaca, N.Y. 14850, USA | Prof. D. M. Jackson, Department of Computer Science | ----- 10/2/70 |

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|---|---|---|
| Colorado State University, Fort Collins, Colorado 80521, USA | Prof. T. J. Boardman, Statistical Laboratory | CDC 6400 13/2/70 |
| Vanderbilt University, Nashville, Tennessee 37203, USA | Dr. S. Hurley, Oxford House | ----- 17/2/70 |
| University of Idaho, Moscow, Idaho 83843, USA | Dr. D. E. Anderegg, 112 Life Sci. Bldg. | ----- 18/2/70 |
| University of British Columbia, Vancouver 8, Canada | Dr. Norman J. Wilimovsky Institute of Animal Resource Ecology | ----- 30/3/70 |
| University of Dundee, 15 Springfield, Dundee | Dr. John Rushforth, University Computing Laboratory | ICL 4120 10/4/70 |
| Consejo Superior de Investigaciones Cientificas Madrid 6, Spain | Fdo. Angel Gil, Centro de Cálculo Electrónico, Serrano 142 | IBM 360/65 27/4/70 |
| The Electricity Council, London, S.W.1. | Mr. D. Norman, Engineering Branch, 30 Millbank | ----- 27/4/70 |
| Imperial Tobacco Group Ltd., Bristol 3, BS3 1QX | Mr. Wyn Paige, Research Department, Raleigh Road | ICL 4/50 7/5/70 |
| Simon Fraser University, Burnaby 2, British Columbia, Canada | Dr. Wolf D. Rase, Department of Geography | IBM 360/50 7/5/70 |

| <u>Installation/Company</u> | <u>Reference</u> | <u>Computer and date of first acquisition</u> |
|--|---|---|
| The Gallup Poll, 211 Regent Street, London, W1A 3AU | Mr. Peter F. Baker | IBM 360 11/5/70 |
| Wollongong University College, Wollongong, N.S.W. 2500, Australia. | Dr. A. C. Cook, Geology Department | IBM 360 14/5/70 |
| Memphis State University, Memphis, Tennessee 3811, USA | Mr. David N. Lumsden, Herff School of Engineering, Department of Geology | ----- 14/5/70 |
| University of California, Los Angeles, California, USA | Prof. Peter M. Bentler, Psychology | ----- 15/5/70 |
| University of Manitoba, Winnipeg, Manitoba, Canada | Dr. P. J. Kaltsikes, Plant Science, E302 Plant Science Building, Fort Garry 19 | ----- 18/5/70 |
| United States Steel Corp., Pittsburgh, Pennsylvania 15217, USA | Mr. E. D. Duggins, Assistant Manager, Engineering and Scientific Computer Services, 1509 Muriel Street | ----- 19/5/70 |
| University of Oklahoma, Norman, Oklahoma 73069, USA | Dr. Paul G. Risser, Botany/Microbiology Department | ----- 9/6/70 |
| C oras Iompair  ireann, 5 Kildare Street, Dublin 2, Eire | Mr. John Markham, Research and Develop- ment, Office of the Manager | IBM 360/50 10/6/70 |
| University of Aberdeen, Aberdeen, AB9 2UD | Dr. Peter Ashton, Department of Botany, St. Machar Drive | ICL 4/50 10/6/70 |

INDEX

A

Agglomerative (see also Hierarchic fusion), 52, 94-95
Agglomerative Group Analysis, 108, 126
Allocation, 97
Almost complete Q-subset, 212
Andean Survey (see also pp. 323-341), 66, 76, 83, 86, 246
ASCOP, 232
Association analysis, 2, 75-78, 126, 251, 253
Attributes (see also Binary data), 10, 115
Average distance, 49, 50, 99, 119, 125, 126, 152, 171, 175, 180, 186
Average linkage, 53, 55, 99, 189, 196
Average similarity (see also Average distance, Average linkage), 51, 53, 55

B

Base subset, 213
Basic classifications, 155
Between-group sum of similarities, 104
Binary: Data, 9, 39-50, 113, 115, 250; Linkage matrix, 217; Matrix (see also Binary data), 10, 39, 72; 2x2 table, 45, 116
Bivariate normal, 168, 175, 183

C

Canberra metric (see also Nonmetric coefficient), 44, 46
Catalogued procedures, 235
Central limit theorem, 167
Centroid, 48, 49, 59, 65, 80, 100, 113, 117
Centroid sorting, 53, 58-59, 61, 65, 67-70, 95, 125, 189, 195, 251
Chaining, 2, 54, 67, 75, 95, 140-142, 190
Characteristic vector, 80
Characters, 133
Chi-square, 65, 72-87
City-block metric, 85

Clumping, 96, 104, 109
Clumps, 107
CLUSTAN (see also pp. 430-437) 42, 127, 236, 248-254, 258
CLUSTAN - limitations, 249
Cluster: Binary data, 49, 115; Continuous data, 48, 113; Diagnostics, 252, 254; Diam 55, 132; Division (see also Monothetic and Polythetic division), 114, 121, 126; F (see also Intercluster simi function), 110, 113; Initia 97, 101; Minimum size crite 183, 252; Nuclei, 144, 187; Parent, 98, 122; Recognitio algorithm, 217-226, 252; Sh 142; Structure, 40, 48, 56; matrix, 39, 113; Variance, 48, 49, 119, 130, 132
Clusters: Chain, 205, 225; Di 113, 139, 256; Elongated, 1 171, 178, 186, 255; Natural also Natural classes), 131, Overlapping, 108, 205; Para elongated, 142, 166, 178; Stragglng, 54; Tight, 55, 186, 255
Cohesion, 96, 104, 126
Cohesion functions, 104, 126
Cole-Wishart algorithm, 207-2 252
Combinatorial: Algorithm, 198 251; Coefficients, 188-204; Flow chart, 201; Solution, 188-204, 251
Complete classifications, 148
Complete linkage, 51, 53, 55,
Complete Q-subset, 212
Component of distance, 8, 21-
Computation time, 82, 111
Confidence test, 65, 79, 183
Contingency table, 143, 149,
Continuous data, 5, 39-50, 11 250

C

Convergence, 101, 170, 176
Conversational-mode, 230, 232, 258
Correlation, see Product-moment
correlation
Cosine, 43, 46, 58, 118, 171
Counter-chaining (see also Chaining),
56
Cut-off point (see also Dendrogram,
Division tree), 66, 74

D

Data matrix, 5, 25, 39
Data storage, 238, 258
De-chaining (see also Chaining), 56
Degrees of freedom, 66, 78, 79
Dendrite-shortest (see also Minimum
spanning tree), 89-91, 252
Dendrogram, 66-70, 72, 162, 181, 207,
253
Density: 80, 143; Estimates, 152, 163,
180, 184; Function, 149-153, 257;
Surfaces, 150, 161
Direct Access, 234, 238
Directional coefficients, 108, 186,
256
Discrete probability function, 63
Diseases (see also pp. 368-377), 138
DISKIO (see also pp. 421-429),
238-248, 258
Disorder, 40, 59, 63, 111
Dissimilarity analysis, 91-93, 96
Dissimilarity coefficients, 40, 107,
189, 204
Distance, 7, 40, 42, 45, 58, 61, 70,
86, 99, 111, 116, 117, 125, 131, 132,
144, 171, 175, 178, 186, 188, 257
Diversity, 64
Division tree (see also Dendrogram),
72, 83-84
Divisive methods (see also Monothetic
and Polythetic division), 94, 114,
121, 126
Dot product, 24, 43, 46, 81, 99, 108,
118, 125, 171, 196
Dynamic file simulation, 247

E

Ecological surveys, 79, 108, 1
Eigenvalues, 29
Eigenvectors, 27
Entropy, 39, 40, 63-66
Error of fit (see also Error s
of squares), 102
Error sum of squares, 48, 50,
66, 85, 87, 89, 101, 108, 11
126, 130, 132, 171, 175, 186
252, 257
Experimental tests, 166-187
Explicit objects, 218

F

Factor: Loadings, 27; Scores,
Factors: Interpretation, 31;
Number of, 28
Flexible, 126, 190
Fragmentation, 2, 75, 79, 86
Frequency vector, 115, 119, 12
Fusion (see also Hierarchic fu
51-71, 97, 111, 114, 120

G

Generalised tripartite procedu
120-127, 255
Graph theoretic model, 255
Gravitation model, 142
Group analysis, 79-84, 108, 12
251, 253
Group average (see also Averag
linkage), 56
Group coefficient, 81-84

H

Hierarchical mode analysis: 14
166, 180, 250, 257; Improved
153-165
Hierarchic fusion (see also Fu
51-71, 120, 125, 189, 251
Hierarchy of clusterings, 223
Histogram, 143, 149, 159
H-R diagram, 129

I

Incore file simulation, 242
Information: 50, 63-66, 70, 77-79,
104; Average, 64; Gain, 50, 63-66,
77-79, 104, 119, 126, 256; Total, 64
Interaction statistic, 81-84
Intercluster similarity function,
113-120, 255
Internal file, 239, 244, 254
Iterative relocation, 96-112, 116, 122,
126, 166, 168, 186, 246, 252, 257

J

Jardine-Sibson algorithm, 206

K

k-mean, 100, 126, 178
k-partition, 205-229, 252

L

Large populations, 150, 171, 224
Latent vectors (see also Eigenvectors),
61
Linkage: Analysis, 54; Methods, 54-58,
125; Parameter, 144, 205; Tree (see
also Dendrogram), 66-70
Local density, 152

M

Matching coefficient, 42, 47, 61
Maximal complete subgraphs, 205, 225
Mean (see also Centroid), 5
Mean group density, 80
Median distance, 60-62, 125, 190
Median linkage, 56-58
Minimum spanning tree, 54, 89-91,
252
Minimum variance, 129-138, 180, 186,
190, 255
Misclassifications, 101, 108
Mode: Analysis (see also Hierarchical
mode analysis), 143, 150; Flow chart,
147, 165; Improved algorithm, 165
Modes, 133, 138, 140, 143
Modular compilers, 235

Molecular models, 33
Monothetic division, 65, 72-87
126, 246, 251
Monotonicity, 67
M-space, 2, 32, 140, 143
Multistate: Characters, 13, 20
Ordered, 13-18, 23; Unordered
13-18, 22
Multivariate normal, 166, 257

N

Natural classes, 131, 134, 138
166, 186, 205, 256
Nearest neighbour (see also Si-
linkage), 55, 99, 149
Noise, 140, 144
Nonmetric coefficient, 44, 46,
171
Normalisation, 13, 19, 250
Normalised correlation (see al-
Cosine), 43, 46
Normal number generator, 167
Null-hypothesis, 65, 79

O

Object file, 239
Optima: Local, 103, 172, 258;
Global, 103, 109, 258
Origin dependence, 169, 171, 2
Oscillation, 98, 101, 107, 110
Output of classifications, 146
Overlap objects, 218

P

Paging diskio (see also DISKIO
241, 248
Part-optimum solution, 101, 10
109, 110, 170, 178, 186
Pattern difference, 47
Plant community, 130
Plant ecology (see also Ecolog
Surveys), 108, 130
Poland, 153
Polythetic division, 87-95, 12
Population-partition, 172, 176
178, 186

P

Potential, 81, 108
Principal components, 25-39, 133, 157, 250, 253
Principal plane, 32, 35
Principal 3-space, 33-39
Probabilistic model, 129-165, 256
Probability-attribute, 11, 22, 49
Probability surface (see also Density function and estimates), 150
Product-moment correlation, 27, 43, 45, 59, 84, 117, 171
Program: CENTRO, 251, 253; CORREL, 250, 252; DIVIDE, 251, 253; DNDRIT, 252; FILE, 250, 252; HIERAR, 251, 253; KDEND, 216, 252; MODE, 250; PLINK, 253; RELOC, 252; RESTART, 253; RESULT, 252; SCAT, 253; STORE, 253
Program packages, 230, 231
Proportional link linkage, 57

Q

Qualitative data (see also Binary data), 10
Quantitative data (see also Continuous data), 5

R

Random start, 101, 110, 171, 178
Reallocation (see also Iterative relocation), 97, 109
Reassignment (see also Iterative relocation), 97, 109
Rectangular distribution, 167
Relocation test, 98, 102, 104, 108, 110, 123, 252
Residue, 101, 252
Reversals, 67, 79
Rotation, 34-39

S

Scatter diagrams, 32-39, 253
Set element potential, 81, 108
Shape difference, 44, 46, 118, 171, 197

Significance test, 65, 78, 79,
Similarity: Analysis, 56, 126,
Coefficients, 39-50, 113-120, 166-180, 189, 204, 250; Funct 48-50, 52-54, 86, 92, 98; Mat 52, 57, 61, 116, 120, 125, 14 188, 205, 228, 233, 250; Rati 43, 46, 118, 171, 175, 186
Single linkage, 2, 51, 53-55, 5 67-70, 85, 95, 139, 144, 189, 207, 225, 230, 257
Size difference, 44, 46, 118, 1 197, 255
Smoothing, 183, 187, 257
Special-purpose programs, 230
Spherical neighbourhood, 144, 2
Sneath's method (see also Singl linkage), 51, 53-55, 139, 189
Sorting level (see also Thresho 51, 139
Standardisation: Binary, 11, 19 Continuous, 8, 19, 30, 137, 1 186, 250
Standard scores, 9, 28, 250
Starting solution, 101, 103, 10 110, 166, 169, 257
Statistical: Languages, 230, 23 Systems, 230, 258
Stopping rules, 76, 79, 97
Structure (see also Cluster), 2 32, 40, 56
Subdivision: Hierarchic, 74, 84 122, 251; Nested, 74, 88, 91, 122, 251
Subroutine systems; 230, 231, 2
Syndrome analysis, 55

T

Taxon, 133
Threshold, 51, 139, 143, 183, 2 225, 252
Transformations, 14-21, 135, 25

U

Unweighted variable group (see Average linkage), 55

U

USER facility, 127, 250

V

Variance, 7, 40, 49, 50, 119, 130, 171,
175

Varimax, 31

W

Ward's method, 59-60, 101, 103, 125,
129, 158, 191-195, 253, 257

Weighted variable group (see also
Centroid sorting), 59

Weighting, 14, 21

Within-cluster distances, 99, 252

Within-group variance, 102, 130

Y

Yates' correction, 77