AN EVALUATION OF CLUSTER ANALYSIS AND RELATED MULTIVARIATE TECHNIQUES FOR OPERATIONAL RESEARCH

by RAY NEWPORT

submitted to the Graduate Business Centre,
City University, for the degree of
Doctor of Philosophy

JULY 1974

AN EVALUATION OF CLUSTER ANALYSIS AND RELATED MULTIVARIATE TECHNIQUES FOR OPERATIONAL RESEARCH

VOLUME 1

Acknowledgements

I am indebted to Al Russell of the Graduate Business Centre for his help and guidance during this project, and in particular to his faith in multivariate methods.

I would also like to pay tribute to the excellent library and computer facilities of the GBC and City University, on which this work has depended so much.

I would like to thank the following people for their help in the case studies given in the Addenda - John Faupel of the Printing and Publishing Industry Training Board for the manpower data, Dr. John Green of the Local Government Operational Research Unit for first introducing me to the sewer pipes problem, and Jock Scholefield of the GBC for information on the team organizing problem.

I am especially grateful for the accurate typing, patience and good humour of Ann Newport.

Preface

The following work is an investigation into methods of Cluster Analysis and Ordination. The main objective of this thesis has been to investigate the capabilities of these methods for practical usage. An important subsidiary aim has been to collect together related work which has been carried out in many different areas - ecology, biology, archaeology, psychology, etc., into one work.

After a brief introduction to the general concepts of multivariate analysis in Section A of the thesis, Section B gives an introductory account of the methods of Clustering, Ordination and Seriation, putting them into the context of the by now better established multivariate techniques. Section C considers Cluster Analysis in depth, explaining and examining various methods reported in the literature, together with methods developed by the author. The suitability of the methods for practical use is discussed and decision rules are set out for the choice of method to be used in any particular study, based on the results of extensive comparative tests of the methods.

In Section D the various ordination methods are considered, giving an overall view and relating the methods to each other. Particular emphasis is paid to the rather neglected metric methods.

Section E, after a survey of published applications of the methods, suggests new areas where the methods previously discussed could be valuable aids for data investigation and problem solving. An Addenda is included which describes several operational research case studies using these methods.

Computer programs are given for the most successful of the newly introduced cluster methods, and an extensive reference section is also included.

LIST OF SYMBOLS USED CONSISTENTLY THROUGHOUT

N , n	the number of observations or objects
M , m	the number of variables
S , s	any similarity measure, unless otherwise specified
S(a,b)	the similarity between the objects or sets a and b
D	any dissimilarity measure
D(a,b)	the dissimilarity between the objects or sets a and b
d	a dissimilarity measure, usually constrained to be in a lower dimensioned space than D

A

INTRODUCTION

A. MULTIVARIATE ANALYSIS AND OPERATIONAL RESEARCH

In various disciplines such as psychology, biology, botany, and, more recently, business studies, there are many situations where we have large sets of data which depend on many variables. The analysis of many of these data sets is complex because of the interaction between variables - one useful statistical approach is the use of multivariate analysis techniques, which can take account of inter-variable relationships. These methods were developed largely in the psychological and biological fields, where large data sets are abundant, and few methods previously existed for their adequate analysis. These methods are now used in most sciences, and their use is currently being examined in the social sciences.

The definition of multivariate analysis (m.v.a.) is rather vague; in its widest sense it could be thought of as encompassing the whole field of statistics, but in practice, boundaries need to be placed on our definition. For our purposes we use the definition due to M. Kendall (1968) which extracts two essential features of m.v.a:

(a) We are concerned with a set of n individuals, each of which bears the value of p different variates. The multivariate character, so to speak, lies in the multiplicity of the p variates, not in the size of the set n.

(b) The variates are dependent among themselves so that we cannot split off one or more from the others and consider it by itself. The variates must be considered together.

Although there exist several mathematical methods of m.v.a., the sum of the methods do not comprise the science. The same remark can be made regarding operational research. Both m.v.a. and operational research are approaches to problems which may involve models but they are primarily ways of looking at problems.

The approach of multivariate analysis is that of a search technique. Information is sought from a set of data as to whether or not it has structure, and the nature of any structure present. Because of the large computer times with some multivariate methods, this analysis of data can be costly and thus, as with other search techniques, one important aspect is to balance the cost of the search with the expected gain.

The operational research worker is concerned with the solution of problems in the control of all aspects of an organized system (often by mathematical means), but not with the decision taking within that system. The trend towards increased complexity of organizations, and the need for greater control led to the early growth of O.R. The continued growth is a measure of the effectiveness of O.R. as an aid to organizational control, and as a method of increasing the knowledge of systems behaviour, which may lead to future benefits.

Most management situations, which are amenable to an O.R. approach, involve many variables. Often, in order for any solution to be found at all, one selects from interacting

variables those which seem most relevant. However any attempt to study the effect of change in one variable in isolation from the rest is often meaningless. Multivariate analysis brings to these problems a battery of techniques, which are essentially multidimensional in their approach, taking into account the interplay between variables. It is surprising that so many published works and degree courses concentrate on methods for deterministic situations. In organizational problems there are relatively few such situations. Multivariate analysis however represents a set of techniques for predictive and descriptive operational research.

History

Multivariate analysis has been in existence as a statistical method for most of this century. Its history can be divided into three parts:

1890 - 1930 Isolated early works on multivariate methods following Galton's correlation coefficient first published in 1888.

1930 - 1950 The discussion on the structure of human ability by Burt, Thurstone, Spearman, etc., in psychology, which demanded new methods to analyse sets of intelligence tests to try and determine the structure of intelligence.

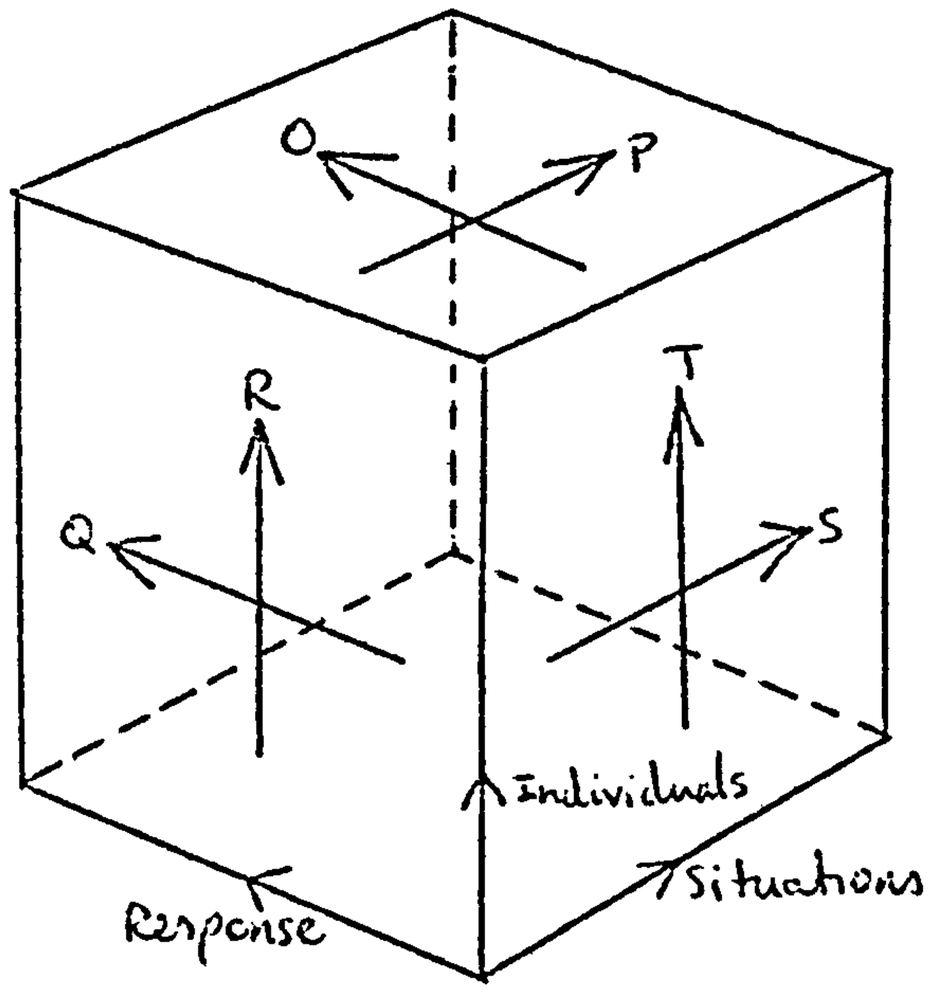
1950 to date The rapid expansion of techniques to other disciplines, and the increase in number of methods available and the size of problem that could be tackled, with the advent of modern computers.

The Data Box

One convenient way of displaying the relationships between some of the multivariate analyses that may be carried out is by use of what has been called by Cattell (1965) the data box. This is based on a classification of organizational phenomena into three classes on which measurements may be made:

1. Individuals (e.g. people, countries, flowers)
2. Responses (e.g. size, temperature, colour)
3. Situations (e.g. times, places, environments)

These form a three-dimensional matrix which can be visualized as a box, the faces of which refer to types of multivariate analyses.



Thus we can describe six types of technique:

Q-techniques analyse individuals according to responses e.g. cluster analysis, discriminant analysis, correlation.

R-techniques analyse the relation between responses using individual scores e.g. principal components analysis, factor analysis.

P-techniques are concerned with variations in response over time e.g. spectral analysis.

O,S,T-techniques are more difficult to visualize but their meaning can be extracted from the diagram of the data box. These methods have been little used, but Cattell (1965) refers to O-technique being used in meteorology and Goronzy (1969a) refers to it being used in clinical psychology.

In a particular analysis, which constitutes the individual, the response, or the situation may not be clear cut and indeed one may decide to use a particular technique twice on the same matrix, by simply transposing the matrix and repeating the analysis.

Cattell (1966) has since proposed a more comprehensive data box in ten dimensions called the Basic Data Relation Matrix (BDRM). The suggested dimensions of the BDRM are:

1. Person or organism
2. Focal stimulus
3. Non-focal stimulus (environmental background other than stimulus)
4. Response or unitary ongoing process
5. Observer
6. State of the organism
7. Variant of the stimulus
8. Phase of the environmental background
9. Style of the response
10. Condition of the observer

Of these, items 1, 2 and 4 refer to the co-ordinates of the original 3D data box. Items 6, 7 and 9 are related to the exact nature of the items 1, 2 and 4. Co-ordinates 3 and 8 refer to subsidiary stimuli or situations, and items 5 and 10 refer to the observer who can be a further influence on the situation.

The listing of the data box dimensions is of particular use in consideration of how the dimensions which we are not including in a study may affect it. Most present techniques only enable a face of the box to be examined and thus other dimensions we usually ignored. Since many of the faces of the box represent techniques which are difficult to picture, and indeed, the exact face we are examining may be doubtful, the BDRM is of more use as a set of variables which have an effect on data, than a method of picturing techniques. The list can also be extended to include, for example, the effect of several non-focal stimuli, several observers, etc.

Multivariate analysis is best explained by the consideration of some of the techniques. We enlarge upon these in the next section.

B

MULTIVARIATE ANALYSIS

1. Methods of MVA
2. Cluster Analysis
3. Ordination
4. Seriation
5. Dissimilarity and Similarity Measures

B.1 METHODS OF MULTIVARIATE ANALYSIS

There are several types of multivariate analyses, all having their own approach to simplifying data blocks. The major methods may be listed:

1. Principal components
2. Factor analysis
3. Canonical correlation analysis
4. Spectral analysis
5. Discriminant analysis
6. Cluster analysis
7. Ordination techniques

Explanations of some of the above methods may be found in M. Kendall (1968), Hope (1968) and Cooley and Lohnes (1971). There have also been several books published recently on the application of these techniques in particular disciplines, such as L. King (1969) on geography, Miller and Kahn (1962) on geology, D. Clarke (1968) on archaeology, and Green and Tull (1970) on marketing.

Principal Components Analysis

This method is perhaps the most well-known and most widely used of the multivariate methods. If we consider data as a set of points in n-dimensional space then principal components analysis is concerned with the rotation and translation of the original co-ordinate axes to a new frame of reference in this space. The first principal component is that linear combination of the original variables which

contributes the maximum to the total variance. The second principal component is orthogonal to the first and is such that it contributes the maximum to the residual variance. Further principal components may be obtained until the total variance has been 'explained'.

The model is therefore given by:

$$P_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i=1,\ldots n)$$

where P_i is the i^{th} principal component and x_j are the original variables, and the P_i's are orthogonal.

The aim of the model is to determine the effective dimensionality of the original variables, and to study the underlying variables of the data set. By using p.c.a., one hopes to be able to reduce the number of variables, and to obtain a more meaningful set. Since the method is dependent on the total variance of the original variables it is customary to normalize the variables before performing p.c.a. in order to have variables measured in the same units.

Details of the mathematics involved in extraction of components can be found in Cooley and Lohnes (1971), Harman (1967), and Seal (1964).

Factor Analysis

Factor analysis is currently one of the most important techniques of multivariate analysis. The aim of factor analysis is the resolution of a set of variables linearly in terms of a smaller number of variables, called factors. We

thus are trying to explain our M dimensional data in terms of a specified number of factors m, and a unique variation (an 'error' term) in each variate. The model is:

$$x_{ik} = \sum_{j=1}^{m} a_{ij}F_{jk} + d_iU_{ik} \qquad (i=1,\ldots n)(k=1,\ldots M)$$

each of the F_{jk} is called a common factor and each of the U_i a unique factor. The common factors are usually orthogonal, although methods for the construction of oblique axes exist.

There are an infinite number of ways of obtaining factors from a set of data, depending on the required form of the factors. Thus several methods exist (see Harman 1966, M. Kendall 1968, Horst 1965, Lawley and Maxwell 1971). Perhaps the most widely used methods are Lawley's maximum likelihood (Lawley 1940) and the Minres method of Harman and Jones (1966). A typical example of a study using factor analysis is Baehr and Williams' (1967) study of personal data and its relationship to occupation.

Canonical Correlation Analysis

This concerns two different sets of data on the same individuals and the method assesses the correlation between the two sets, and the linear function of each set of variables which gives that correlation. This is repeated as in principal components to find the largest correlation orthogonal to the first and so on. The aim of c.c.a. is to determine the redundancy in two sets of variables and the connected factors. This is covered in Cooley and Lohnes (1971), M. Kendall (1968) and Hope (1968).

Spectral Analysis

This is the analysis of time series, especially that of sinusoidal series, and the removal of autocorrelation in multivariate data, which is used in forecasting, and in pattern recognition and signal detection. The methods and problems of this type of analysis have recently come into prominence with the introduction of a new technique by Box and Jenkins (1970) (see also Chatfield and Prothero 1973, and Box and Jenkins 1973). Spectral analysis is also discussed in a book by M. Kendall (1973).

Discriminant Analysis

If a set of data consists of two or more known classes (i.e. two types of plant, people who failed or passed an exam, etc.) then the data matrix is divided into that number of groups and gives an identification routine so that a new set of observations can be assigned to these classes so that the probability of incorrect assignment is minimized. This is normally performed for only a few classes (2, 3 or 4) and is designed for normally distributed classes which overlap. For a discussion of the method see Cooley and Lohnes (1971) or M. Kendall (1968), or the original work of R. Fisher (1937). Some interesting recent examples are Brainerd (1973), Massey (1971) and Graham (1970).

Cluster Analysis

This is concerned with the existence of groups of similar objects within a given set of data. The purpose of the study might be to determine if such groups do exist or

not, or to divide a set of data into the 'best' groups for a particular purpose. The techniques will be discussed at length in this work. There are few (if any) major published works on cluster analysis. The best references are chapters from books on pattern recognition such as Meisel (1972) and Duda and Hart (1973), there is also a good review by Bolshev (1969).

Ordination

This is the representation of data in ordinate form, normally with the view of drawing the objects as points in one, two or three dimensions, so that their structure can be visually examined. Both principal components analysis and factor analysis can be used as ordination methods, but other newer techniques also exist (such as multidimensional scaling), which are primarily designed for producing ordinate representations.

Classifications of multivariate methods have been attempted by M. Kendall and Babington Smith (1950), Sheth (1971) and Kinnear and Taylor (1971). In all three, the methods are divided firstly into dependence and interdependence techniques. Of the above techniques, spectral and canonical correlation analyses are of the dependence type, and principal components and factor analysis are interdependence models. The other three techniques do not fall into either of these categories, since they are concerned both with dependence and interdependence. We thus feel that the division is an artificial one.

Of the methods we have outlined above, we have selected cluster analysis and ordination as the subject of this work. The reason for the selection of these two techniques is the lack of, and need for, research into these areas. All the other methods which we outlined have been the subject of lengthy analysis by other workers, and over the years one or two methods of performing each technique have come out as the 'best' methods. Also each method has been applied to many fields of interest, has been programmed into readily available packages, and been discussed in several books. In contrast, there have been proposed over a hundred different cluster analysis methods, and most of these have not been compared or analysed to any great extent. In fact, the progress of cluster analysis has been hampered by the lack of a 'best' method. Ordination methods have not as yet suffered from over-plentiful methods. However, especially with the new multi-dimensional scaling techniques, there is a lack of comparative studies between the methods.

Thus one reason for our selection of these two methods is the need for a comparison study in each case. Another area of investigation from which each technique would benefit is the practical applications in which each could be used.

The rationale in examining both methods together is that they are complementary techniques for examining data structure, which do not have underlying assumptions of dependence, etc.

One of our aims in this work will be to bring together the large amount of literature on the subject, and the methods which have been proposed, into one volume. The need for techniques for examining data structures in many sciences has lead to the development of cluster analysis and ordination in many fields, most workers having been unaware of parallel developments in other sciences.

Our second aim, following this introduction of the techniques, will be the comparison of a variety of cluster analysis methods, and an investigation of the properties of ordination techniques.

The third part of the work will be designed to show applications of the methods, as operational research techniques. Case studies will be given, and a discussion of other applications, to show the power of cluster analysis and ordination as management decision aids.

B.2 CLUSTER ANALYSIS

General

In order to reduce large masses of data into more compact and usable sets, it is often necessary to arrange items into groups. For example, in order to discuss the behaviour of human beings it would be almost impossible (and of doubtful use) to discuss the behaviour of each individual - thus we group them into classes, by various attributes - by country, age, weight, size of family, etc. - depending upon our need in a particular instance. Thus if we were concerned with human social behaviour we might group into classes such as age, sex or social class, and if we were interested in genetic characteristics we might group according to colour of hair, height or weight. In order for these or other groupings to be of use, we need the following two axioms to be observed:

1. All members of the population are assignable to a group. (A group may consist of a single member, or no members at all.)

2. All elements of a group possess a property (or properties) which all elements in other groups lack.

For instance, when we classify individuals according to their age, then the people in a group possess the property of having a certain age, which no other people possess, and also each person has an age and is hence assignable to a group. Groupings which abide by these two axioms are called dissections. The boundaries between groups are often

arbitrary, for instance if we group people according to the age groups: under 21, 21-40, 41-60, over 60, then there is no suggestion that the ages 21, 41 and 60 are anything more than a convenient division point, we might easily have chosen the groups: under 18, 18-35, 36-53, over 54.

Any set of data can be dissected, but if we try and obtain groupings with less arbitrary divisions between them, and find 'natural groupings', then the groups contain more information, and may give more of an indication of the structure of the data. Thus we place a stronger condition on our dissection to obtain groups that will give insight to the data, or at least a more meaningful dissection. This condition replaces axiom 2:

2'. We require members of each particular group to possess a degree of similarity which is not possessed by items not from the same group as each other.

This type of grouping is called <u>clustering</u>. (Note that overlapping groups whilst obeying axiom 2', do not obey axiom 2, and that any non-overlapping groups do form a dissection.) With clustering we are searching for evidence that the data is multi-modal, and thence grouping the observations. We can probably best visualize this as looking for 'swarms' of points in two dimensions. These 'swarms' may be visualized in three dimensions (but of course cannot be displayed so easily) and one can hence use the analogy for higher dimensions.

The problem may be stated more formally as follows:

Given a set B of N observations $(b_1, b_2, \ldots b_N)$ measured on a set of M attributes such that b_i is defined by the attribute profile $(a_{i1}, a_{i2}, \ldots a_{iM})$ for i=1,...N. We wish to partition B into K subsets B_k(k=1,...K) such that -

$$\bigcup_k B_k = B \qquad \text{(Axiom 1)}$$

If we wish our groups to be non-overlapping then we have

$$B_i \cap B_j = \emptyset \quad \text{(for all i,j such that } i \neq j)$$

We define a clustering function between the observation b_i and a subset β of B where β comprises the p elements b_{β_m} (m=1,...p) by:

$$S(i,\beta) = f\left(g_1(a_{i1}, a_{\beta,1}, \ldots a_{\beta_p 1}), g_2(a_{i2}, a_{\beta,2}, \ldots a_{\beta_p 2}), \ldots \right.$$
$$\left. \ldots g_M(a_{iM}, a_{\beta,M}, \ldots a_{\beta_p M}) \right)$$

where $g_j(\quad)$ is some function relating the j^{th} attribute of observation b_i with the j^{th} attributes of those $b_i \in \beta$ and f is some function combining the attribute functions.

We can thus express Axiom 2' as

$$S(i,q) > S(j,r)$$

for all i,q such that $b_i, b_q \in B_k$, all j such that $b_j \in B_L$ and all $b_r \notin B_L$.

The different choices of the various functions in the expression for S gives rise to a wide range of types of clustering method. An approach used by many methods is to base the clustering function between a point and a group on a defined relationship between a pair of points. Thus we have

$$S(i,j) = f_1\left(g_1(a_{i1}, a_{j1}), g_2(a_{i2}, a_{j2}), \ldots g_M(a_{iM}, a_{jM})\right)$$

for all $i,j \in B$

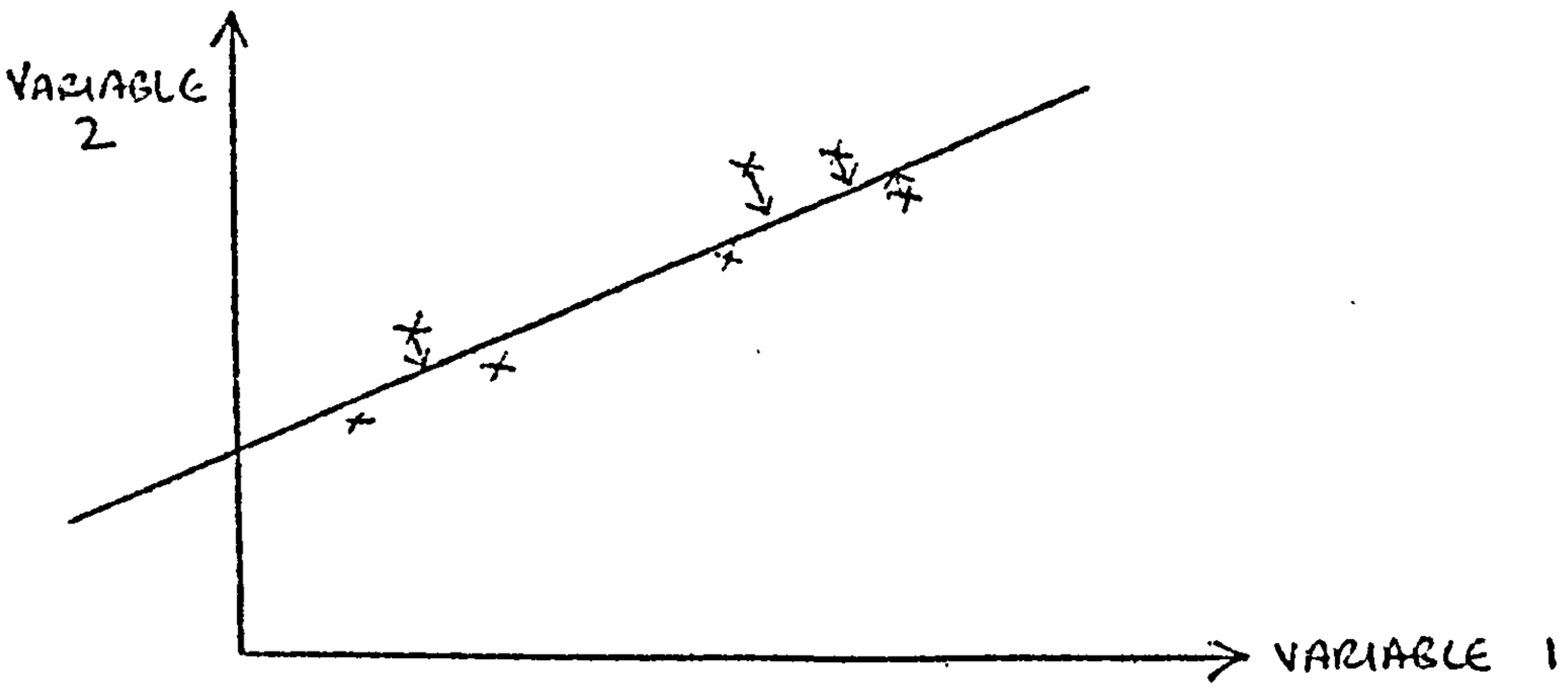
and $S(i,\beta) = f_2(s(i,\beta_1), s(i,\beta_2), \ldots s(i,\beta_p))$

The first expression is simply a similarity function and thus the procedure in a great number of methods is to define

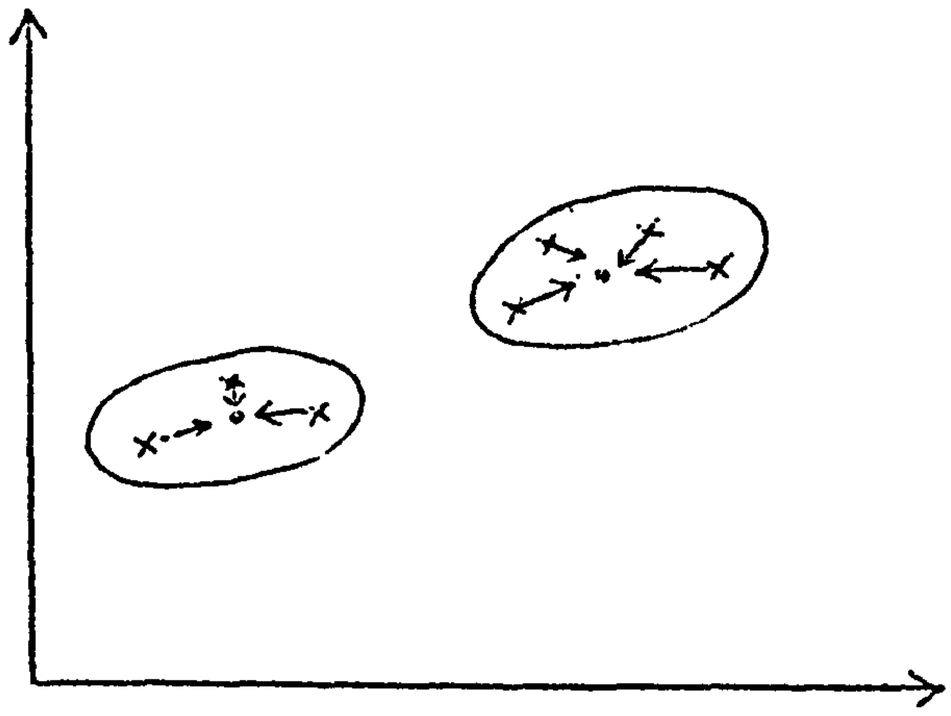
a similarity function between points, and from this to construct groups. Often the functions g_i (i=1...J) are identical or simply multiples of each other, and similarly with h_m (m=1...n) so the expressions simplify further in practice.

The relationships between principal components analysis and a certain type of clustering called least squares cluster analysis can be illustrated. Whilst p.c.a. attempts to summarize the observation vs. variables matrix by reducing the number of variables, with cluster analysis we attempt to reduce the number of observations. However, p.c.a. always gives an optimum result, which is not true for least squares classification.

E.g. If we are able in p.c.a. to reduce two-dimensional data to one dimension then graphically we are approximating a series of points to a series of linear points.



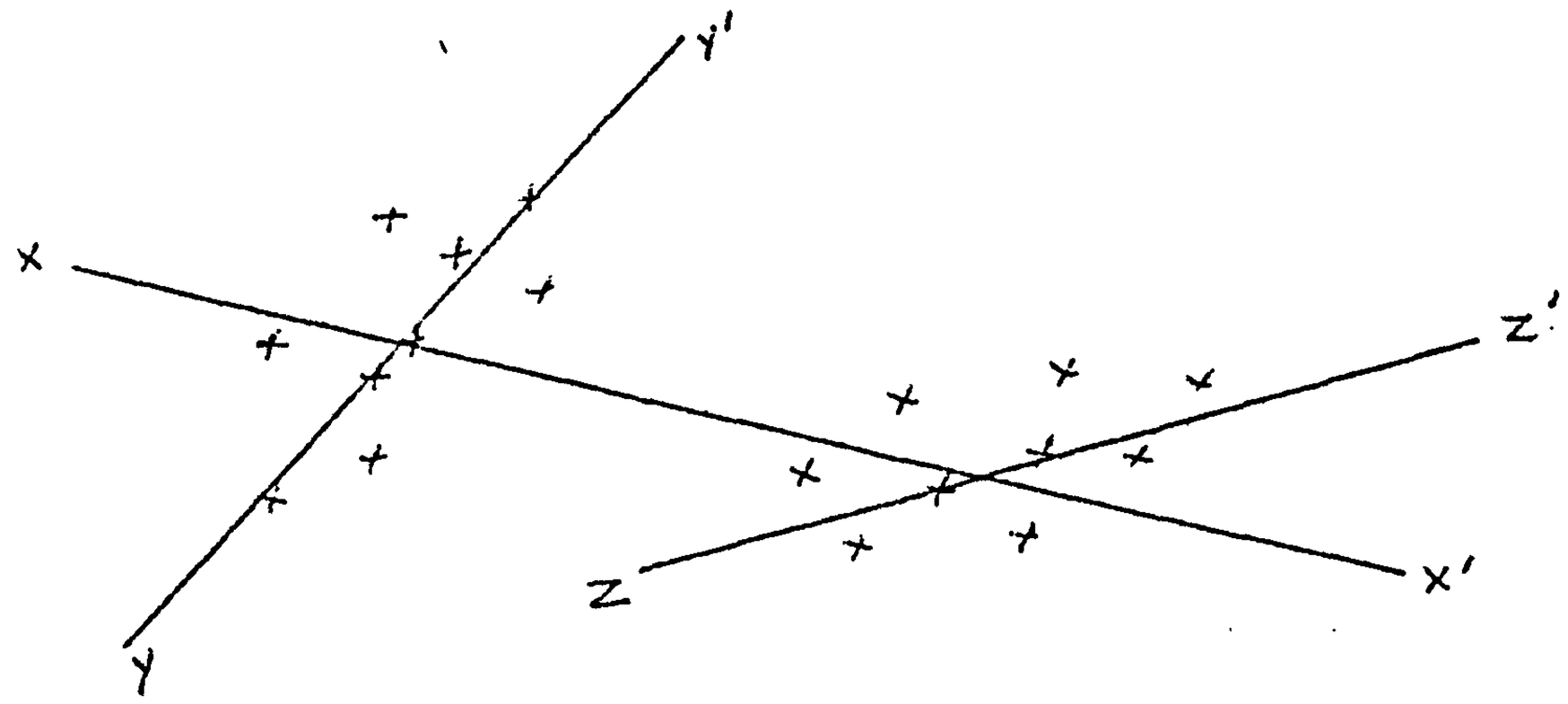
With the same set of points, if we cluster the seven observations, we obtain two groups.



With p.c.a. we have reduced the 2 variable 7 observation matrix to a 1 by 7 matrix and with cluster analysis obtained a 2 by 2 matrix.

In a particular study of a set of data one might well use more than one multivariate technique - e.g. p.c.a. and cluster analysis. However, if data has clusters present then principal components may give misleading results since each cluster may have different underlying dimensions.

E.g.



In the above diagram a principal components analysis would yield a first component of XX', whilst each cluster has its own major underlying dimension of YY' or ZZ'.

Applications

The number of fields in which cluster analysis is being used is increasing rapidly. Also in each of the fields where it has been used, its use is becoming more widespread. Some aspects of most sciences have been subjected to clustering methods in the past, and now the methods are being applied in the business and social science fields.

Some examples of the use of clustering include:

1. Palaeontology, Botanical and Zoological Sciences: the main application in this field has been in taxonomy - the grouping of plants, animals, insects, etc., into species. Previous to the use of mathematical techniques, detailed measurement of specimens (a measurement is called an operational taxonomic unit or OTU), together with the past experience of the taxonomist, produced possible groupings into species. The problem is clearly one which cluster analysis can tackle - finding homogenous groups from many measurements of specimens - and thus it is not surprising that this field was one of the first to use clustering. Indeed, some of the first numerical methods were propounded by such practitioners of the zoological sciences as Sneath, Sokal, Rohlf and Michener. One of the first major works in this field of numerical taxonomy is Sokal and Sneath (1963) which covers the whole process of taxonomy and discusses the

clustering methods then available. Earlier articles which introduced new methods are Sneath (1957) and Sokal and Michener (1958). (See also Sneath 1962, 1964a, 1964b, 1967; Rohlf and Sokal 1965, 1967; Rohlf 1965, 1967, 1970; Kubica et al 1973; and a book by Blackith and Rayment 1971.) Similarly Cheesman and Berridge (1959) attempt grouping of bacteria from chromatography results.

Another early application was in ecology, where the object of study is the relationship between different plant species, and between them and their environment. By dividing an area of land into equal sized squares (called quadrats or stands) each square can be examined and it can be noted which plants are present in each particular square. From this, by clustering methods, plants can be grouped, so that plants which tend to co-exist are in the same group. From this the effects of, say, burning off grass, or of nearby canals, on vegetation can be assessed, and the effects of one species on another. The early work in this field was carried out at about the same time as the first investigations in taxonomy and the ecology methods were then quite different. The pioneers of these methods were Lance and Williams and their co-workers, currently in Australia. Their papers include Williams and Lambert (1959), Williams and Dale (1965), Lance and Williams (1967). However, perhaps the earliest work in this particular application is Sorenson (1948) who made a study of the vegetation on Danish commons. Other methods have been introduced by Rogers and Tanimoto (1960), Fisher (1937) and Hall (1967a, b).

2. Psychology: One of the uses of clustering in this field has been to try and cluster individuals into groups on the basis of the results of psychological tests, either to group people into 'types' as in McQuitty (1957, 1964), for use in personnel administration (Ward and Hook 1963), or to allocate people to jobs, as in the armed services (see Thorndike 1953). Another use is in trying to group psychological tests into similar sets on the basis of test results, and thus to gain insight into the nature of the particular tests, as in Harman (1966). In psychology cluster analysis is used as a method of data investigation in order to seek structure in the data; this type of application is very different from the taxonomy problem in which the groupings are the required results. Thus in this type of data investigation, the clustering methods are used similarly to factor analysis, as a method for possible hypothesis generation. A typical application of this type is Miller (1969) who investigated the perceived similarity between meanings of words.

3. Earth Sciences: The problem, in geography, of defining 'regions' is one which can be partially resolved by the use of cluster analysis. The definition of a region is very similar to that of a cluster: an area, the parts of which are very similar to each other but dissimilar to the parts of other regions. An often desired property of the resultant groups is that they must form contiguous areas (there is a discussion of this in Johnston 1970). This can be achieved in two ways, either by clustering without this constraint and

observing the resultant groups, which may well turn out to be contiguous (as shown in King 1969) or by incorporating such a constraint in the clustering algorithm. Another geographic application is simply as a method of classification, as in the classification of American business districts by Berry (1967), or British towns by Andrews (1971) and Moser and Scott (1961). Richter (1969) has clustered the geographical location of firms to investigate industry grouping.

Another problem similar to that of the regionalization problem is that of political districting, where an area must be divided into equal-sized parts for parliamentary constituencies or wards. Here, apart from the constraints of contiguousness and equal size, there is normally that of compactness. Whereas geographical regions are often long thin meandering areas, it is normally considered necessary to have political regions as compact as possible. Given a completely evenly spaced population the solution is that of tesselated hexagons (although this is still unproven). See Weaver and Hess (1963), Harris (1964), Thoreson and Lüttschwager (1967), Kaiser (1966), Bunge (1966) and Garfinkel et al (1969).

Another geographical aspect in which classification methods have had some recent success is that of land use analysis which is of particular use in environmental planning. Clustering has been used to find areas (not contiguous) in cities, in which similar activities occur. For example, Goddard (1970), by analysing taxi flows in London split the

city into five regions - the West End, Westminster, Soho, The City and Bloomsbury. Alexander (1972) has divided the centre of Perth, Australia into five regions on the basis of the business activities present. In an earlier paper, Goddard (1968), grouped London into five regions also on the basis of the location of types of business activity and grouped the business activities into eighteen groups. See also Gittus (1964), Jones (1968), Dear (1969) and Golder and Yeomans (1973).

Pedology is another of the earth sciences to which clustering has been applied in order to discover soil regions. Grouping is carried out by measuring the soil composition, colour, type of stones, roots present, etc. The main exponent of modern numerical methods in soil science is Rayner (see Rayner 1965, 1966, 1969, and also Bidwell and Hole 1964 and Grigal 1969).

4. Life Sciences: Apart from the use of classification in biological taxonomy, the other main life science in which cluster analysis has been used is that of anthropology. One of the areas of study has been the evolution of the American Indian tribes - by obtaining information on the social customs and behaviour of each tribe, it is suggested that by grouping together tribes which are similar on these counts, one can trace the origins of each tribe. Research has been carried out along these lines by Kroeber (1939) and Clements (1954). Linguistics is another subject to which cluster analysis has been applied in order to detect the evolution of

particular tribes or races, either by comparing languages or dialects for linguistic similarity (see Levelt 1970). Driver (1965) in a review article discusses the progress of numerical classification and other statistical methods in anthropology up to 1963.

5. Archaeology: This field is related to anthropology, and clustering has been used in a similar way, to discover past relationships between cultures, by measuring similarities between artefacts. The use of cluster analysis has been discussed in Hodson, Sneath and Doran (1966) who compare the relationships between 30 brooches found in Iron Age cemeteries in Switzerland, as measured by numerical methods and by the intuitive groupings by four archaeologists and an anatomist, and conclude that the numerical methods came out at least as good as the best intuitive results. Another well written piece is that of Clarke (1962) who applied simple cluster analysis to early British beaker pottery. Other applications are found in the book of readings by Hodson, Kendall and Tautu (1971). (See also Von Hagen-Bordaz and Bordaz 1970.)

6. Information Retrieval: In information retrieval the user's requirements are often that he wants related information to a particular subject. This can be stated as requiring to know other similar information to a particular given piece of information, i.e. one wants to find the other members of the cluster to which the possessed information belongs. For instance one could measure the similarity between journal articles by the number of references to other articles they had in common, and produce clusters from this

data, or better, analyse articles and classify each into several descriptors (called keywords) and measure similarity by the number of descriptors in common. Since one piece of information may be of use in several different fields the emphasis in this field has been on clusters which overlap, and which, in information retrieval, have been called clumps - the subject has been discussed by Minder et al (1973). A large amount of the English work in this field has been done by Needham and Jones at the Cambridge Language Research Unit. A great deal of work has been done on the type of classification needed in information retrieval (this is covered in Jones 1971), but little has been attempted in terms of practical results; one of the methods used is explained by Needham (1965).

7. Medicine: One can consider the task of a doctor or psychiatrist as the allocation of patients to groups corresponding to diseases, according to their symptoms. Any patient with a particular disease or disorder will have symptoms similar to those of other patients with that condition. Thus cluster analysis could possibly be used for fast diagnosis, or as an aid for human diagnosis. The numerical methods are of more use in the psychiatric field where the diagnosis problem is more difficult. By investigation of data on patients with known disorders, cluster analysis may be used to find the major discriminating variables between disorders and thus diagnosis of new patients may be achieved rapidly. This use in psychiatry is explained in Kaskey et al (1962), and in several works by Lorr (1966, and Lorr et al

1963, 1965), and an overview of statistical methods in psychiatric research is given by Moran (1969). See also Strauss et al (1973), Everitt et al (1971), Pilowsky et al (1969) and Paykel (1971). Applications in diagnosis of liver diseases are given in Baron and Fraser (1968), and the use of clustering in investigating child deaths has been investigated by Carpenter (Royal Statistical Society, Multivariate Study Group Meeting, 9th October 1973).

Another application in medicine is that of clustering particular cases of a disease, in space and time, in order to detect epidemics early. This is discussed by Knox (1964), Armitage (1971) and Pike and Smith (1968).

8. Signal Detection and Pattern Recognition: In the physical sciences, there has been a large amount of work done in analysing noisy data such as in radio signals, or pictures from satellites (see Haralick and Dinstein 1971). Here cluster analysis can be used to reduce this noise. Another similar problem is in machines which are required to 'read' or interpret other visual material. With machine readers, clustering can be used to reduce input characters to 26 groups, which hopefully represent each of the letters of the alphabet. There is a large volume of work on this type of subject, which also includes work in related fields such as classification into known distributions, discriminant analysis, etc. See Sebestyn (1962a, b, 1966), Ball and Hall (1966), Rutovitz (1966), Casey and Nagy (1968), Switzer (1968), Sammon (1970), Dorofeyuk (1971), Meisel (1972) and Duda and Hart (1973).

9. Business: Because of the varied aspects of management, and the large amounts of data available, business is one of the largest areas for cluster analysis applications. The main discussion of the use of clustering methods is left until later in this work, but at this point we can outline one or two areas in which clustering is of obvious use.

In marketing, because of the abundance of data from surveys and the lack of detailed knowledge of many aspects of marketing such as buyer behaviour, advertizing effectiveness, product market, etc., many multivariate methods have been applied. Examples include: clustering towns in order to select one from each cluster as a representative set in order to test market products (see Green, Frank and Robinson 1967, Morrison 1967); clustering stores, in order to find similar stores to test product prices (see Day and Heeler 1971); and clustering different models of the same product to ensure your company is competing in each sector of the product market (see Green and Tull 1970, and Frank and Green 1968).

In investment analysis the grouping of shares into sets which behave similarly over time can enable one to build a portfolio which minimizes expected risk, i.e. one selects one investment in each group, to protect the portfolio from a particular slump in one area of the market (see King 1966, Farrar 1962 and Russell and Taylor 1968).

Other Applications

Although the major fields of current cluster analysis usage are outlined above, numerous other areas have been

investigated by these methods - a few examples are listed below:

Geology - Miller and Kahn (1962), Parks (1966), Gower (1970).

Sociology - Chabot (1950), Alexander (1963), Beum and Brundage (1950), Coleman and McRae (1960), Forsyth and Katz (1946), Laumann and Guttman (1966).

Criminology - Wilkins and McNaughton-Smith (1964).

Experimental Design - Cochran and Cox (1957), Kish (1965), Kennard and Stone (1969), Marriott (1970), Calinski (1971), Golder and Yeomans (1973).

Economic Aggregation - Skolka (1964), Fisher (1969).

DNA Classification - Ehrlich (1964), Silvestri and Hill (1964), Fitch and Margoliash (1967).

Discarding Variables in p.c.a. - Jolliffe (1972, 1973).

History

The early work on clustering was carried out in two main fields - those of psychology and of zoological classification (numerical taxonomy). Cluster analysis was first used in the great debate on the structure of human ability which followed Spearman's 1927 book 'The Abilities of Man' which brought together all the ideas of the structure of the mind's processes which had begun to germinate in the 1920's. In the ensuing discussing over the twenty years after Spearman's book, psychologists such as Thurstone and Burt used early multivariate techniques, especially factor analysis in order to try and add weight to their own favourite theories. Some of the researches involved the use of factor analysis as a

clustering method, trying to group intelligence tests according to the precise ability they were testing, such as mathematical or verbal ability.

Classification methods in their own right were first used in the late 1930's (see Holtzinger and Harman 1938) and it was about then that the term cluster analysis was first used. The first major published work on method was Tryon (1939). These early methods were simple and fairly rough since they had to be computed manually - they normally involved simple investigations of correlation matrices. Because of this limitation on methods, the subject advanced very slowly after the publication of these few hand methods, and it was not until the advent of the high speed computers in the 1950's that the subject was to advance to any great extent.

The first breakthroughs of the computer age were in zoology, where the first computer-based methods began to appear in the late 1950's (see Sneath 1957, Sokal and Michener 1958 and Michener and Sokal 1957). These methods were used as a simulation of the way in which the taxonomist was able to group species, and resulted from the realization of the fact that taxonomy was a simple clustering problem. With the publication of the first few methods and with larger and faster computers, larger problems could be solved by the existing methods, and more complex methods were introduced which could only be implemented by use of the new computers. At about the same time as the use of clustering in taxonomy,

other fields were also experimenting with other forms of clustering, especially in ecology and information retrieval, and also workers in psychology had continued their early work. Because of this growth in parallel, with each stream of investigation having little knowledge of the work in other streams, came a great deal of duplication of work, and differing terminology. However, as these areas of interest became aware of each other, methods discovered in one subject were used in others and cross-fertilization expanded the use of cluster analysis. Thus cluster analysis became a subject in its own right and was another mathematical tool to be used in many kinds of research.

In the last five to ten years the subject has expanded rapidly in two directions - new applications and new methods, and along with this expansion has been a related increase in the theoretical difficulties. One of the more unique and troubling properties of this new technique of cluster analysis is that over recent years the number of methods in use has increased rapidly, and there have been few attempts to rationalize the use of particular methods. If there has been one factor which has inhibited the growth of the use of clustering, it is the fact that there exists no one method which has been proved 'best'.

In the last six or seven years, operational research scientists have begun to look at cluster analysis and other multivariate methods and to use these techniques as methods of data analysis in the management field. Marketing was the

first area to be subjected to these methods, being an area in which data is abundant and knowledge about that data is scarce, and the first published works appeared in 1966 and 1967. Another management function currently using clustering is the personnel field where studies in psychology are of relevance to selection and placement.

Methodology

One definition of cluster analysis is that we are searching for 'a degree of similarity between members from the same group, which is not possessed by members not both from the same group' and another definition is that we obtain clusters 'which are in some sense optimal'. The exact meaning of these definitions is one of the central problems in cluster analysis - what do we mean by 'a degree of similarity', what do we mean by 'similarity', what does 'in some sense optimal' mean, and in what sense? The answers to these questions will vary with the particular application and thus the method used will be dependent on the application. In general, however, all cluster methods reach their conclusions by passing through four stages - Measurement, Similarity, Method and Analysis of Results. Measurement is the obtaining of the required data and the process of putting this in the required form for our investigations, Similarity is the production of a measure of similarity by which we will be able to compare the objects under study, Method is the process in which the similarity measure is used on the data and clusters are obtained, and Analysis of Results is the

testing and consideration of the clusters found. We shall examine these four aspects in detail, in the above order.

Measurement

In order to judge the extent of the similarity between a pair of objects, we must measure each object on various attributes such as weight, temperature, colour, etc. In fact when considering the similarities between all pairs from a set of objects we must measure each object on all attributes that may have some bearing on the similarity between objects from the point of view of our specific investigation. The first problem is thus deciding which variables are pertinent and which are not. If we exclude variables that are relevant then we may invalidate our results, and if we introduce unnecessary variables, we at best cloud our results, and at worst arrive at meaningless clusters derived in part from the irrelevant variables. The particular clustering method used on the data could possibly help minimize the risk of large errors from the selection of the wrong set of variables, but the basic problem is one of experimental design. Such problems are best overcome by detailed knowledge of the objects under study, from experience with similar objects and experience with clustering techniques.

Having decided which variables to measure, the next problem is how to measure them. We may conveniently define measurement as the process of representing properties by numbers. There are an infinite number of types of scale by

which one could measure, but five well-known types have been selected as examples. These are, in order of 'exactness':

1. Nominal Scale: This is the giving of numbers to objects to represent them as being either the same or different in a particular property, without any ordering of the objects being possible. For example, if objects from a particular set could be one of the three colours - blue, brown yellow, then we could not order the objects on this property, but we could label all blue objects with the number 1, brown with 2 and yellow with 3, to show they were different. Another good example is the numbers on the shirts of football players, all players with the number 9 are centre forwards but no more similar to players numbered 8 than those numbered 6.

2. Ordinal Scale: This type of scale is a ranking of objects according to their degree of possession of a property. For example, one could order a set of plants by the darkness of their leaves by simple comparisons between them.

3. Interval Scale: An interval scale is one in which distances between points on the scale can be meaningfully compared. For example, if we measure the temperature on a set of days, we have a meaningful difference between any two days' temperatures.

4. Ratio Scales: These scales are interval scales with a zero on the scale being a zero of the property. For example, measures such as height and weight, because they have a zero which corresponds to no height or no weight, we are able to say that 2 pounds is twice as heavy as 1 pound, whereas we cannot say that $20^{o}C$ is twice as hot as $10^{o}C$.

5. Absolute Scales: Absolute scales have one unique scale on which one can measure, such as counting the number of legs an animal has, or the number of leaves a plant

possesses. Ratio scales such as height can be measured meaningfully on many different scales such as feet or metres.

Note that each of the above scales is a special case of all those scales preceding it. With interval scales, the relative positions of objects on that scale are fixed and thus similarities are preserved. Thus data on absolute scales (number of toes on an animal), ratio scales (height of a tree), and interval scales (boiling point of a liquid) are readily usable for cluster analysis because most similarities are invariant under any linear transformation. However, nominal and ordinal scales cause difficulties, since they do not fix objects to a relative positional point. With these types of scale we must either approximate them to an interval scale or find other ways round the problem - we discuss this below.

Ordinal Scales - the individuals may all be ordered or they may be in ordered groups (for example a person may have to rate his opinion of a radio programme into one of the categories excellent/good/fair/poor/bad which would form an ordinal scale of groups of programmes). Ordinal values could be used as if they were on interval scales, but by grading the objects subjectively one could arrive at a better approximation to an interval scale (in the above example the person would be asked to give each programme a mark out of fifty according to how good he considered it to be). Alternatively, once the measurement on the ordinal scale was

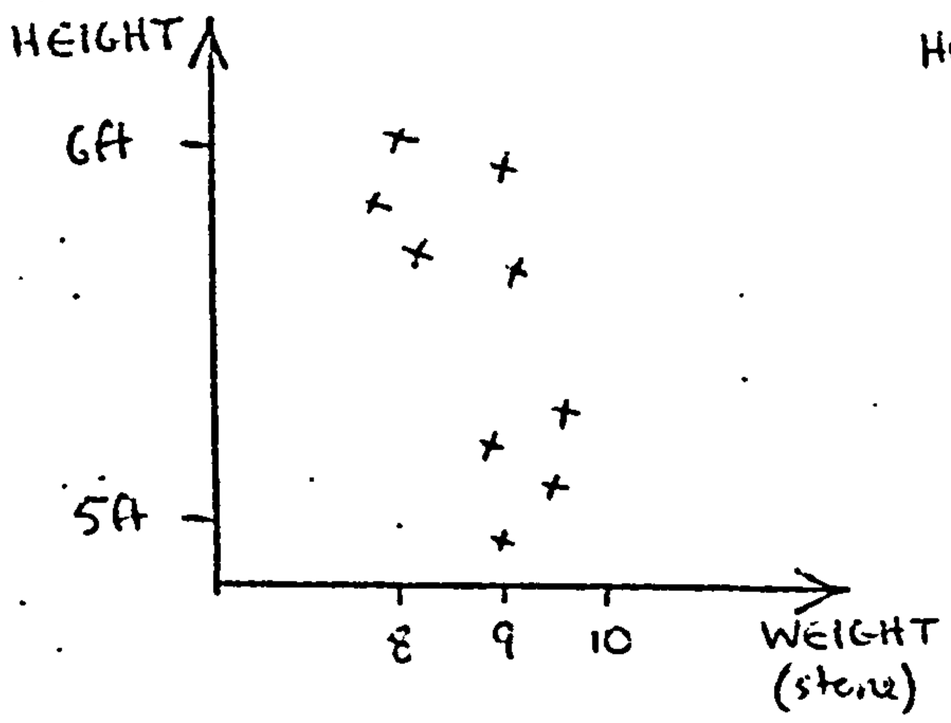
made, further investigation may help one decide how 'close' the objects are on that scale, and produce a subjective scale (i.e. instead of approximating to an interval scale by using 1, 2, 3, 4, etc. one would consider how close 1 was to 2, 2 was to 3, etc. and perhaps use 1, 3, 4, 7, etc. as a better approximation). A further alternative would be not to try and measure that particular variable but to try and measure a related variable, which would give the same information, but which could be measured on an interval scale (for example, depending on the particular application, perhaps the number of listeners to each particular radio programme, or the number of letters about the programme, etc., could be used in our previous example).

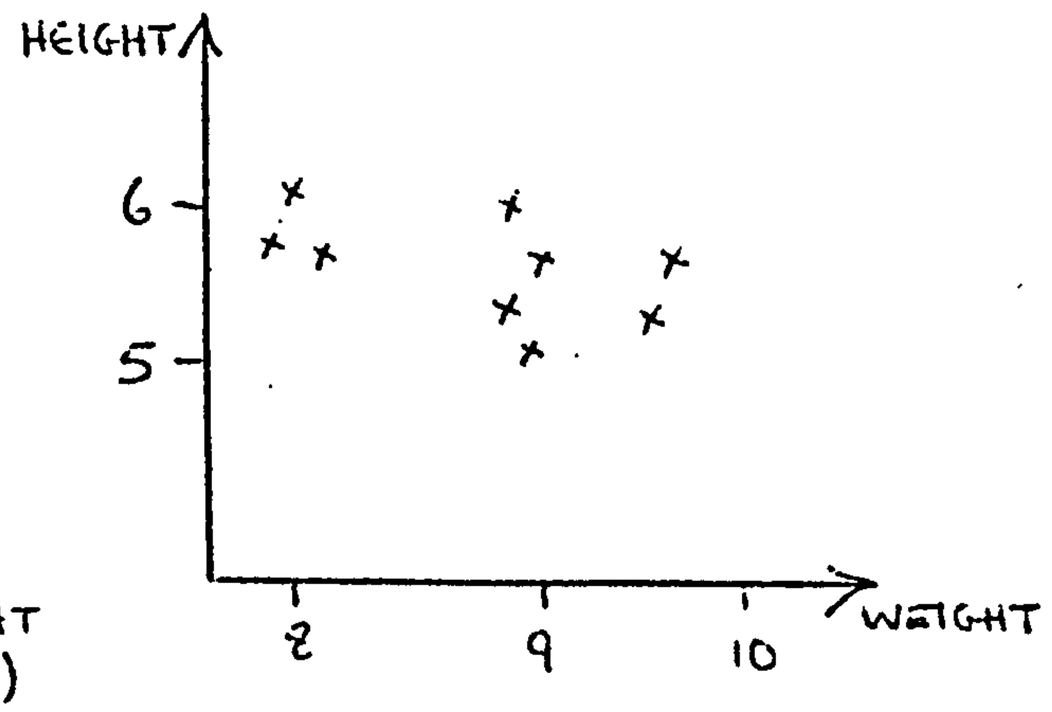
Nominal Scales - since these scales do not give an ordering of objects measured on them they are even more difficult to convert to interval scales. The problem is of putting such variables as colours for example into an order. If we have variables which are dichotomous, i.e. objects either possess or do not possess a particular property (this is called <u>binary</u> data) then these are on interval scales, so one solution is to split our nominal data into dichotomous data. For example birds of different colours could be measured on the nominal scale black = 1, blue = 2, brown = 3, etc., this could be converted into several binary variables such as black/not black, blue/not blue, brown/not brown, etc. Alternatively, we could try and measure alternative related variables. For example we could measure the birds on brightness of plumage.

If we have doubt about whether certain variables should be included in the study, or about the way we have scaled or measured them, then the cluster analysis should be performed both with and without these variables, in order to determine their effect on the results of the analysis. A possible future aid in the scaling of nominal and ordinal data is the current psychological work on for example the scaling of colours using ordination.

We can divide scales into three types - <u>binary</u> data (sometimes termed dichotomous, or two-state data) which we have previously referred to, <u>multi-state</u> data which is data that can possess only a finite number of numerical values, and <u>continuous</u> data which can hold any value in a particular (possibly infinite) range. Note: the above problem of converting non-interval scales to interval scales is often discussed as the problem of trying to convert multi-state data into either binary or continuous data. This is not quite equivalent since some multi-state data is on an absolute scale and thus it is not necessary to transform the data in any way.

Once we have our data in the form where we have selected all relevant variables and measured each object on a suitable scale then we have the problem of the correct scaling between the variables. This is a related, but often more complex problem. This problem can best be illustrated by an example; consider the following pair of diagrams, representing the same set of weight, height vectors:





Since height and weight are different scales and cannot be equated, we could equally well choose to represent our data in either of the two cases above. In both cases we have two clearly defined clusters, but the membership of the clusters is not constant. M. Kendall (1966) has suggested that rank correlations should be used in discriminant analysis as a distribution-free method, but the use of rank correlations in clustering is discouraged because any gaps between clusters would be reduced. However this could be a useful check for misclassification after clusters have been extracted from the data. Since one cannot equate measures such as height and weight, the normal procedure in order to reduce this problem is either to normalize each variable to unit variance, or to use a method which is designed to find groups of any shape (see later).

Another problem is that if we have a mixture of discrete and continuous variables then data may tend to group itself according to the discrete variables. We suggest this could be overcome by converting the discrete variables to continuous ones by randomizing and performing several analyses with different random numbers.

It may be that we consider some variables to be more important than others in our analysis and we wish to give these more weight by scaling them up and thus produce clusters more accurately and efficiently. Sneath has proposed (1962) (after Adanson - a famous eighteenth century botanist), that all variables should be given equal weight. He gives as an example (see Sneath 1967) of false weighting (of observations in this case), that if one were calculating an average height of British adults, one would not, for example, count policemen and bishops twice because they are 'more reliable'. However consider the same example, and suppose we had obtained the height measurement of several hundred people, but had only been able to measure half as many women as men, then we would be perfectly correct in counting women twice to obtain a balanced sample. Certainly if little is known of the relative importance of the variables then any weighting is dangerous.

The question of weighting is related to the problem of correlated variables - if several variables in a study are correlated, then an implicit extra weighting is given to the variables. For example, if human body dimensions are studied, then measurements of thigh length, leg length, overall height are all correlated as measures of tallness. This correlation could be corrected for by use of certain similarity measures (e.g. Mahalanobis D^2), or by using principal components analysis or factor analysis to determine the underlying dimensions of the variables. However a part of the correlation present can be due to the

clusters in the data. For example if we wished to try and cluster apples and oranges into two groups, and we measured the diameter and circumference of each fruit, then these variables would be highly correlated because of approximate sphericity, whereas if we measured the fruits on the binary variables: segmented (0) or not (1), coloured orange (0) or not (1), smooth skinned (0) or not (1), then we would again have high correlations between the variables, but this would not mean that the variables were all similar measures of an underlying dimension, but rather that they are discriminating features. The possible ill-effects of correlated variables can perhaps best be eliminated by careful selection of initial variables, and by inspection of the correlation matrix of the variables. Another useful aid in the consideration of the input data is to print histograms of variables.

A recent problem is in trying to adapt current methods to n-dimensional matrices (where $n > 2$). For example we might have a series of data on several business firms, such as turnover, sales, etc., for a set of years. This problem is as yet largely unresolved and the current method of procedure is to take several two-dimensional matrices from the larger one. A similar but more complex problem arises in pedology where several places are used from which to take soil samples, and at several depths. These samples are analysed as to chemical composition, etc., and so a three-dimensional matrix of places against depths against

composition variables. The problem here is further complicated by the fact that a sample taken at a depth of 2 feet may correspond to a sample taken elsewhere at 4 feet, because of the shift of the land. This is related to a biological concept called homology which is concerned with the changing physical form of animals in evolution, and thus in comparing animals one has to consider how similar they are, taking into account the fact that parts may be in different places or new parts may have grown or old ones disappeared.

The measurement phase of cluster analysis, and indeed of most other statistical methods, is of extreme importance. A method is only as good as the data it uses. More emphasis should be placed on reliable information - from data or by some form of cross-validation. It is hoped that, as this thesis is concerned more with method than data, the reader will not forget the importance of these preliminary stages.

Having considered the major problems in proceeding from the initial stage of having a selection of objects to be clustered to the stage where data has been gathered, investigated and processed into the desired form, we can continue to consider the second phase of cluster analysis - obtaining similarities.

Similarity

The usual step from the data matrix is to produce a similarity matrix, since clusters are defined as groups of 'similar' objects. It is, however, possible to obtain

similarities data directly from the objects under study by, for instance, rating each pair of objects subjectively on overall similarity. It is also possible to proceed directly from the data matrix to the method. Since any measure of similarity between two objects reduces two vectors to a single number, information is lost in the process of replacing the data matrix by the similarity matrix, and thus the choice of similarity measure is of some importance. Since ordination measures also rely to a great extent on the choice of similarity measure we will leave the consideration of the various measures to Section B.4.

Any group of objects measured on a set of attributes can be considered as a set of points in space - each dimension being one of the dimensions of that space. Any objects which are 'similar' will have 'similar' vectors of attributes and thus will be close together in this space. Thus distance measures are types of similarity measures (strictly speaking they are dissimilarity measures, since the smaller the distance between points, the greater their similarity).

The types of measures depend to some extent on the data in the data matrix. There are four main categories:

(a) binary data;

(b) multi-state nominal;

(c) multi-state ordinal;

(d) continuous data.

The actual choice of measure depends somewhat on the choice of method because some methods work better with some measures than others, and some methods are based on the properties of particular measures.

We leave further discussion of similarities to Section B.5 and proceed to the third stage of cluster analysis - the method.

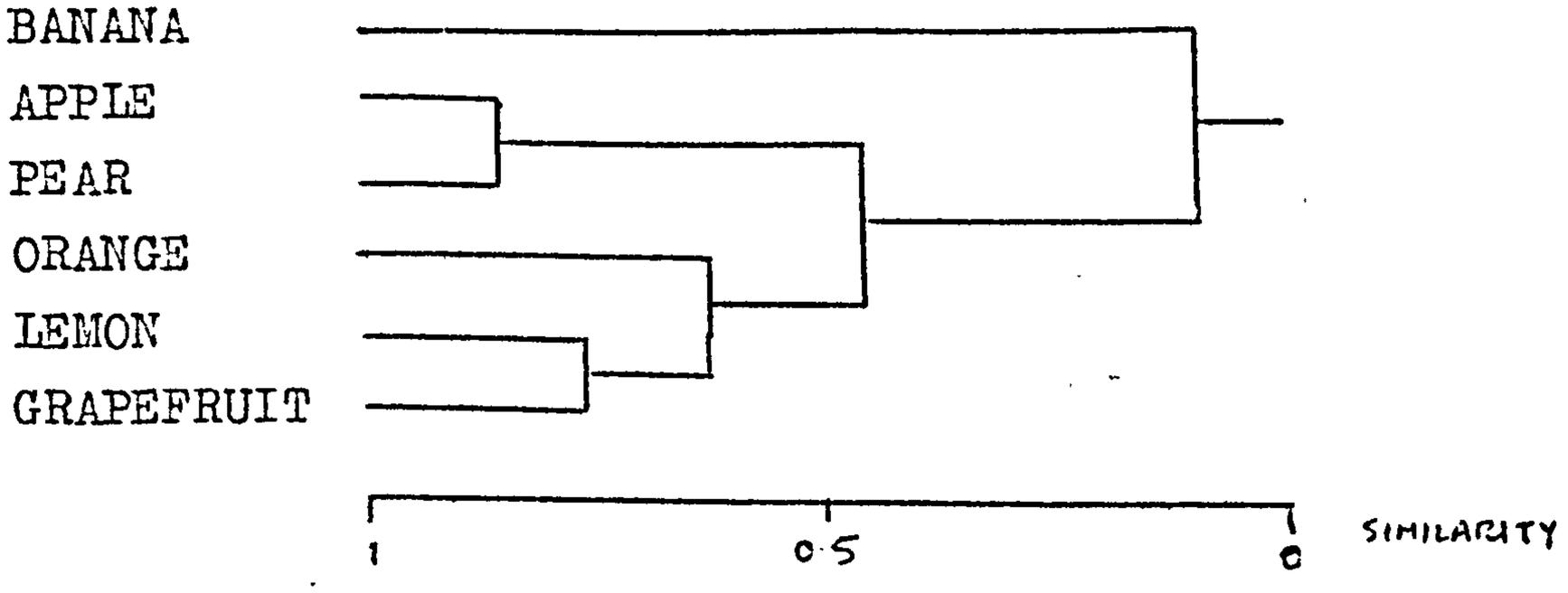
Method

The methods of clustering are numerous, and apart from differences within the methods themselves, external considerations such as the type of data used, type of similarity used and the type of output required also produce differences in method.

The type of data in the data matrix may restrict the choice of method, for example special methods have been developed for binary data, and with some methods the problem of standardizing variables is more important than with other methods. The similarity measure used may also restrict the methods which can be used, since some methods are not compatible with all measures, indeed some are based on one in particular. Most methods may be adapted to deal with either similarities or dissimilarities data simply by reversals of sign, or by other simple linear transformations. Storage space in the computer can often be reduced by not storing the whole similarity matrix but by computing similarities, when required, from the stored data matrix. This can, however, increase the calculation time and is only advantageous when

the number of variables is less than the number of observations, although the number of variables can be reduced by the use of principal component analysis. The output required may also reduce the possible methods that can be used - for example output may be desired to be in a hierarchical form, especially in taxonomic investigations where animals or plants are grouped into genus then species and then sub-species. Alternatively overlapping groups might be required.

Hierarchical methods produce a linking of objects which forms a tree diagram called a dendrogram. This has form:



The similarity between objects is shown by the position on the scale where objects link in the dendrogram. The same dendrogram can be drawn in several ways - for example if Apple and Pear are reversed then the trees are equivalent, similarly Banana could be at the end of the list and still linked at the same similarity to the other points. In fact the tree is perhaps best visualized as a mobile, where inter-object similarity can only be measured along the tree. Note that one of the properties of a dendrogram is that the similarity values decrease as the points merge.

The nature of the specific data used may cause problems for some methods. Outliers (an observation which is very different from all other observations) may cause difficulties because they can seriously distort the similarity matrix, and sometimes the method. Once any outliers are found, which is normally fairly simple (histograms of variables may help detect them), they should be eliminated and the analysis repeated without them. The shape of any clusters present will have a bearing on the accuracy of the method - whilst most methods are designed so they can detect hyperspherical clusters, some are not designed to find clusters of more elongated shapes. Also, if clusters are of unequal sizes, large clusters may tend to 'swallow up' nearby clusters.

The procedures of the different methods vary tremendously and explanation of particular methods is contained in Section C.2, but general comments may be made.

Most methods are performed by mathematical algorithms. A large number of these are executed in a hierarchical manner, either by gradually combining objects into groups (agglomerative algorithms), or by splitting the set of all objects into smaller groups until all groups contain one object (divisive algorithms). Note that although an algorithm may proceed by hierarchical agglomeration or division this does not necessarily imply that a dendrogram can be produced from it. Since most methods use algorithms and specific methods tend always to be performed by very similar algorithms, there is sometimes confusion between properties of the method and properties of the algorithm

used. Thus some methods have been erroneously referred to as divisive or agglomerative methods (see Lance and Williams 1966). This has been pointed out by Jardine (1970) who has given examples of methods which may be performed by various types of algorithm.

Methods which produce dendrograms normally proceed by agglomerative algorithms for computational convenience, and the main difference between such methods is the criteria used in deciding which existing groups are to be amalgamated at any stage. Similarly with divisive algorithms criteria by which groups are split into smaller groups are the main difference between methods. Another type of method which can be performed by agglomerative or divisive algorithms (but not necessarily), is the iterative relocation methods. These are different from the hierarchic type in that they allow objects to change groups. For example, iterative relocation methods performed by agglomerative algorithms begin with each object in a group on its own then gradually join groups until all objects are in one group, and at every stage when a pair of groups have joined, each object, or set of objects are examined to see whether they would be better placed in a different group and if so they are moved to that group. Thus with iterative relocation not only must a criterion be used in amalgamating groups but also to decide if objects should change groups.

In some applications such as information retrieval, where a particular piece of information may be related to two

different fields of interest, and thus one would like it to be grouped with information on each field, overlapping groups are necessitated. These methods are at present very slow and normally restrict the degree of overlap possible.

Another major part of any cluster method is to decide whether clusters are present or not, and if so, to determine how many. This question is basically that of how 'good' does a group have to be to be called a cluster. There is a marked lack of significance tests or published studies with methods run with random data. One of the difficulties is defining our null hypothesis - if we use random data should it be uniform through space, from a single group based on a unimodal distribution, and if so what type of distribution. Many methods, especially hierarchic ones output the value of an objective function at each stage of the clustering and clustering is indicated by a large jump in this function at any stage.

The process used in choosing a particular method for a given study is initially that of narrowing the possibilities by the type of input and output wanted, those available, those capable of handling the size of data, and those which will be executed fast enough. In order to proceed from this short-list one must either survey the sparse information on relative merits, or do preliminary selection tests on data similar to that of the proposed study. One useful aid is to use more than one method to compare results. To discuss analysis of results we move on to our fourth and final stage of clustering.

Analysis of Results

Because of the shortage of significance tests for cluster results, the analysis of results is a very important stage. In order to test the validity of results it is advantageous to be able to check with alternative data. If alternative data is not available the present sample could be divided into two parts and the analysis performed separately on each part. One of the most important guides in considering results is knowledge of the original objects, and using personal judgment to assess the results. It is helpful to output the similarity matrix, sorted so that members of the same cluster are adjacent. In order to test the rigidity of the clusters, sensitivity analysis can be performed - adding small random elements to the data matrix and re-analysing.

Assessing the characteristics of each cluster is, at present, more a job for the analyst than for the method, although the centroid of each cluster and some measure of its dispersion can easily be output from the computer program.

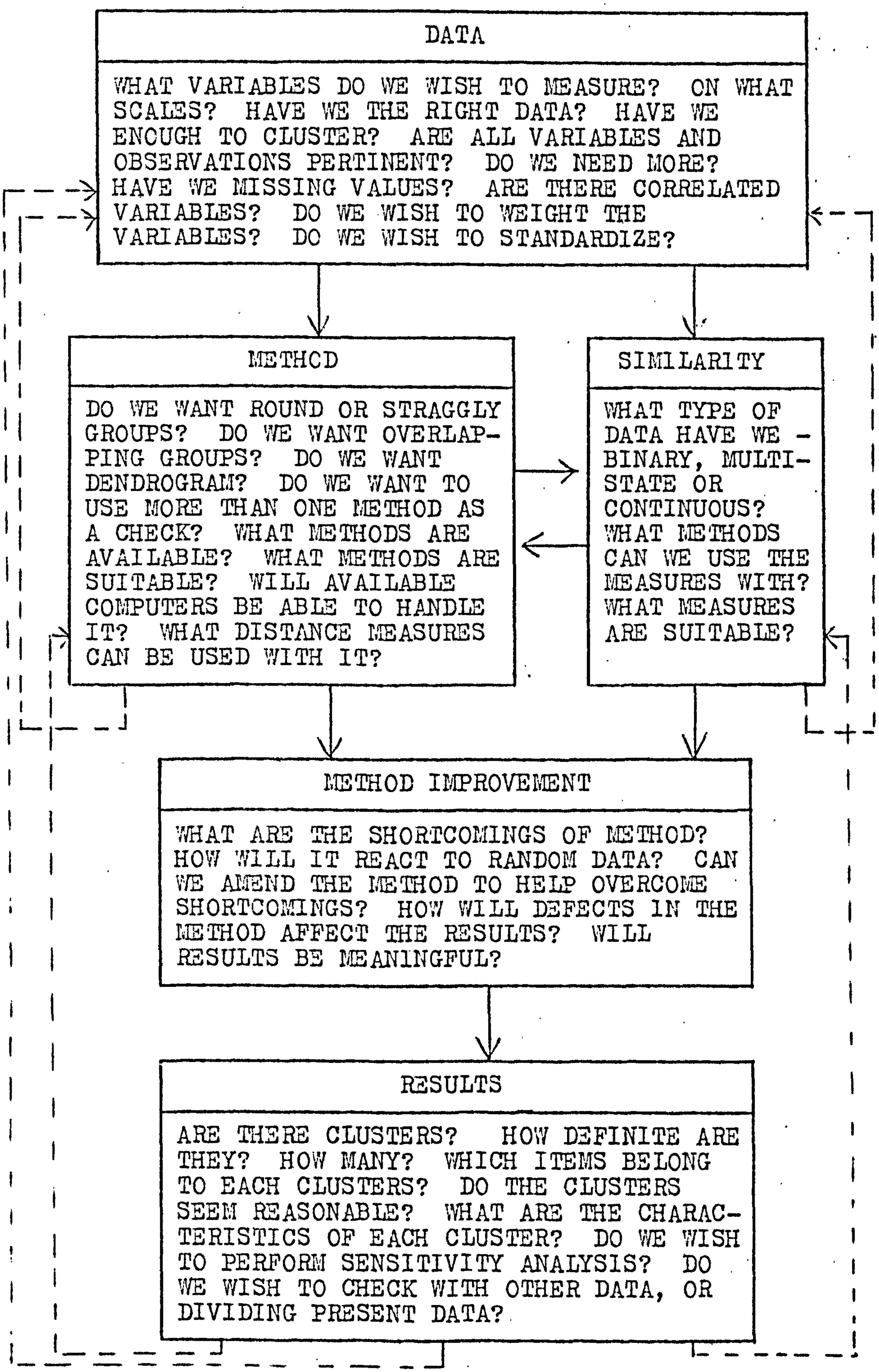
The process of drawing conclusions is neatly summed up by Herne (1973):

> "...belief in an interpretation of a numerical analysis grows diffidently and slowly only when data sample after data sample lead repeatedly to practically the same conclusions. Confirmation, confirmation, confirmation, ..., and then faint belief".

This may be overstating the case, since at any stage of our investigation we will have a 'best estimate' of the structure of the data, which may be better than no conclusions at all, but serves as a useful caution to all data analysts.

General Procedure

Once we have decided to use cluster analysis in our investigation of a certain set of objects we must, in general, follow the four stages outlined above; these can be explained diagrammatically as follows:



Main feedbacks are shown as dotted lines.

Programming

As with most multivariate methods, the major use of cluster analysis is to reduce large volumes of data to a manageable and perhaps more meaningful set. Thus it is important that a method should be able to handle large numbers of observations and variables, and thus programming considerations become important.

If the method requires storage of the full similarity matrix then storage is of the order (N^2). However in most cases only the lower (or upper) off-diagonal elements need be stored thus requiring storage order ($\frac{1}{2}N(N-1)$).

However if M is small it may be possible to adapt the program to calculate similarities as required from a stored data matrix (and as mentioned previously M can be effectively reduced by use of p.c.a.). Thus storage of order (NM) is needed. This can however increase calculation time.

One or two of the simplest methods which are in existence (e.g. single linkage, see page 133) can be programmed so that storage of the order (N) is necessary. However these methods are only elementary and are not so exact as the more complex methods. However this can increase the problem size which can be attempted by a large amount.

For example, if storage of 20K is available:

	N	M
STORE N^2	141	unlimited
STORE $\frac{1}{2}N(N-1)$	200	unlimited
STORE NM	200	100
	300	66
	400	50
	500	40
STORE N	20,000	unlimited

External storage can of course increase the problem size which can be tackled by a particular computer but this increases execution time.

One approach to clustering large populations is to use a simple method which requires storage of order (N), to reduce the size of clusters to a size which can be tackled by a better method which can solve problems of order ($\frac{1}{2}N^2$) or (NM).

An alternative approach is that of Ross (1970) who has discussed the possibility of analysing a set of groups of observations, and merging the results, or analysing a reference set and introducing other observations one by one. Ross obtained encouraging results using the merging method.

Programs which have been published include Wishart (1971), Veldman (1967), Bonham-Carter (1967), Mather (1969), Sibson (1973), Sparks (1973). Other related programs, to output various clustering representations are published in Ling (1971) and McCammon (1970) (see McCammon 1968).

B.3 ORDINATION

General

A special set of techniques which can be used in similar instances to cluster analysis in order to examine data structure, is the set of so-called ordination methods. These techniques, although having differing approaches, all attempt to produce a mapping of observations (or objects) in a space of low dimensionality, in order that the structure of the data may be visually inspected. This new configuration can be obtained from one of two types of data: data from a higher dimensioned Euclidean space, or data from non-metric data (which as it stands cannot be shown as a Euclidean configuration of points). The method can be used on data sets of observations which are known to have a certain underlying number of dimensions, or if a certain dimensionality is required in a specific case. The importance of the representation of data in ordinate form is summed up well by Gnadesikan and Wilk (1969), they say:

> "One of the most important strategies of data analysis is, and always has been, graphical presentation and pictorialization".

Factor analysis and principal component analysis can both be used as ordination methods, and in fact, they were the first ordination methods to be introduced. Other ordination methods may be used similarly to factor analysis, and indeed some are termed as 'non-metric factor analyses'. In data structure investigation, ordination has a wider purpose than the search for underlying structure of variables

in factor analysis, for example, if clusters exist then these may be seen by inspection of the final ordination configuration.

Ordination methods, as well as producing low dimensioned mappings of points, also produce information from which the best number of dimensions to use, in a particular investigation, can be determined. This is normally facilitated by use of a measure of how much the original data has been changed to reduce the dimensionality. This measure is calculated for several different dimensions. From a graph of this measure against dimensions, the 'best' or lowest meaningful dimensionality may be estimated by inspection. This measure is sometimes called the <u>stress</u> of the configuration.

The methods may be divided into two distinct types: metric and non-metric. Non-metric methods are those designed for use with non-metric input data, often from a rank order distance matrix. These methods are normally called the multidimensional scaling methods, although this term is sometimes in fact applied to ordination methods of any type. The non-metric methods seek only to preserve the rank order of the original distance matrix, in a space of given dimensionality. The most well-known method of this type is the method of Shepard (1962a, b) which was formalized by Kruskal (1964a, b). The metric methods of ordination which include factor analysis and p.c.a. are based on input data which is metric, but which has more dimensions than required.

Applications

Ordination methods as such are a recent development and nearly all the progress in this area has been achieved in the last decade. This is possibly due to the large amount of interest aroused by Shepard's paper in 1962. The areas of application, as yet, are mainly those in which cluster analysis has been used with some success. However, as the subject is one of the newest multidimensional methods, the practical applications have not been numerous, although they are increasing rapidly. At present, applications have been suggested or attempted in the following fields:

1. Psychology: Non-metric methods can be said to have been founded in the area of psychological scaling. The early investigations were normally based on methods which could be executed manually, and hence have been partly superseded by more recent computer methods. An excellent overview of the pre-Shepard psychology work is contained in Coombs (1964). Since Shepard's innovative paper, many areas of psychology have been subjected to multidimensional scaling. Ordinations have been made of people's perceptions of colours (Shepard 1966, Doehlert 1968), Morse code (Shepard 1963), facial expressions (Abelson and Sermat 1962), social attitudes (Messick 1956) and job status (Burton 1972).

2. Ecology: The use of ordination in ecology has largely arisen from the use of heuristic manual methods, such as that of Bray and Curtis (1957), and the use of factor analysis. An overview of early references is contained in their paper. More recent studies have been based on more mathematical methods, examples include Austin and Orloci (1966), Bannister (1968) and Anderson (1971).

3. Archaeology: In this field the innovative use of ordination has stemmed mainly from the interest in such techniques of D. Kendall and Hodson. Ordination methods are seen in this area to be a useful way of gaining more information from incomplete data. Examples of the use of the techniques are very wide, see several of the papers in Hodson, Kendall and Tantu (1971), Hodson, Sneath and Doran (1966) and Tobler and Wineburg (1971).

4. Marketing: Here ordination was first used in the form of factor analysis, but the majority of work has been accomplished recently with non-metric methods. A lot of research has been done in trying to produce an ordination of products (called a <u>product-space</u>), from which relationships between products may be assessed. For examples of this application, and others, the works of Green are of particular note, and of abundance (see Green and Tull 1970, Green, Frank and Robinson 1967, Green and Rao, 1971, 1972, Green and Carmone 1969, 1970). A discussion of ordination in marketing is contained in Neidell (1969).

Applications have been suggested in other specializations but, as yet, these have been isolated exploratory papers, for example Anderson (1971b) discusses the use of ordination in geology, Thompson and Woodbury (1970) consider medicine and Green and Maheshwari (1969) have used non-metric methods in investment analysis. Soils have been subjected to ordination by Bidwell and Hole (1964), and a sociology study on the existence of social classes is contained in Laumann (1966).

The initial procedural steps of ordination are very similar to those in cluster studies. In both cases the process of obtaining data input to the methods, normally in the form of a distance matrix, is a very important stage. However the method stage is very different, because of the less concise structure which one is investigating, and because of the differing aims of each method. The analysis of results is also very different from that in cluster analysis - the result of an ordination requires much more user interaction to abstract its meaning, and for this reason it is suggested that its use will be best performed by those with knowledge of the underlying concept of the method, and with detailed background knowledge of the data under study.

Optimization

Some of the techniques of ordination involve the minimization of non-linear functions, normally quadratic functions. One of the most well-known methods to solve such problems is the search technique of the method of steepest descent. This method begins with an initial set of values for the variables, which can be either random or estimations, and seeks to improve upon the value these give to the objective function. The direction of maximum improvement in this function from this starting point is simply given by the gradient of the objective function at this point.

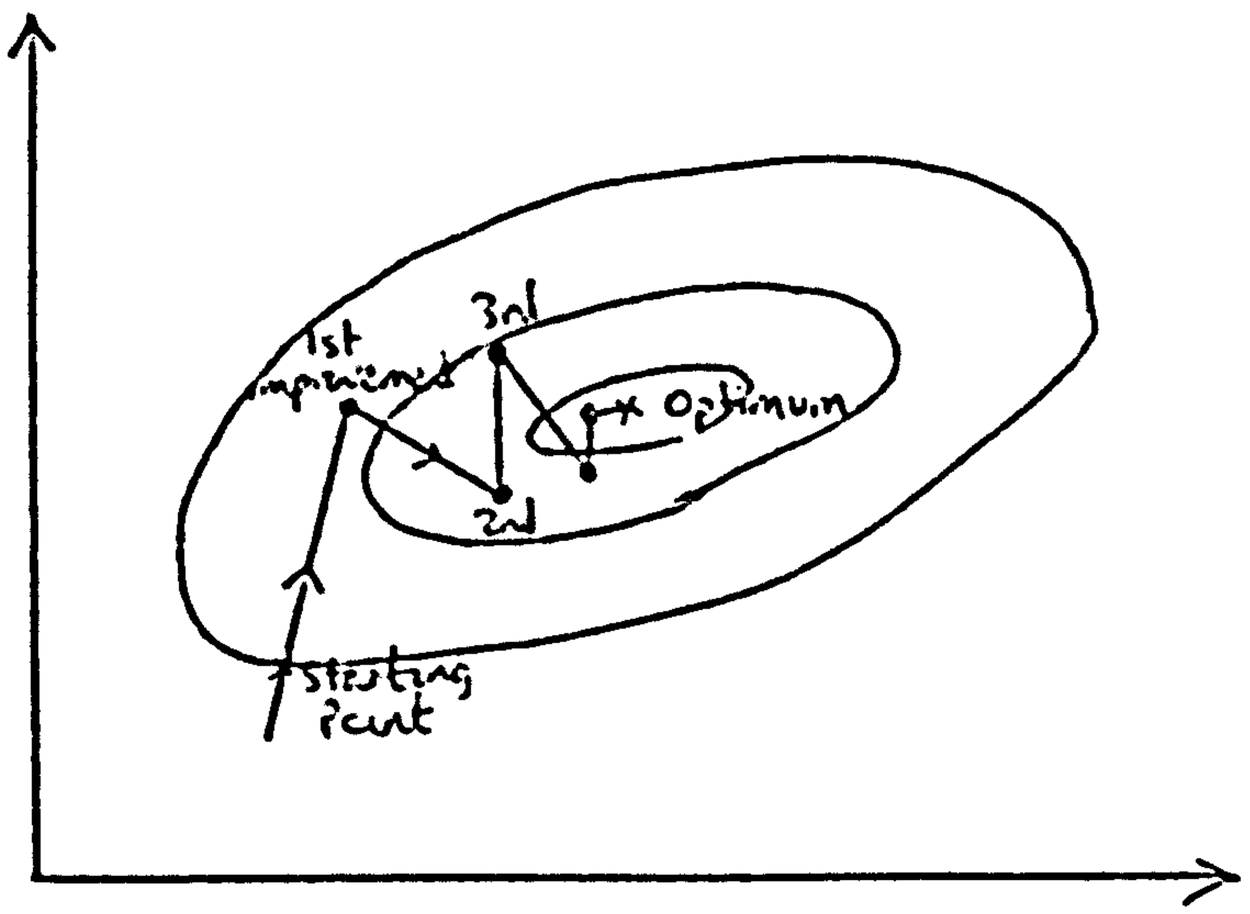
This if our objective function is $f_p(x_i)$ and our starting point is given by the values a_i of our variables x_i then we calculate the value of $-\text{grad}(f_p(x_i))$ at this point

and obtain an improved estimate of the optimum by:

$$f_{p+1} = f_p - \lambda \text{grad}(f_p)$$

λ is a parameter which controls the distance which is travelled along this line of maximum improvement. λ may be either estimated or by gradually increasing λ until f_{p+1} just begins to increase.

The procedure may be illustrated in two dimensions by the following diagram:



One of the disadvantages of this procedure is that local optima can sometimes be found. This is normally partially avoided by following the procedure several times with different starting positions.

For further details of this technique of optimization and others see Kruskal (1964), Spang (1962) and Wilde and Beightler (1967). Other methods which are faster in some instances are given by Fletcher and Powell (1963), Rosenbrock (1960).

B.4 SERIATION

For certain types of problems in multivariate data reduction it may be necessary to order objects on a one-dimensional scale. Techniques which perform this type of analysis are called <u>seriation</u> methods. The aims of these methods are very similar to some scaling methods (as used in psychology), they only produce rank orderings and do not give any measures of relative closeness on a scale. The methods of psychological scaling and seriation overlap to a large extent, and the use of different names and approaches is due to different historical roots - archaeology where temporal ordering is of importance, and scaling methods have been developed independently in psychology.

The principle on which archaeological seriation is based is that if objects are from a similar time period then they will be more similar than objects from totally different periods. This concept (which is at least partially due to Robinson 1951 and Brainerd 1951) implies that if we have a matrix of similarity between archaeological artefacts, such as beakers (see Clarke 1962) or brooches (see Hodson et al 1966), and the order in which the artefacts appear in the matrix corresponds with the historical order then we would expect to have similarities increasing towards the diagonal of the matrix. In other words, we would expect, in a perfect case:

$$s(i,i) \geqslant s(i,i+1) \geqslant s(i,i+2) \geqslant \ldots\ldots \geqslant s(i,n)$$

and $s(i,i) \geqslant s(i,i-1) \geqslant s(i,i-2) \geqslant \ldots\ldots \geqslant s(i,1)$

(where $s(k,j)$ represents the similarity between artefacts k and j).

Thus in a real situation with an unordered matrix, by interchanging the rows and columns corresponding to objects, one should be able to find the ordering which best reproduces the above form. This form is termed the Robinson form by Kendall (1969), but is more generally known in mathematics as the Toeplitz decreasing pattern (Renfrew and Sterud 1969).

The development of seriation can perhaps best be described with the aid of a diagram depicting the references between major papers on the subject. This is shown in Figure 1. (Note: if paper B refers to paper A and paper C refers to both A and B, then the link C-A is not shown.)

The first appearance of seriation comes under the name of sequence-dating and was performed by Petrie (1899) by a 'masterly combination of subjective and objective methods' (Kendall 1963). However a more scientific approach was proposed (apparently independently) over fifty years later by Robinson (1951) and Brainerd (1951).

The Brainerd and Robinson papers initiated a chain of heuristic computer programs to arrange a matrix into near-Robinson form - see Ascher and Ascher (1963), Kuzara et al (1966), Hole and Shaw (1967), Craytor and Johnson (1968), and Johnson (1968), each of which attempts to improve upon the results of the preceding paper (and increases the computer time). The seriation problem is very much related to psychological scaling in one dimension (as in Torgerson 1958 and Coombs 1964) although this was not apparently recognized until the middle nineteen sixties, when Kruskal's program MDSCAL was seen to be able to be used for seriation (see

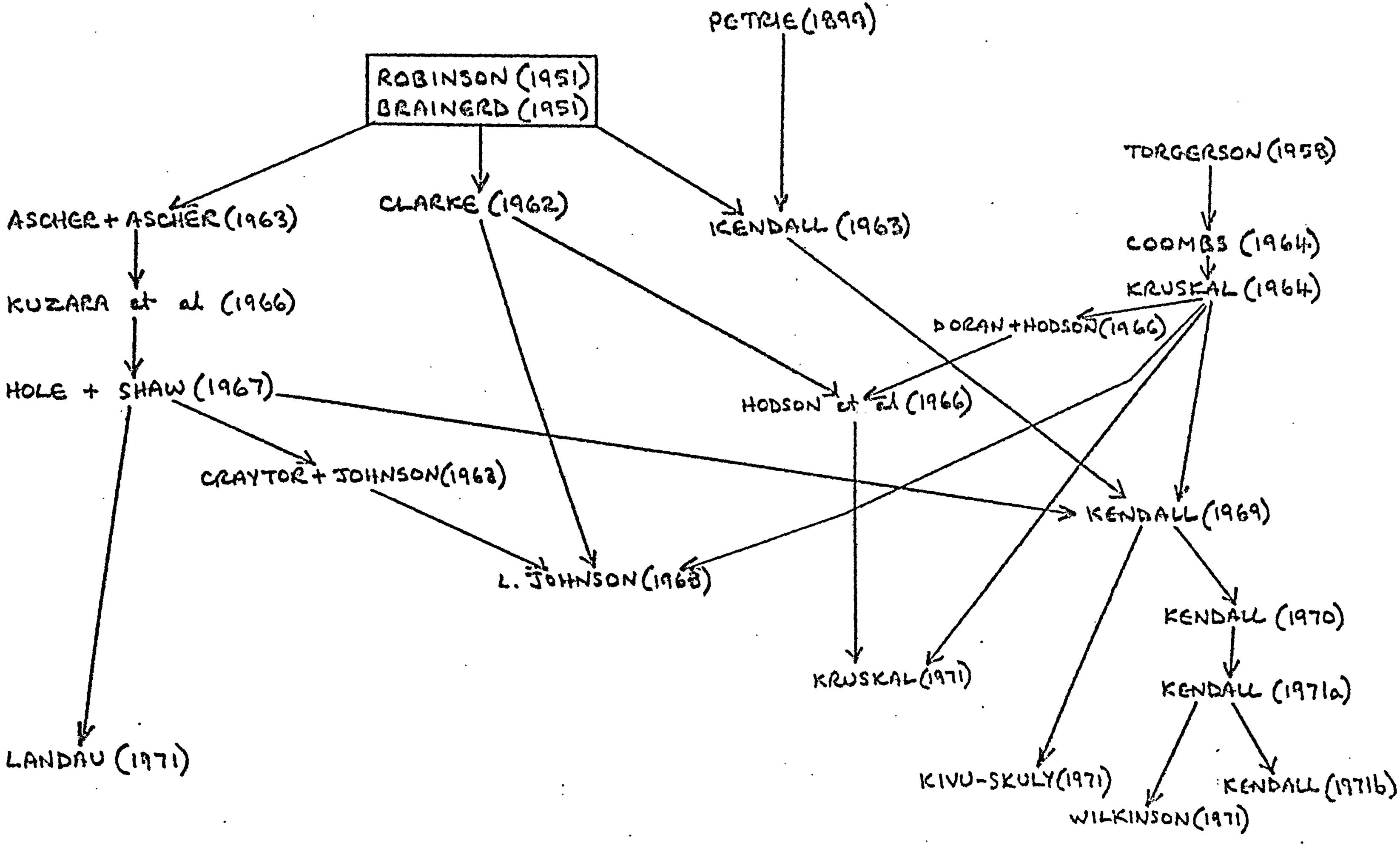

PETRIE (1899)
ROBINSON (1951)
BRAINERD (1951)
TORGERSON (1958)
ASCHER + ASCHER (1963)
CLARKE (1962)
KENDALL (1963)
COOMBS (1964)
KRUSKAL (1964)
KUZARA et al (1966)
DORAN + HODSON (1966)
HODSON et al (1966)
HOLE + SHAW (1967)
CRAYTOR + JOHNSON (1968)
KENDALL (1969)
L. JOHNSON (1968)
KENDALL (1970)
KRUSKAL (1971)
KENDALL (1971a)
LANDAU (1971)
KIVU-SKULY (1971)
KENDALL (1971b)
WILKINSON (1971)

FIGURE 1.

Doran and Hodson 1966, Hodson et al 1966, Cowgill 1968). Recent work by Kendall (1969, 1970, 1971) and papers contained in Hodson et al (1971) have continued the application of one-dimensional scaling to archaeological data and have begun to consider the implications of higher-dimensional scalings (see Kruskal 1971). Wilkinson (1971) has related the travelling salesman problem to seriation.

A difficulty with all methods mentioned so far is that of local optima. Since the methods cannot consider all possible orderings of the objects (10 objects can be ordered in 10! = 3,628,800 ways, 20 objects can be ordered in over 2×10^{18} ways), the heuristic programs consider subsets of the maximum number, and may thus find local optima. The psychological scaling methods of MD-SCAL type (see later) employ hill-climbing methods which also can result in sub-optimal solutions.

With the heuristic programs, as different orderings are considered, a criteria has to be used to determine if a particular reordering is 'better' than the previous ordering. With this type of method it can be defined by two considerations:

A. Which reorderings are evaluated.

B. How improvement is measured.

We proceed to define the major methods by these criteria:

Robinson (1951) (see also Troike 1957 and Belous 1953)

A. Reorderings by inspection.

B. For the initial stages a reordering takes place if the number of negatively signed differences (the number of cases where the similarities do not increase towards the diagonal) can be reduced. In the later stages the magnitude of the similarities is used by trying to decrease the ratio of the sum of squares of all the negative differences, and the sum of squares of all differences.

Ascher and Ascher (1963)

A. Agglomerative building of the matrix. Take the first two objects and consider if the third object is best placed in front, between, or at the end of the first two. Then take the fourth and find the best position for this, and so on. If any objects cannot be placed so that the matrix has Robinson form then they are discarded until all other objects which can be put into the matrix with Robinson form have been included, and then the 'best' position for these discarded objects is found. No exchanges are attempted.

B. A 'best' position is determined by the minimum number of negative differences.

Kuzara et al (1966)

A. Every object is tried in every position relative to the others until the 'best' is found. This is repeated until no improvement is found. This procedure is initiated several times with different starting positions.

B. 1. The sum of negative differences over the sum of all differences.

2. The sum of squares of negative differences over the sum of squares of all differences (as in Robinson).

Hole and Shaw (1967)

A. Alternating between two strategies - pairwise interchange which considers swapping each of the $\frac{1}{2}n(n-1)$ pairs of objects for improvement, and successive rotation, trying each object in all n possible places in the matrix, until no improvement can be found.

B. Sum of squares of negative differences.

(Note: Kivu-Sculy 1971 discusses the method with further elaborations.)

Craytor and Johnson (1968), Johnson (1968)

A. Every object is tried in each position (as in Kuzara et al).

B. Sum of all differences.

Clarke (1962) has also used a sorting routine to produce seriations but does not fully report his results. Tugby (1965) suggests the use of mechanical sorting devices. Kendall (1963) also discusses and comments on Robinson's paper.

The foundations of seriation from the psychological field date back to the first multidimensional scaling experiment by Richardson (1938) which lead to the non-metric multidimensional scaling method due to Shepard (1962) and Kruskal (1964). The first use of this method in archaeology was in Doran and Hodson (1966), although the aim of this study was not primarily seriation. Hodson, Sneath and Doran (1966) used the method for seriation and obtained

encouraging results. Kendall has also used the method with some success (Kendall 1969, 1970, 1971).

(Surprisingly, the fact that other one-dimensional ordination procedures might produce good seriations appears to have been passed by, even other non-metric multidimensional scaling methods have not been used.)

An unusual approach of Kendall's work is that he does not produce a one-dimensional ordination but employs a two-dimensional plot so as to incorporate more user control and to reduce the possibility of local minima (Kendall 1971).

A more appropriate procedure would be to use the stress values from the program for different dimensions to determine the true dimensionality of the data and investigate that ordination. Kruskal (1971) suggests that one dimension may not be enough to describe the data, and the example given in Doran and Hodson 'refused to yield a low strain configuration in one dimension'. Wishart and Leach (1970) have attempted a one-dimensional MD-SCAL on Platonic texts and below is their similarity matrix rearranged in the order they suggest.

PHAEDRUS	X									
REPUBLIC	0.45	X								
SYMPOSIUM	0.67	0.07	X							
TIMAEUS	1.08	1.01	1.11	X						
SOPHISTES	1.23	0.78	0.80	0.15	X					
CRITIAS	1.22	0.63	0.57	0.27	0.15	X				
7TH EPISTLE	1.28	0.70	0.71	0.40	0.27	0.19	X			
POLITICUS	1.68	0.90	0.92	0.40	0.27	0.16	0.27	X		
PHILEBUS	1.73	0.61	0.58	1.03	0.63	0.44	0.34	0.33	X	
LAWS	2.23	0.87	0.79	1.45	0.91	0.69	0.57	0.53	0.08	X

It can be seen that apart from Republic and Symposium there is almost perfect Robinson form, but these two cannot really be fitted anywhere in the diagram with satisfaction. Boneva (1971) has also analysed Platonic prose and did not proceed to a one-dimensional ordination. (Griffith 1967 has referred to the use of clustering in literature studies.)

The suggestion that a one-dimensional solution may not give a true representation is not new, Ford (1954) suggested that the regional might be more than the temporal variation.

It has been suggested that if clusters are present in the data then this can be seen by inspection of the seriated matrix. A process similar to seriation has been used in sociological studies for determining group structure in human social behaviour (see Coleman and MacRae 1960, Chabot 1950, Beum and Brundage 1950 and Forsyth and Katz 1946). The use of seriation as cluster analysis is not recommended, but if a

seriation is available or required then it may give some indication of whether clustering would be a fruitful approach - Hodson (1970) sums up with the words -

> "it is difficult to accept seriation as a serious approach to find clusters".

B.5 DISSIMILARITY AND SIMILARITY MEASURES

As was discussed in Section B.2, the similarity stage in any cluster analysis is a very important stage. This stage is possibly even more important in an ordination study. The need for the similarity matrix arises from the definition of a cluster - requiring 'similarity' between cluster numbers - this can be assessed directly from a similarity matrix as it enables us to compare a pair of objects by a single measure of closeness. Ordination is also necessarily based on similarity, since that is what one seeks to represent in an ordination study.

The inputs to the similarity phase are normally from a data matrix of several observations measured on several variables, but they may be obtained directly from the observations themselves. An example of directly obtained similarities is if subjects are asked to rank a set of pairs of objects in order of similarity. In the cases where the data matrix is bypassed, the distance matrix is not obtained by calculation, thus in this section we shall restrict ourselves to distances obtained from data blocks. The similarity matrix obtained is the input for the next phase - the method. This may constrict the similarities phase because some methods require a certain type of similarity input, thus in a particular study, before calculating similarities, one must consider the possible methods.

Metrics

Jardine and Sibson (1971) define a dissimilarity coefficient as a function $P : P \times P \rightarrow R$ $(R \neq 0)$ such that:

1. $d(a,b) \geq 0$ for all $a,b \in P$
2. $d(a,a) = 0$ for all $a \in P$
3. $d(a,b) = d(b,a)$ for all $a,b \in P$

(Note in Jardine and Sibson (1968) they define it differently including:

$d(x,y) = 0$ if and only if $x = y$

which is a stronger condition than above. For our purposes we use their later version which allows two observations to have the same values in the data matrix. This can easily occur for example in biology where two animals of the same species may well be identical on a set of measurements.)

However, these are only necessary conditions and not sufficient - for any symmetric matrix with zero diagonal and no negative elements fulfils the requirements of the above definition. To be a meaningful dissimilarity coefficient the matrix must have a relationship with the observations which preserves our intuitive meaning for dissimilarity. This relationship is difficult to define, and indeed its definition will vary with the particular measure in question.

If we add a further condition, we obtain the set of coefficients which we call metric coefficients or simply distance measures (although this latter term has lost its meaning through misuse). This condition is:

4. $d(a,c) \leq d(a,b) + d(b,c)$ for all $a, b, c \in P$

This is the so-called triangle equality which if we imagine the observations as points in space requires the longest side of the triangle formed by the lines joining three observations to be no greater than the sum of the other two sides.

Metric coefficients are of importance because most of the widely used dissimilarity measures are of this type.

A similarity coefficient can be defined by analogy to a dissimilarity coefficient by the axioms:

1'. $s(a,b) \leqslant K$ for all $a, b \in P$

2'. $s(a,a) = K$ for all $a \in P$

3'. $s(a,b) = s(b,a)$ for all $a, b \in P$

(Note: This is at variance with the definition given by Harrison (1968) who states that "a similarity function only takes values in the interval $[0,1]$". This definition is clearly at fault for it excludes, amongst others, the correlation coefficient which takes values in the range $[-1,1]$.)

There is no direct equivalent of the metric coefficient with similarities.

If, further to our conditions necessary to define a metric space we use another stronger condition, to replace axiom 4:

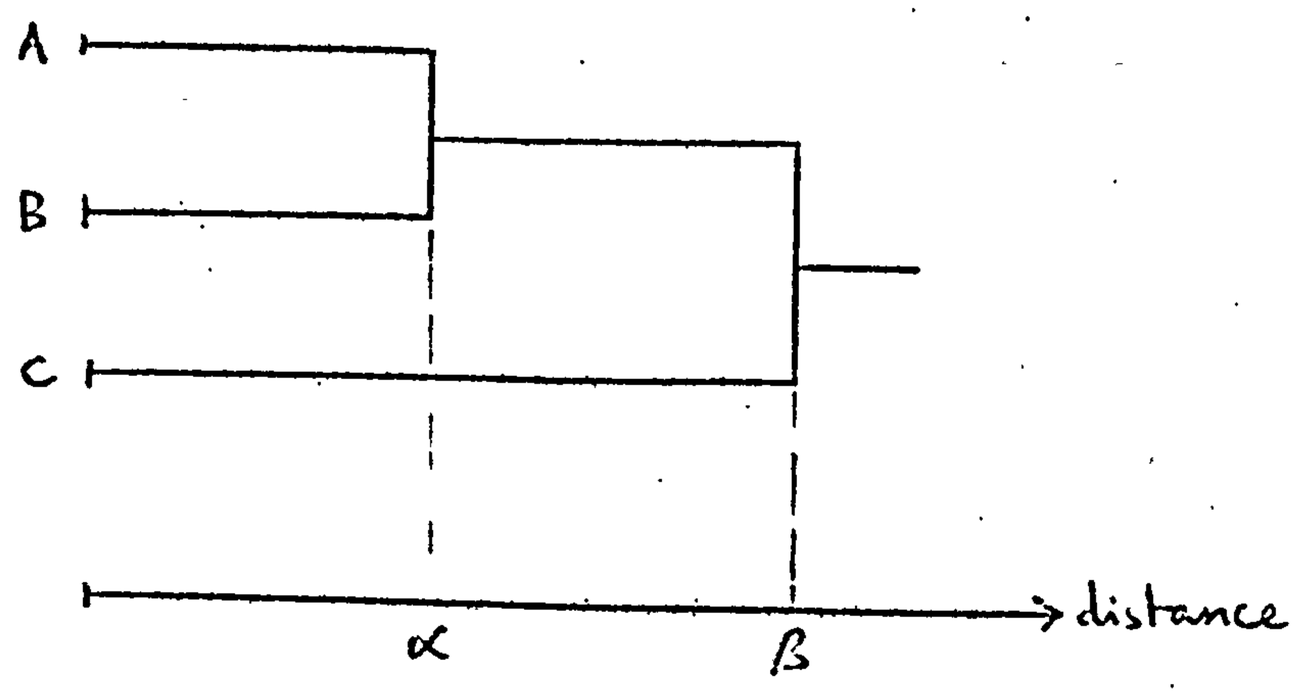
5. $d(a,c) \leqslant \max(d(a,b), d(b,c))$ for all $a,b,c \in P$

then we have an ultrametric space and the above relationship is called the ultrametric inequality. Axiom 5 implies (if

we consider the axiom applied three times to any triad) that any three observations form a triangle with the two largest sides equal.

The importance of an ultrametric space is that it defines a hierarchic dendrogram. We can show that a hierarchy obeys the ultrametric inequality as follows:

Any three points in a dendrogram can be arranged in the following form:



We have $d(A,B) = \alpha$, $d(A,C) = \beta$, $d(B,C) = \beta$

and since $\beta \geq \alpha$

we have

$$d(A,B) \leq \max(d(A,C), d(B,C)) \qquad \alpha \leq \max(\beta,\beta)$$

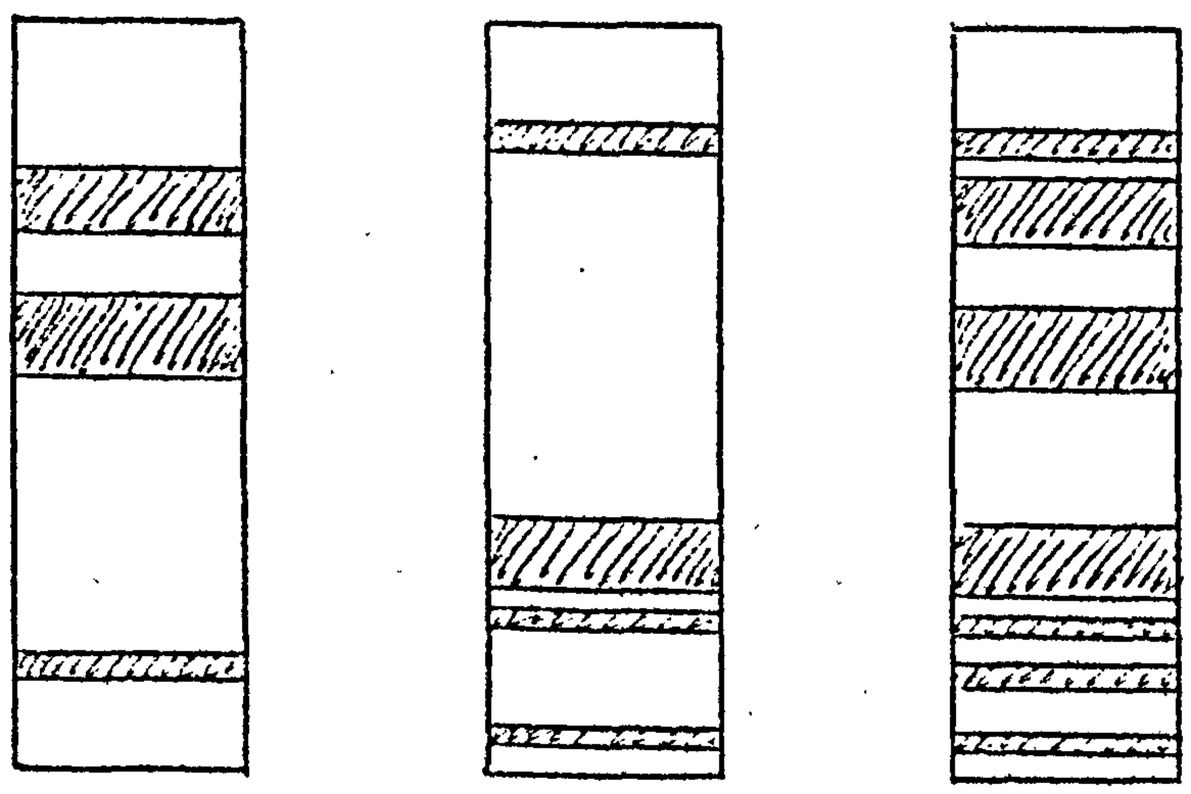
$$d(A,C) \leq \max(d(A,B), d(B,C)) \qquad \beta \leq \max(\alpha,\beta)$$

$$d(B,C) \leq \max(d(A,C), d(A,B)) \qquad \beta \leq \max(\beta,\alpha)$$

Hence any dendrogram is ultrametric.

Thus the process of forming a dendrogram can be considered as the process of transforming the given distance matrix into a distance matrix which obeys the ultrametric inequality in a way which is in some sense optimal. The clustering method of Roux (1969) is based on this fact and systematically reduces to the 'sub-dominant ultrametric' of the distance matrix.

Metric measures are not always appropriate, especially in some particular applications. For example, in linguistics one word can have two different meanings and thus be very similar to two very dissimilar words. Another example is given by Ornstein (1965) where spectra, such as the following three, may occur in an experiment.



1 2 3

Here spectrum 3 is similar to both 1 and 2, but spectra 1 and 2 bear no relation to one another.

A useful transformation given by Majone (1968) and Majone and Sanday (1971), transforms from one metric to another which has range 0 to M. The expression is:

$$d'(a,b) = \frac{M \times d(a,b)}{1+d(a,b)}$$

which can easily be shown to obey the triangle inequality.

Types of Similarity Measure

As discussed in Section B.2 there are several types of data which may be produced for input to the similarity stage. Each variable may be measured in one of four general types, and a data matrix can thus contain any combination of these types.

The types are:

1. Binary (or Two-State or Dichotomous)
2. Unorderable Multi-State
3. Ordered Multi-State
4. Continuous

We shall continue to discuss each of these types individually before proceeding to the question of mixed data types.

1. Binary Measures

If we consider the following 2 by 2 contingency table we can express most similarity coefficients in the terms of the elements within.

		Observation X		
		1	0	
Observation Y	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	a+b+c+d=n

Thus a is the number of variables on which both X and Y score 1, b is the number of variables on which X scores 0 and Y scores 1, etc. Thus a, b, c, d and n are non-negative integers.

In order to describe the similarity between X and Y, a fairly obvious measure would be to use the ratio of the number of variables in which X and Y have the same value, to the total number of variables.

I.e. $\frac{a+d}{n}$ which has been called the Simple Matching Coefficient by Sokal and Michener (1958).

However, in many applications such as botany and biology, all the variables may be of the so-called presence-absence type. With this type of variable, 0 signifies that an observation does not possess this variable (or attribute) and 1 signifies that this attribute is possessed. In this case d is less meaningful than a since two objects that possessed none of the attributes would attain maximum similarity with the Simple Matching Coefficient. Thus in these cases one could use

$\frac{a}{n}$ which is the coefficient due to Russell and Rao (1940). Alternatively, since in many studies observations are so diverse that many variables are needed a particular object may only score on a small percentage of attributes, thus

$\frac{a}{a+b+c}$ may be more meaningful in these cases. This coefficient can be traced back as far as Jaccard (1908).

Kulczynski (1927) has suggested $\frac{a}{b+c}$ which is related to the first two coefficients.

Two other coefficients of a similar nature have been suggested. That of Dice (1945) is $\frac{2a}{2a+b+c}$. This is an apparent attempt to reweight the denominator of the Jaccard coefficient so that presences have the same weight as absences. The other which Sokal and Sneath (1963) call the Tanimoto coefficient (and give the reference Rogers and Tanimoto 1960 - although in this reference the similarity measure used is Jaccard's) is

$\frac{a+d}{a+2b+2c+d}$ which weights mismatches double in the denominator.

Other workers (such as Cheetham and Hazel 1969) still refer to this coefficient with the reference Rogers and Tanimoto 1960, but we believe this arises from an error by Sokal and Sneath, and also since no simple logical basis can be propounded for counting mismatches twice, we regard this coefficient as of little value.

Other similar coefficients have been suggested, but these bear simple relationships with the above measures. For example, the Coefficient of Floral Community (which, unlike the measures discussed so far is a dissimilarity measure) which is

$$\frac{b+c}{2a+b+c}$$

This is simply $1-S_D$ where S_D is the measure of Dice.

The Coefficient of Floral Community can be shown not to be a metric measure by a simple counter-example

		variable	
		1	2
object	1	1	1
	2	0	1
	3	1	0

$d_{12} = d_{13} = \frac{1}{3}$ $\qquad$ $d_{23} = 1$

$d_{12} + d_{13} < d_{23}$ $\qquad$ thus d is not a metric function

Clements, Schenk and Brown (1926) have proposed the coefficient

$$S_H = \frac{a+d-b-c}{n}$$

but this is a variation of Sokal and Michener's coefficient, since

$$S_{sm} = 1 + \frac{S_H}{2}$$

Stephenson, Williams and Lance (1968) have proposed the Number of Features of Difference (NFD) as a dissimilarity coefficient

$$NFD = b + c$$

However this measure cannot be compared for different studies since any increase in the total number of attributes measured must increase the NFD, thus if we standardize to $\frac{b+c}{n}$ we obtain

$$NFD = 1 - S_{SM}$$

Thus we can consider all these measures as being based on the following set:

1. $S_{SM} = \frac{a+d}{n}$

2. $S_{RR} = \frac{a}{n}$

3. $S_J = \frac{a}{a+b+c}$

4. $S_D = \frac{2a}{2a+b+c}$

The relationships between these can be shown graphically. If we consider the case where n=10 and graph similarity against a, we can obtain a set of points depending on the particular values of d and (b+c). See Figures 2 and 3.

From these graphs it can be seen that

$$0 \leq S_{RR} \leq S_J \leq S_{SM} \leq 1$$

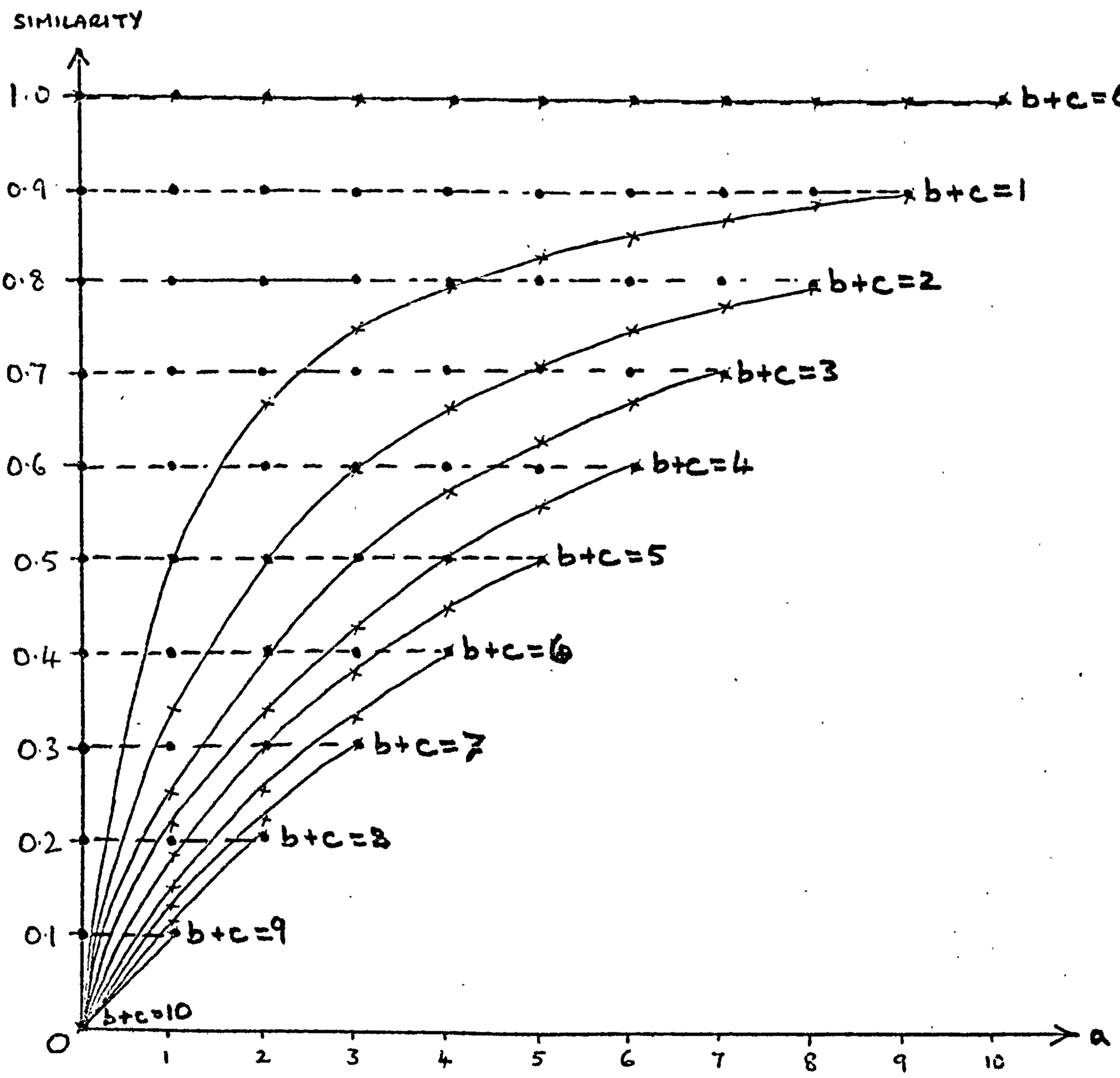
and

$$S_J \leq S_D \leq 1$$

This can easily be proved mathematically from the coefficients.

The graphs for other values of n are similar to thos shown, for example with n=20 on the graph shown the a scale is relabelled in steps of 2 up to 20 and the curves become b+c=20, b+c=18, b+c=16, ...b+c=0.

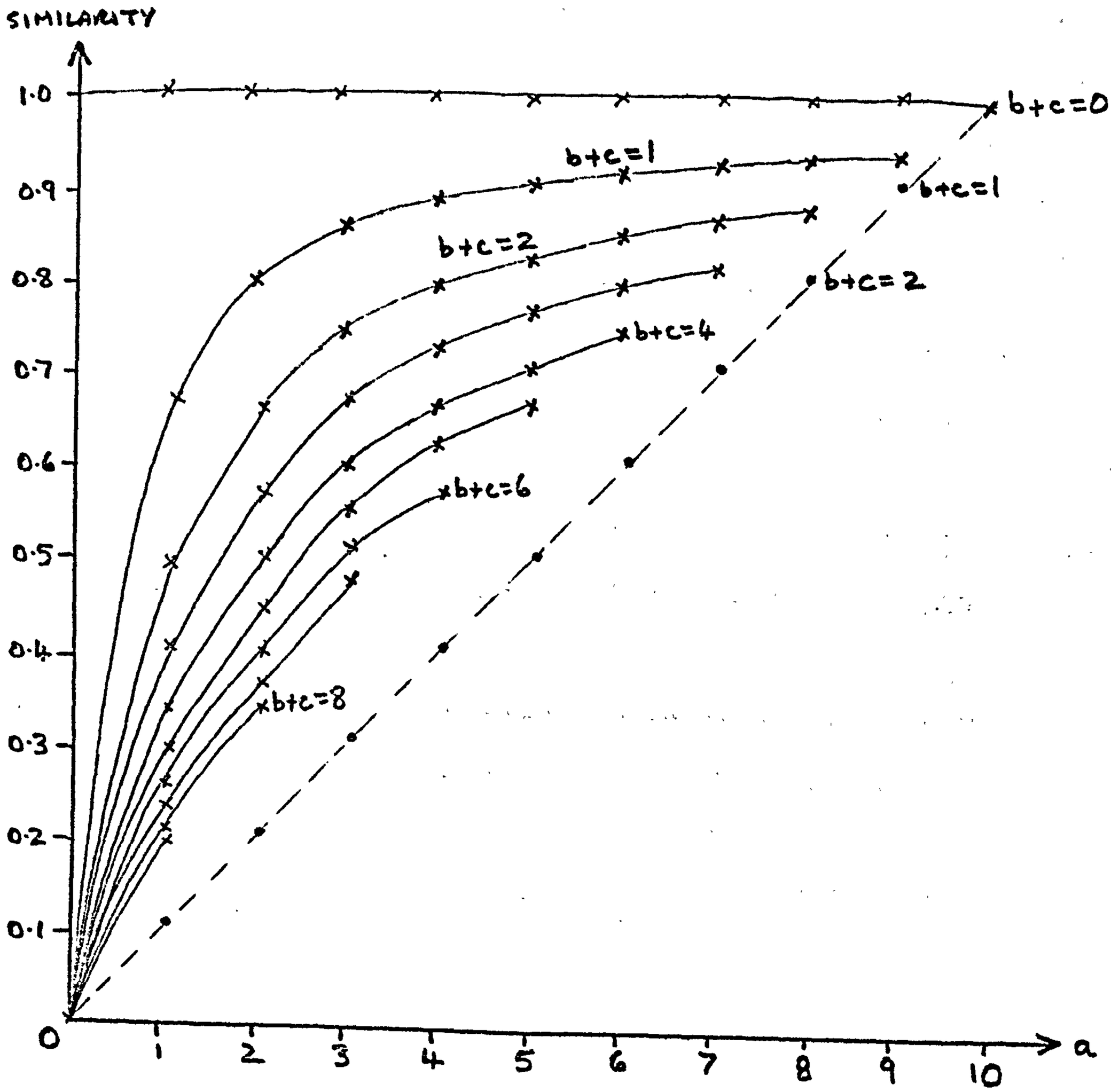
The curves drawn on the graph have no meaning in themselves as a takes only integer values, but serve to connect points in the same sequence.



S_{SM} ●----● S_D ×——×

GRAPH of similarity v. a
for $a+b+c+d = n = 10$

FIGURE 2



$\times$——$\times$ S_D

$\bullet$- - - -$\circ$ S_{RR}

GRAPH of Similarity v. a
for $a+b+c+d = 10 = n$

FIGURE 3

It can be seen that S_J and S_D produce similar curves - both are fairly insensitive to changes in a when b+c is low and a is over n/3, and sensitive to changes in a when a is less than n/3. All four methods give similar results with a slightly less than n, and may be markedly different with $a < n/2$.

In choosing between the four coefficients it can be seen that S_J and S_D are little different except that S_D is more sensitive to error when a is small, thus S_J is preferred. S_{SM} should not be used with presence/absence data since d does not contain as much information as a. Thus S_{RR} and S_J are left for consideration. Each has its own advantages. We may argue that the addition of a new attribute to our list should make no difference to our similarity between 2 objects if neither possesses this new attribute. S_J is unaffected by this new attribute, but S_{RR} is decreased. Alternatively we may argue that with S_J the values are not comparable between different studies as its value does not alter as d alters.

The distribution of S_{SM} has been examined by Goodall (1967). He has shown that Sokal and Sneath's (1963) estimate of the variance of the coefficient is not correct and its true variance is that of the Poisson binomial distribution.

We proceed to consider other similarity measures not of the simple form so far discussed. Another set of binary measures is based on χ^2. Perhaps one of the first specifically for binary data was the fourfold point correlation coefficient or Pearsons phi coefficient, where

$$\phi = \frac{ad-bc}{\sqrt{(a+c)(b+d)(c+d)(a+b)}}$$

(Note that $\phi^2 = \chi^2/n$)

Measures based on χ^2 are applicable to multi-level data (see later) and more general formulae may be reduced to their special 2 by 2 cases, for instance Tschuprows coefficient becomes the same as and that of Cramer (1946) is χ^2 for the 2 by 2 case.

Pearson has also proposed $\frac{\phi}{\sqrt{1+\phi^2}}$ which has been called the coefficient of mean square contingency.

Other writers have used the same numerator as that of Pearons . A famous example is Yules coefficient of association

$$Q = \frac{ad - bc}{ad + bc}$$

and a variation called Yules coefficient of colligation

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

It can be shown (Kendall and Stuart 1967) that

$$Q = \frac{2Y}{1+Y^2}$$

and Kendall and Stuart state that hence nothing much is to be gained from Y. In fact Wilson (1931) states that Yule came to the conclusion that neither of his two methods was safe.

All the above measures include terms with bc, thus if b is increased by 1 and c is decreased by 1, the similarity measure generally changes. This seems intuitively wrong for a similarity measure. For the case where data is not presence/absence and a variable may thus be coded in the reverse way by replacing 1 by 0 and 0 by 1, the measures including terms of bc change with any reverse coding.

E.g.

X	0	1	1	1	0
Y	1	0	1	1	0

becomes

X \ Y	1	0
1	2	1
0	1	1

Thus $\phi = \frac{1}{6}$ $\quad Q = \frac{1}{3}$

whereas if we reverse the first variable we have

X	1	1	1	1	0
Y	0	0	1	1	0

which gives

X \ Y	1	0
1	2	2
0	0	1

Thus $\phi = \frac{1}{\sqrt{6}}$ $\quad Q = 1$

Thus for presence/absence data the methods are suspect and for other data the coefficients are of no value. Goodman and Kruskal (1954) write:

> "the fact that an excellent test of independence may be based on χ^2 does not at all mean that χ^2 of some simple function of it is an appropriate measure of degree of association".

Dagnelie (1965) has used the coefficient $\frac{1}{n}(ad-bc)$, and Freeman (1970) has used $(ad-bc)^2$ and $|ad-bc|$, which are virtually identical.

The coefficient of Forbes (1907) is

$\frac{na}{(a+b)(a+c)}$ which when 1 is subtracted from it, as suggested by Cole (1949) becomes

$$S_F = \frac{ad-bc}{(a+b)(a+c)}$$

which is the same as ϕ, but with a replacing d in the denominator. Cole also dismisses the 'coincidence index' falsely attributed to Dice (1945), $\frac{2a}{(a+b)(a+c)}$ as being simply Forbes' original coefficient multiplied by $\frac{2}{n}$, however Forbes' measure is dependent on d, whereas the other measure is independent.

Michael (1920) has used the measure

$S_m = \frac{4(ad-bc)}{(a+d)^2 + (b+c)^2}$ which is again of similar form to the above measures.

Thompson (1916) has used

$\frac{a}{\sqrt{(a+b)(a+c)}}$ which is the square root of that used by Sorgenfrei (1959).

The coefficient due to Cole (1949) depends on the relative magnitude of a, b, c and d.

$$S_c = \begin{cases} \dfrac{ad-bc}{(a+b)(b+d)} & \text{when } ad \geqslant bc \\ \dfrac{ad-bc}{(a+b)(a+c)} & \text{when } ad < bc \text{ and } a \leqslant d \\ \dfrac{ad-bc}{(b+d)(c+d)} & \text{when } ad < bc \text{ and } a > d \end{cases}$$

Hurlbert (1969) has shown that this can be expressed in the form

$$S_c = \frac{ad-bc}{|ad-bc|} \left| \sqrt{\frac{\text{Obs } \chi^2}{\text{Max } \chi^2}} \right|$$

where Obs χ^2 is the value of χ^2 from the contingency table and Max χ^2 is the value of χ^2 as large (if $ad \geqslant bc$) or as small (if $ad < bc$) as the marginal totals (a+b, a+c, b+d, c+d) will permit. Hurlbert shows that the measure is biased in that it is influenced by any observation that scores on either many or few attributes.

Hurlbert suggests

$$C_8 = \frac{ad-bc}{|ad-bc|} \left| \sqrt{\frac{\text{Obs } \chi^2 - \text{Min } \chi^2}{\text{Max } \chi^2 - \text{Min } \chi^2}} \right|$$

to remedy this bias.

Wallis (1928) has suggested the use of

$$S_S = \frac{a}{\min(a+b, a+c)}$$ and that due to Braun-Blanquet (1932)

is given by $\dfrac{a}{\max(a+b, a+c)}$

Savages Coefficient of Difference is simply one minus this last expression.

Hole and Hironakas (1960) coefficient

$$S_{HH} = \frac{2 \min(a+b, a+c)}{2a+b+c}$$

is a simple relationship of other measures.

$$S_{HH} = S_K / S_S$$

Fager (1963) has proposed the measure

$$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\max(a+b, a+c)}}$$

the first expression of which is the same as that of Thompson and the second expression is similar to that of Braun-Blanquet.

Kulczynski (1927) has used $\frac{a}{2}(\frac{1}{a+b} + \frac{1}{a+c})$ which is a similar type of coefficient.

Sokal and Sneath have suggested an average taxonomic distance which reduced to $\sqrt{\frac{b+c}{n}}$ for binary data, this is simply the square root of $1 - S_{RR}$.

Finally, perhaps the most complicated measure that has been used is Preston's (1962) Resemblance Equation, where the similarity Z is calculated from the following expression:

$$(a+c)^{1/Z} + (a+b)^{1/Z} = (a+b+c)^{1/Z}$$

Unfortunately Z cannot be calculated directly from this expression.

We summarize our list of coefficients in Table 1, excluding those which are variations of others, (and Table 2 shows, for completeness, these remaining coefficients).

We have already discussed the first four measures in detail and the discussion on χ^2 covers measures , S_P and C_8. With measures 6 to 18 in Table 1 it can be seen that the relative numerical values of b and c are of importance. We have considered the case of non-presence/absence data and shown that the similarity may be altered by the reversal of the coding of a particular attribute. Consider now the following two examples of presence/absence data:

A. A set of insects are under study. They are coded as to whether they possess various attributes e.g. do they possess wings, antennae, jointed legs, etc.

B. An area of marshland is divided up into quadrats and each quadrat is examined to see which plants, e.g. reeds, grasses, etc., are found in that square. The plants are thus scored (0) if they do not exist together in a particular quadrat and (1) if they do co-exist.

Suppose the contingency table in both cases was as follows:

		insect/plant Y	
		1	0
insect/plant X	1	2	0
	0	4	6

In case A, the insects X and Y agree on only two out of a possible twelve attributes and thus one would expect the similarity between them to be small and indeed $S_J = 1/3$, $S_{RR} = 1/6$ and $S_D = 1/2$. One might assume that for the

1.	SOKAL & MICHENER (1958)	$S_{SM} = \frac{a+d}{n}$
2.	RUSSELL & RAO (1940)	$S_{RR} = \frac{a}{n}$
3.	JACCARD (1908)	$S_J = \frac{a}{a+b+c}$
4.	DICE (1945), BURT (1958)	$S_D = \frac{2a}{2a+b+c}$
5.	TANIMOTO	$S_T = \frac{a+d}{a+2b+2c+d}$
6.	PEARSON	$\phi = \frac{ad-bc}{\sqrt{(a+c)(b+d)(c+d)(a+b)}}$
7.	YULE	$Q = \frac{ad-bc}{ad+bc}$
8.	DAGNELIE (1965), FREEMAN (1970)	$S_{DA} = \frac{1}{n}(ad-bc)$
9.	FORBES (1907), COLE (1949)	$S_{FO} = \frac{ad-bc}{(a+b)(a+c)}$
10.	COLE (1949)	$S_C = \frac{2a}{(a+b)(a+c)}$
11.	(McEWAN &) MICHAEL (1920)	$S_M = \frac{4(ad-bc)}{(a+d)^2 + (b+c)^2}$
12.	THOMSON (1916), OTSUKA & OCHIAI (1957)	$S_{TH} = \frac{a}{\sqrt{(a+b)(a+c)}}$
13.	HURLBERT (1969)	$C_8 = \frac{ad-bc}{\lvert ad-bc \rvert} \left\lvert \sqrt{\frac{\text{Obs } \chi^2 - \text{Min } \chi^2}{\text{Max } \chi^2 - \text{Min } \chi^2}} \right\rvert$
14.	WALLIS (1928), SIMPSON (1943)	$S_W = \frac{a}{\text{Min } (a+b,\ a+c)}$
15.	BRAUN-BLANQUET (1932)	$S_B = \frac{a}{\text{Max } (a+b,\ a+c)}$
16.	FAGER (1963)	$S_{FA} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\text{Max}(a+b, a+c)}}$
17.	KULCZYNSKI (1927), DRIVER & KROEBER (1932)	$S_{K2} = \frac{a}{2}(\frac{1}{a+b} + \frac{1}{a+c})$
18.	PRESTON (1962)	Z where $(a+c)^{1/Z} + (a+b)^{1/Z} = (a+b+c)^{1/Z}$

TABLE 1

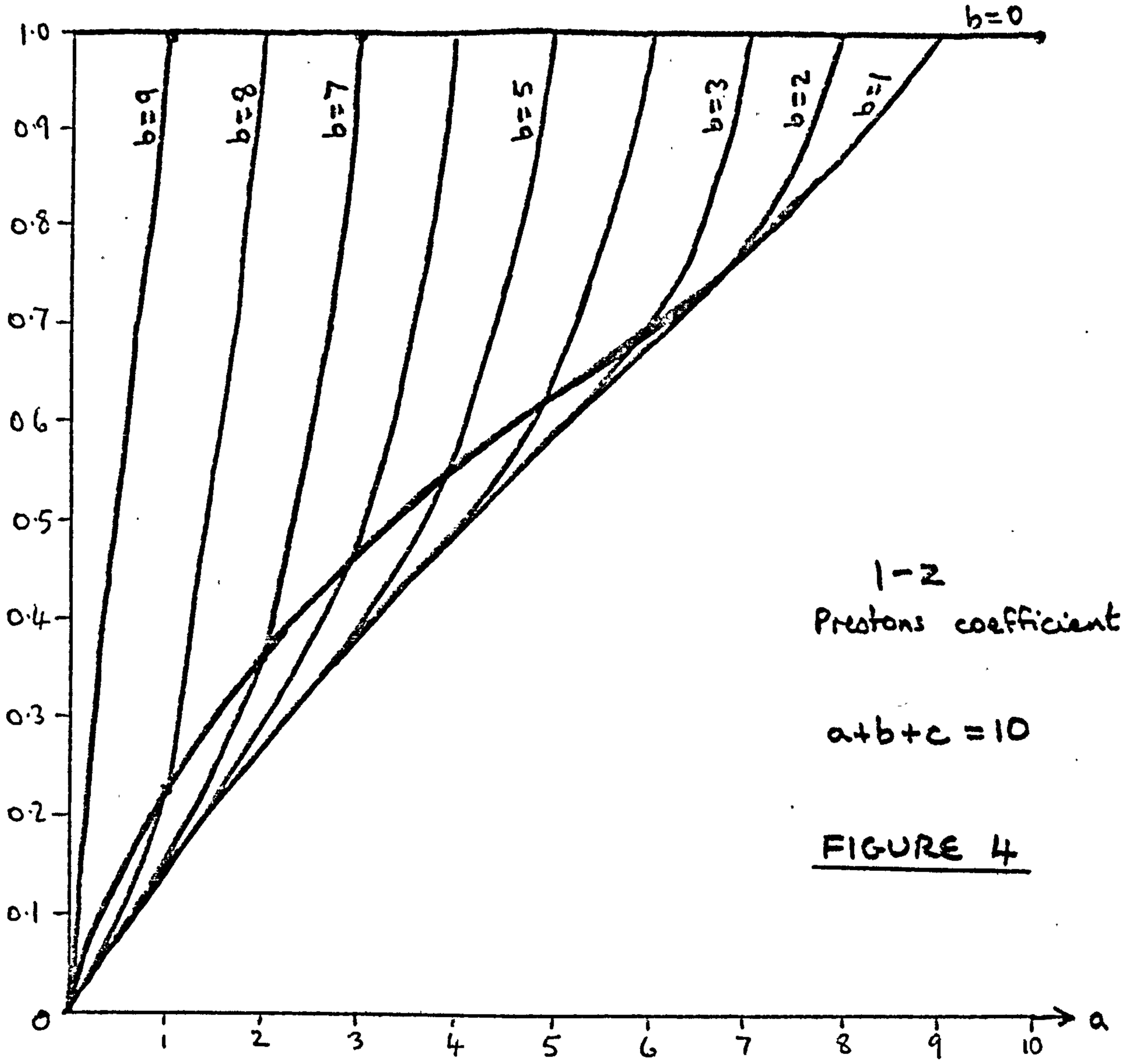
			Variation of
1.	KULCZYNSKI (1927), HARRISON (1968)	$S_{K1} = \frac{a}{b+c}$	$S_{RR}/(1-S_{SM})$
2.	COEFFICIENT OF FLORAL COMMUNITY	$S_{FC} = \frac{b+c}{2a+b+c}$	S_D
3.	CLEMENTS ET AL (1926), HAMANN (1961)	$S_{CS} = \frac{a+d-b-c}{n}$	S_{SM}
4.	STEPHENSON, WILLIAMS & LANCE (1968)	$NFD = b+c$	S_{SLI}
5.	PEARSON	$S_p = \frac{\phi}{\sqrt{1+\phi^2}}$	ϕ
6.	YULE	$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	Q
7.	SORGENFREI (1959)	$S_s = \frac{a^2}{(a+b)(a+c)}$	S_{TH}
8.	COLE (1949)	$S_c = \frac{ad-bc}{\lvert ad-bc \rvert} \left\lvert \sqrt{\frac{Obs\,\chi^2}{Max\,\chi^2}} \right\rvert$	C_8
9.	SAVAGE (1960)	$S_{SA} = \frac{\max(b,c)}{a+\max(b,c)}$	S_B
10.	HOLE & HIRONAKA (1960)	$S_{HH} = \frac{2\min(a+b,a+c)}{2a+b+c}$	S_D/S_M
11.	SOKAL & SNEATH (1963)	$S_{SS} = \sqrt{\frac{b+c}{n}}$	S_{RR}

TABLE 2

two plants, the same statement would be true - they co-exist in only two out of twelve places, and thus they are dissimilar. However, the argument used in ecology is that plant X exists in only two places and in both of these plant Y also exists, and thus there is a very strong relationship between X and Y, and just because plant X is rarer than Y, this should not be disregarded. Some measures show this strong relationship, e.g. in this example Q = 1.0 and $S_W = 1.0$.

As each coefficient which is suited for this special purpose should attain its maximum value on this example, we may eliminate some of the similarity measures from present consideration. In fact, four measures reach their most similar value in this case, Q, C_8, S_W and Preston's measure, which is insoluble for b or c equal to zero, but as b or $c \longrightarrow 0$, then $Z \longrightarrow 0$ which is complete association, since Z is a dissimilarity measure. The graphs of 1-Z and S_W are shown in Figures 4 and 5 for the case a+b+c = 10. Since Q and C_8 are dependent on d, two graphs are shown for the case a+b+c = 10 with d = 4 and d = 8, these are in Figures 6 to 9.

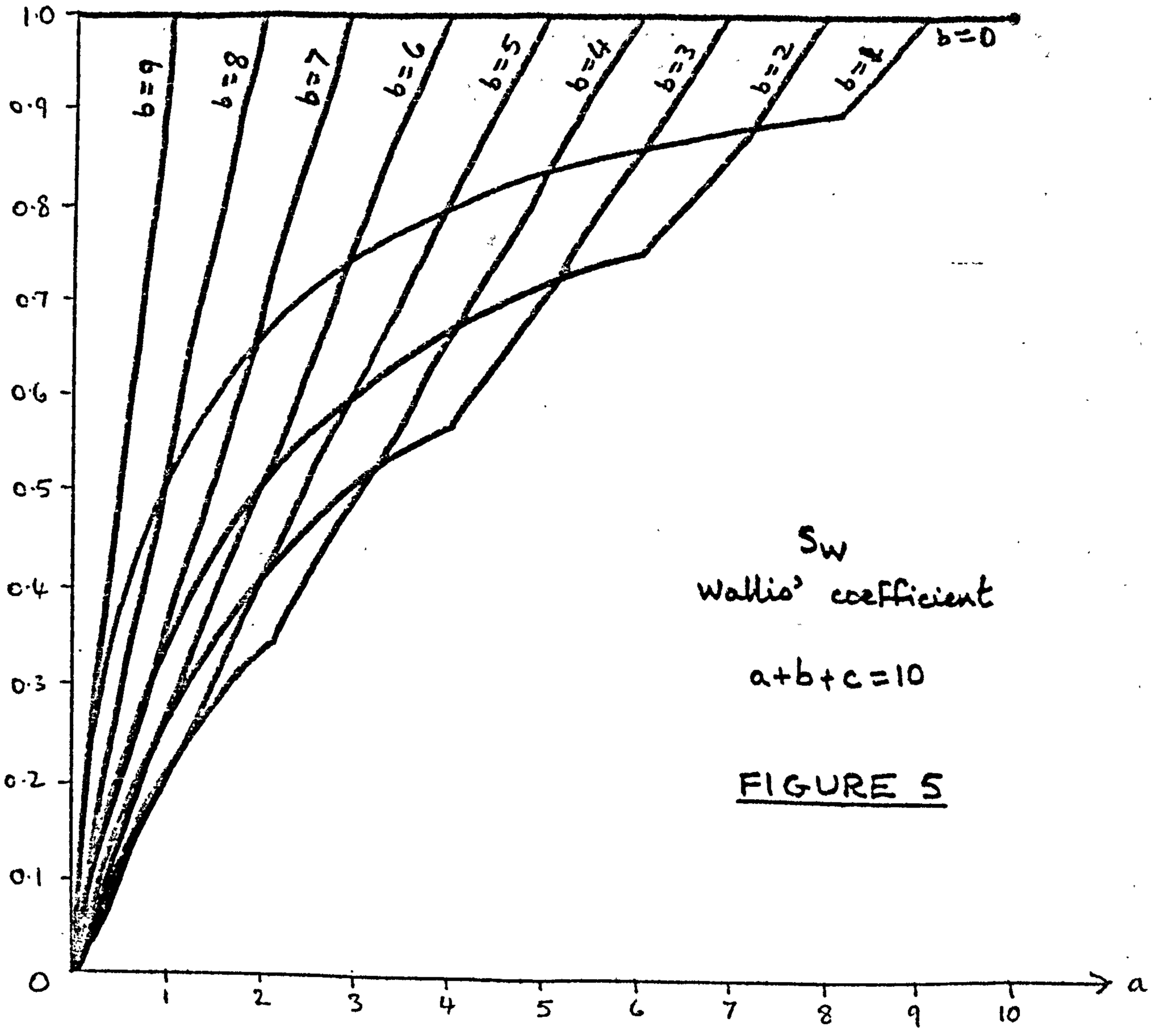
As can be seen the graphs for Q, S_W and 1-Z are very similar with the main differences occuring when a and b are low.



1-2
Preatons coefficient

a+b+c=10

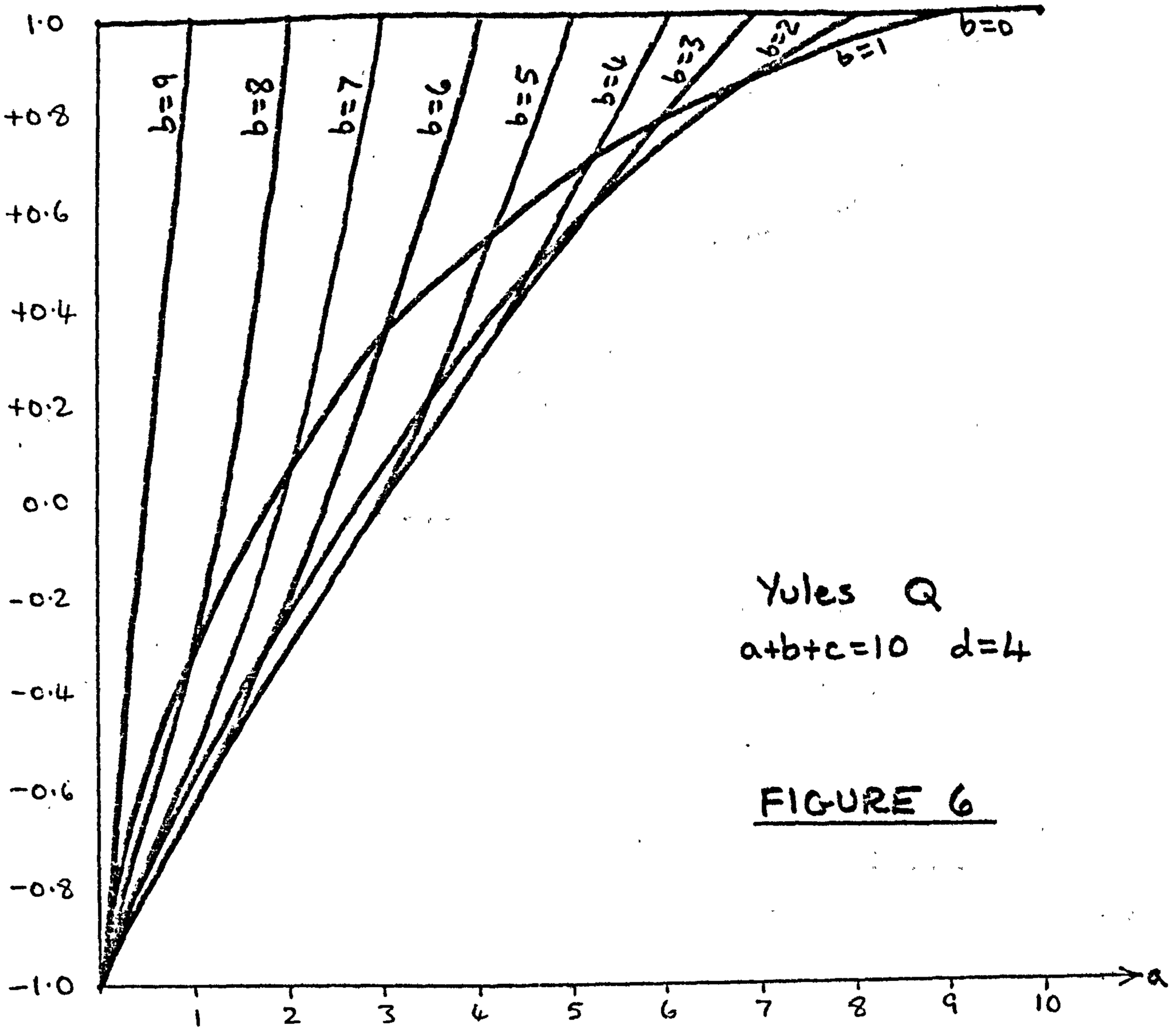
FIGURE 4



S_W
Wallis' coefficient

a+b+c=10

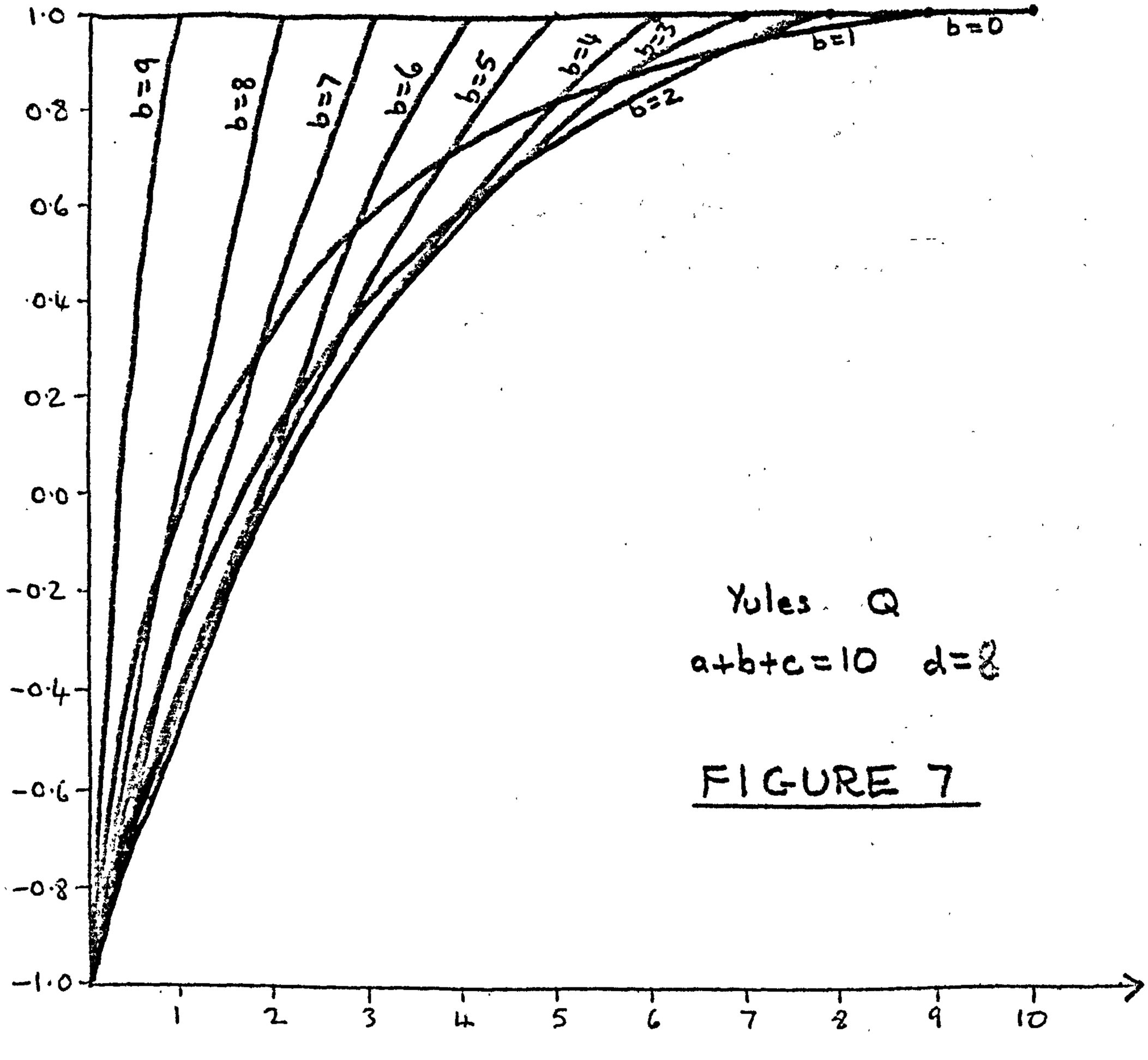
FIGURE 5



Yules Q

a+b+c=10 d=4

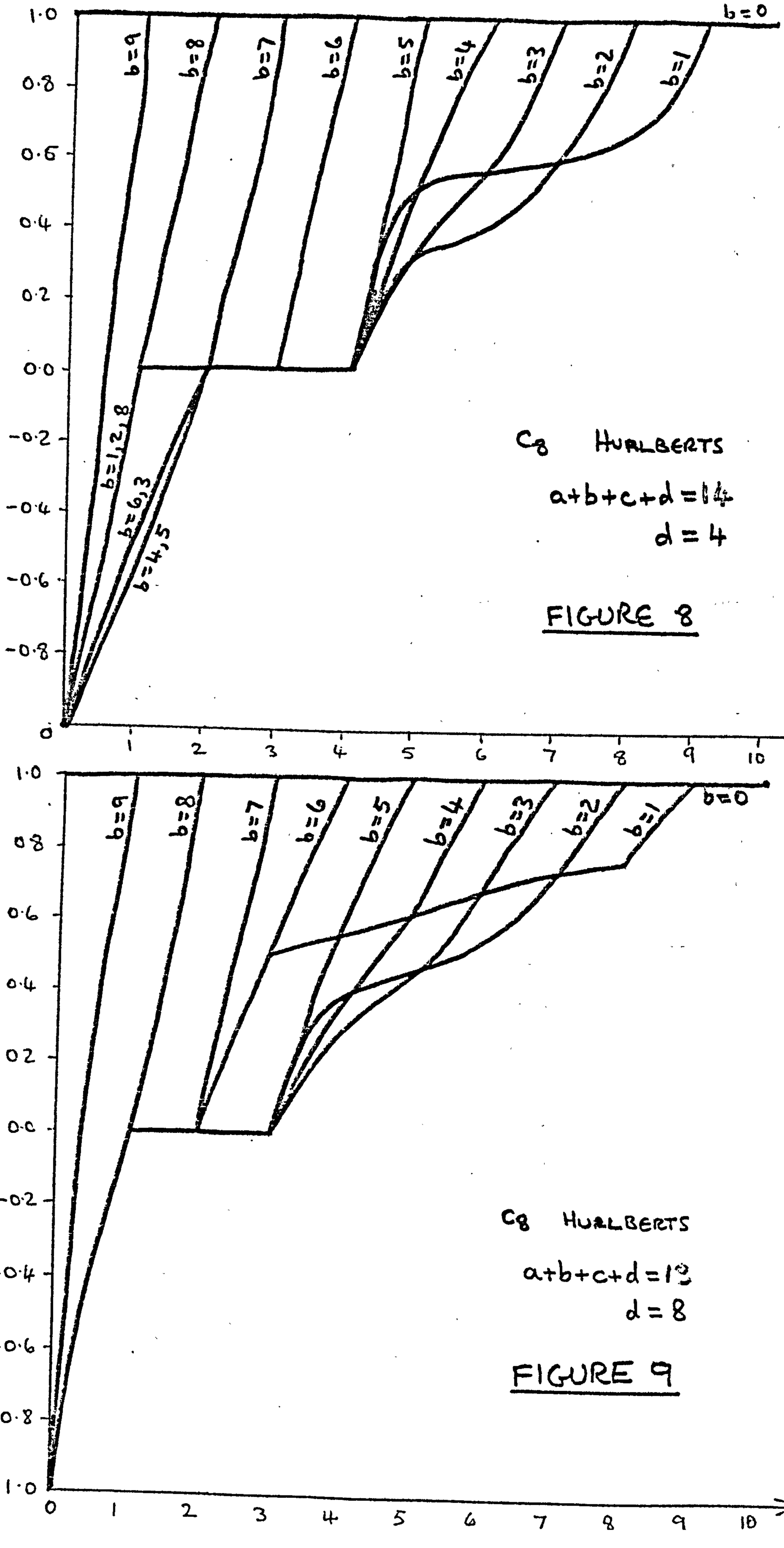
FIGURE 6



Yules Q

a+b+c=10 d=8

FIGURE 7



C_8 HURLBERTS

$a+b+c+d=14$

$d=4$

FIGURE 8

C_8 HURLBERTS

$a+b+c+d=12$

$d=8$

FIGURE 9

However the calculation of Z is difficult, and best performed graphically (although a partial table is given in Preston's paper) - this is a disadvantage. From the graphs of C_8 the main point of note is the discontinuity at zero similarity - this appears to be due to the disadvantage of χ^2 as a measure of association, but illustrates the use of the measure for testing independence, since extreme values are only assumed for extreme values of a. Thus in choosing between the four measures, S_W and Q seem the most useful in practice for this type of application.

Figures 10 to 22 show the other coefficients for our specified examples. It can be seen that S_C is very similar to S_{FO}, which is not surprising as their equations are similar, and both of these coefficients resemble the behaviour of Q, but scaled according to the value of n. S_{K2} is very similar to ϕ and S_{DA} is also of the same type, but with straight lines for fixed values of b. S_{FA} hardly varies at all with b for fixed a.

S_M is very similar to ϕ, but decreases slightly for high values of a, in fact in the two cases below:

a	0
0	d

a-1	1
0	d

the second case gives a higher similarity than the first for all cases where $a < a^2 - d^2$. This is contrary to the intuitive notion of similarity.

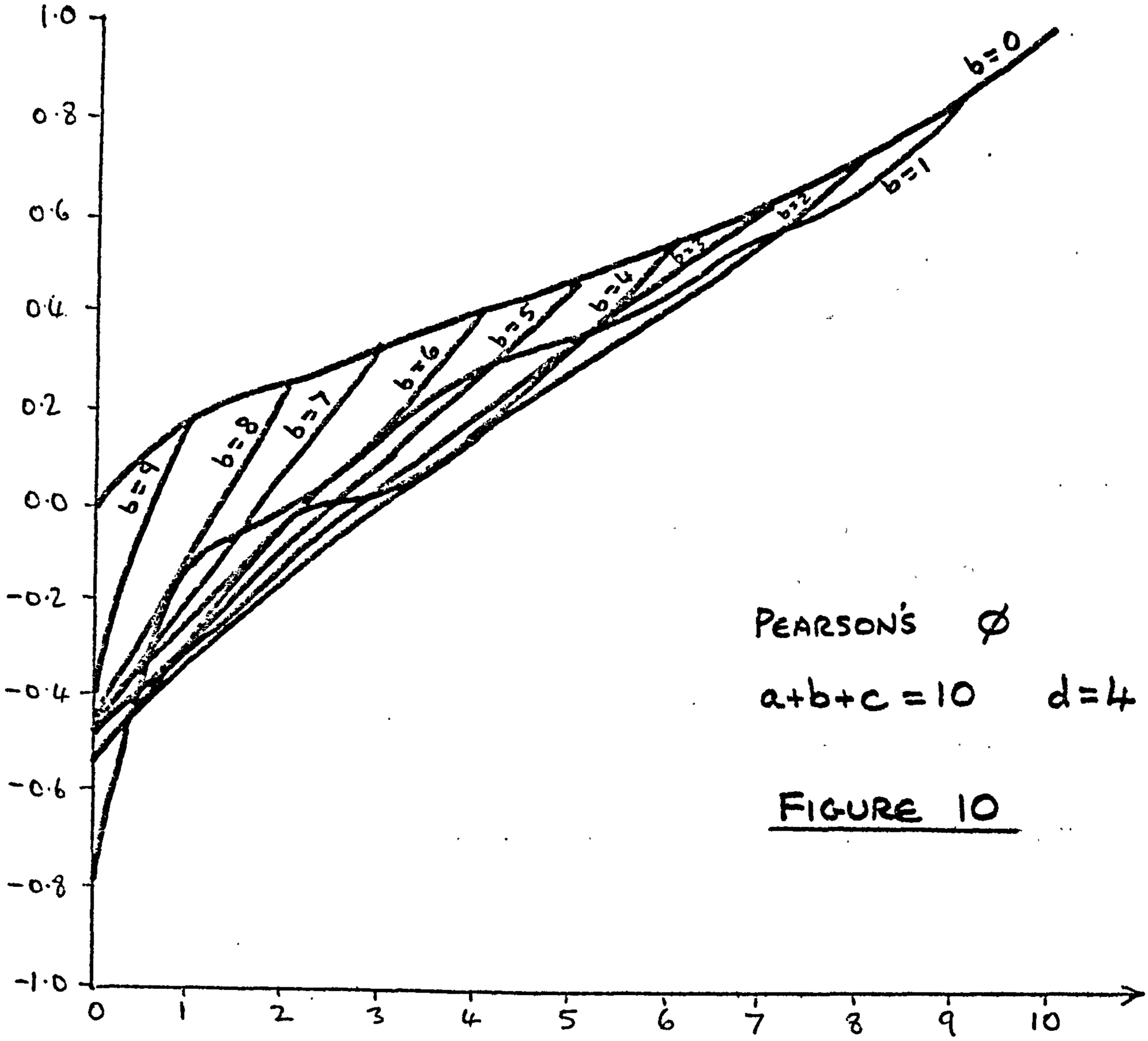


FIGURE 10

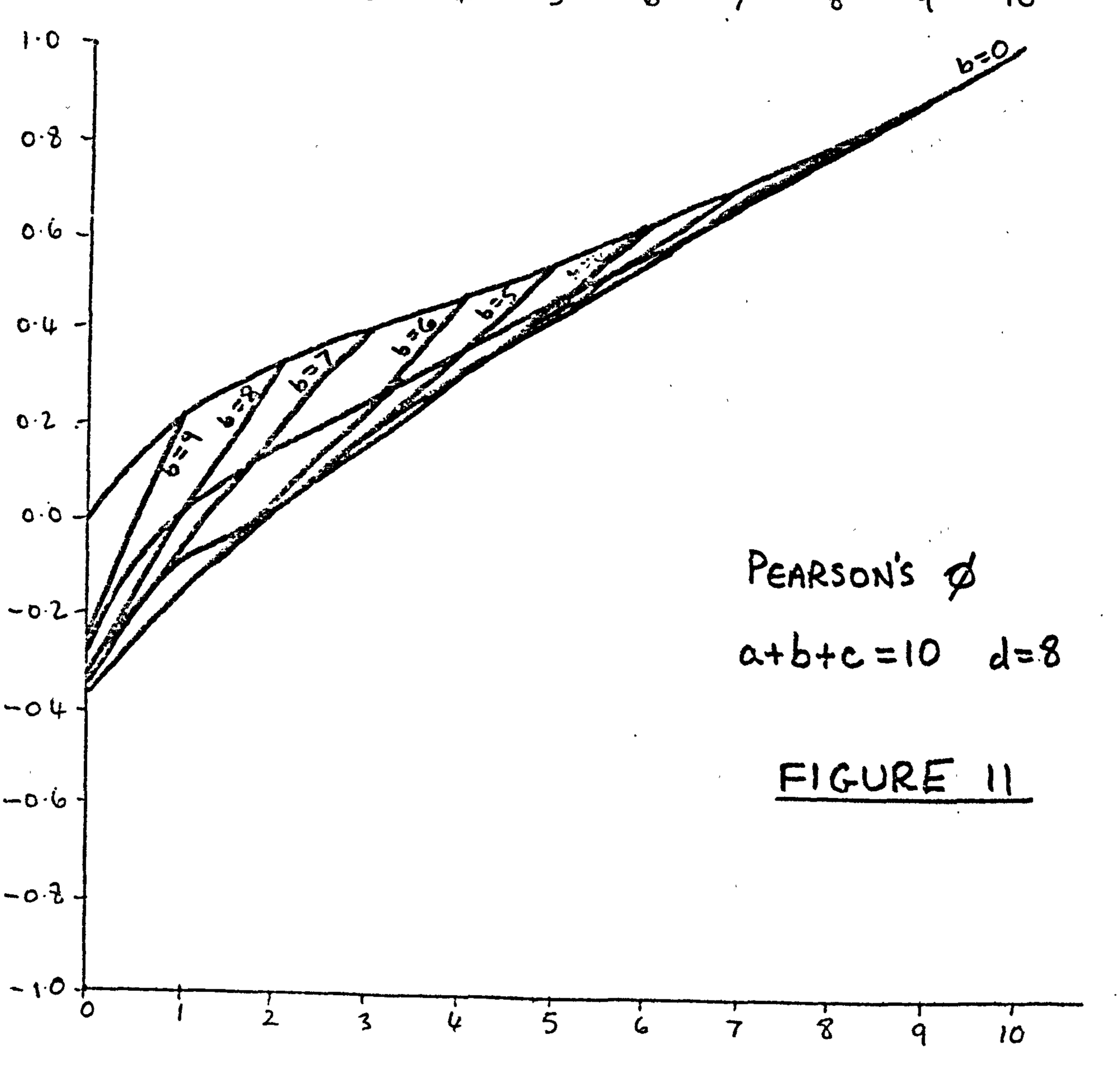


FIGURE 11

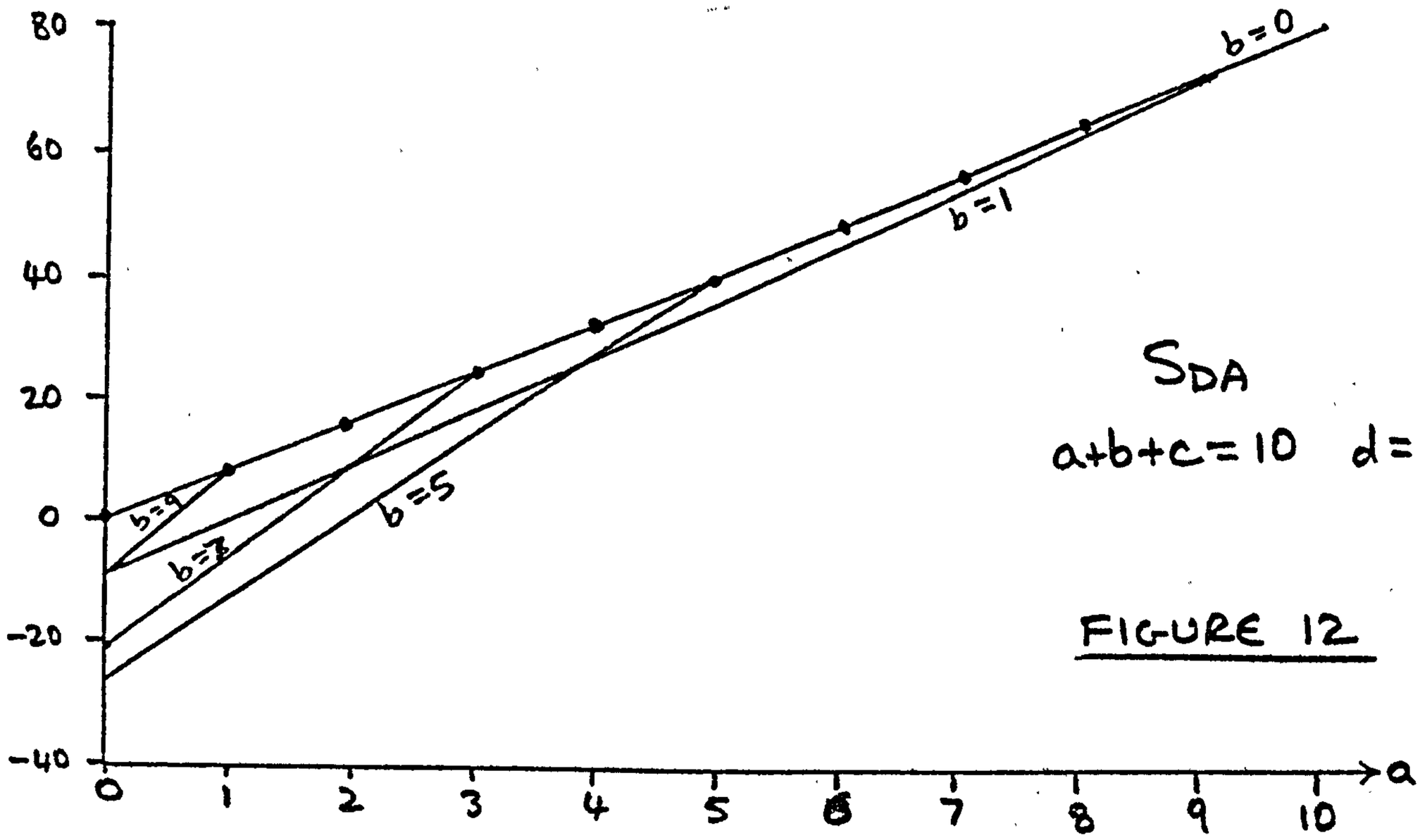


FIGURE 12

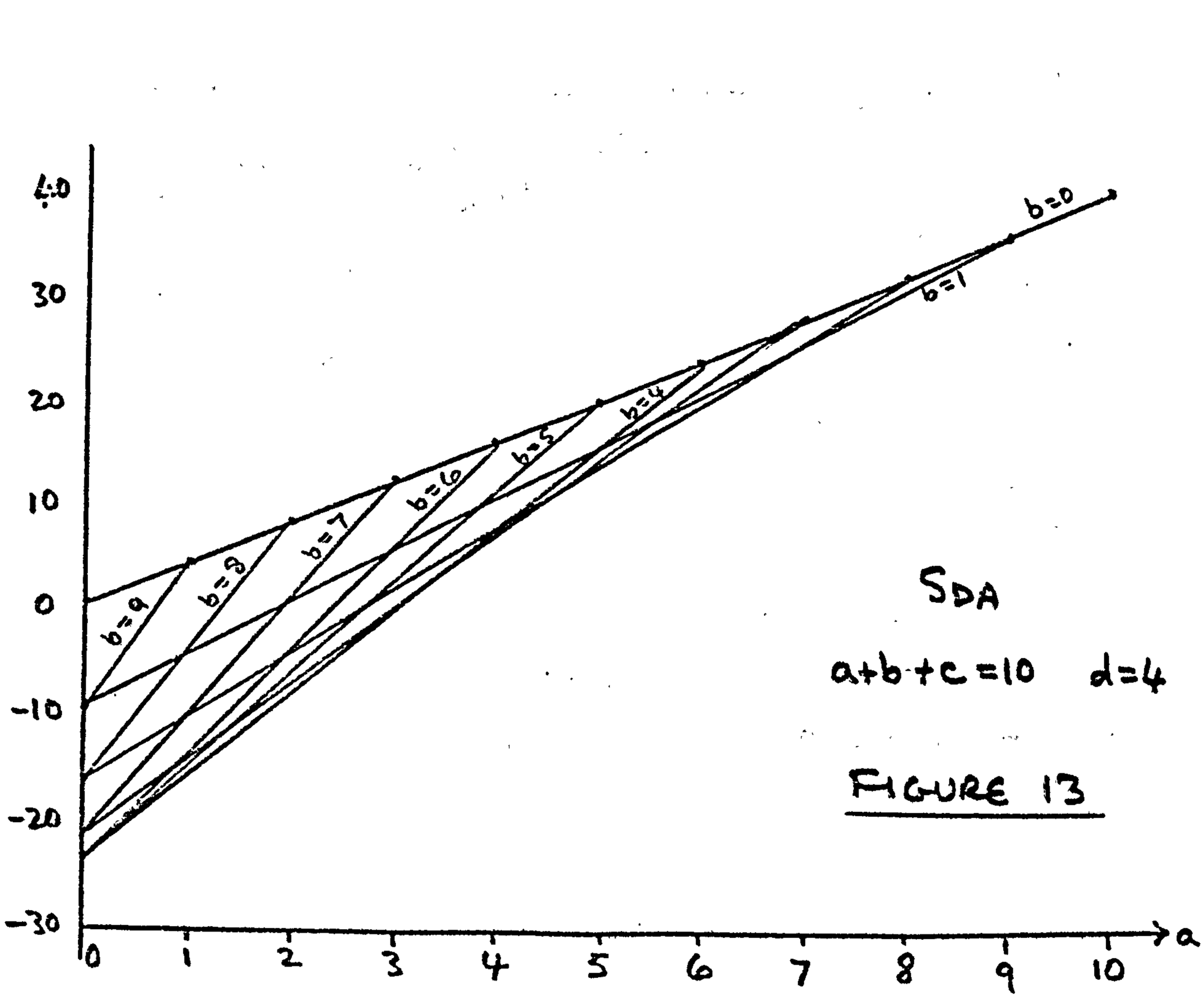


FIGURE 13

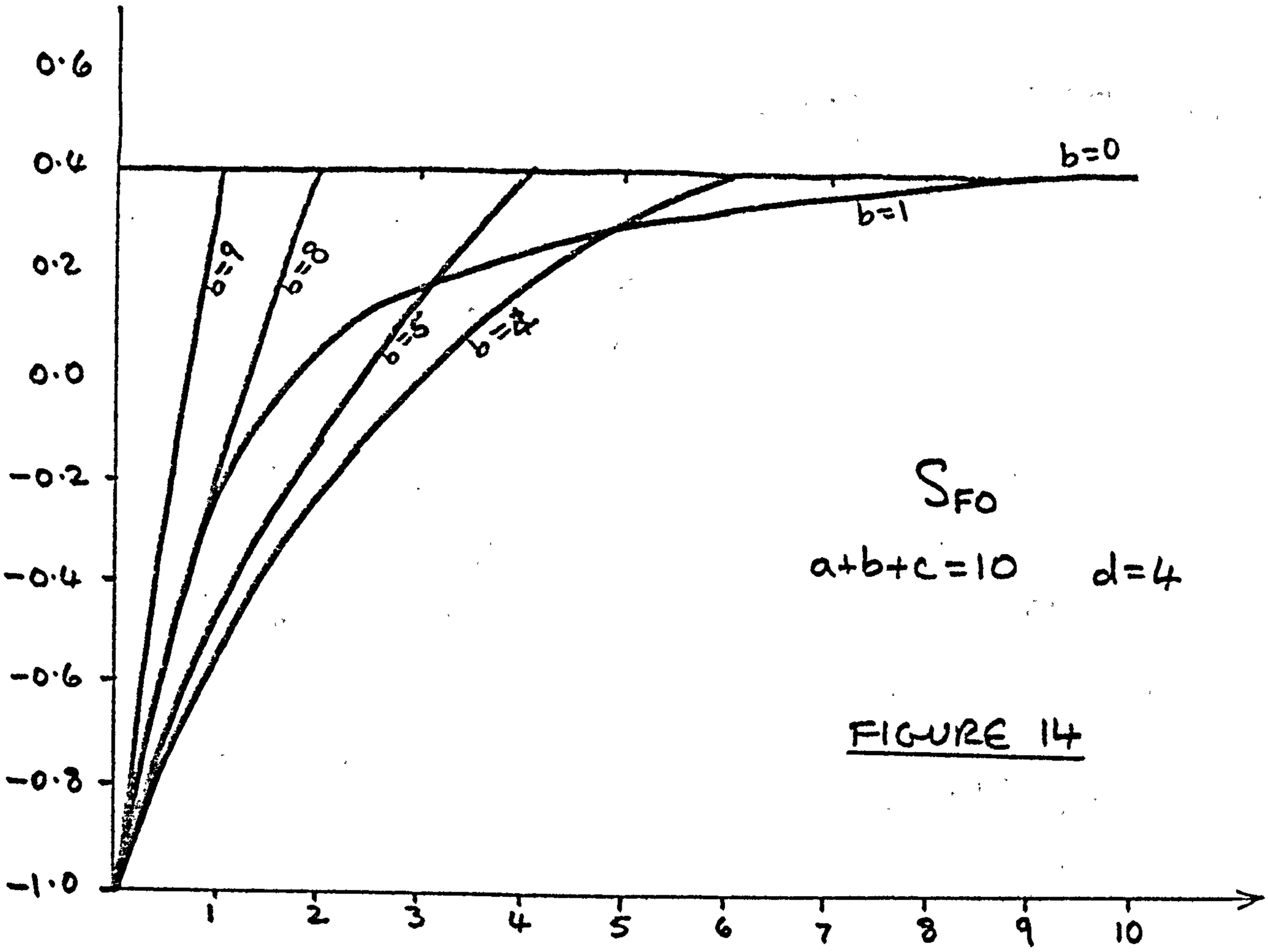


FIGURE 14

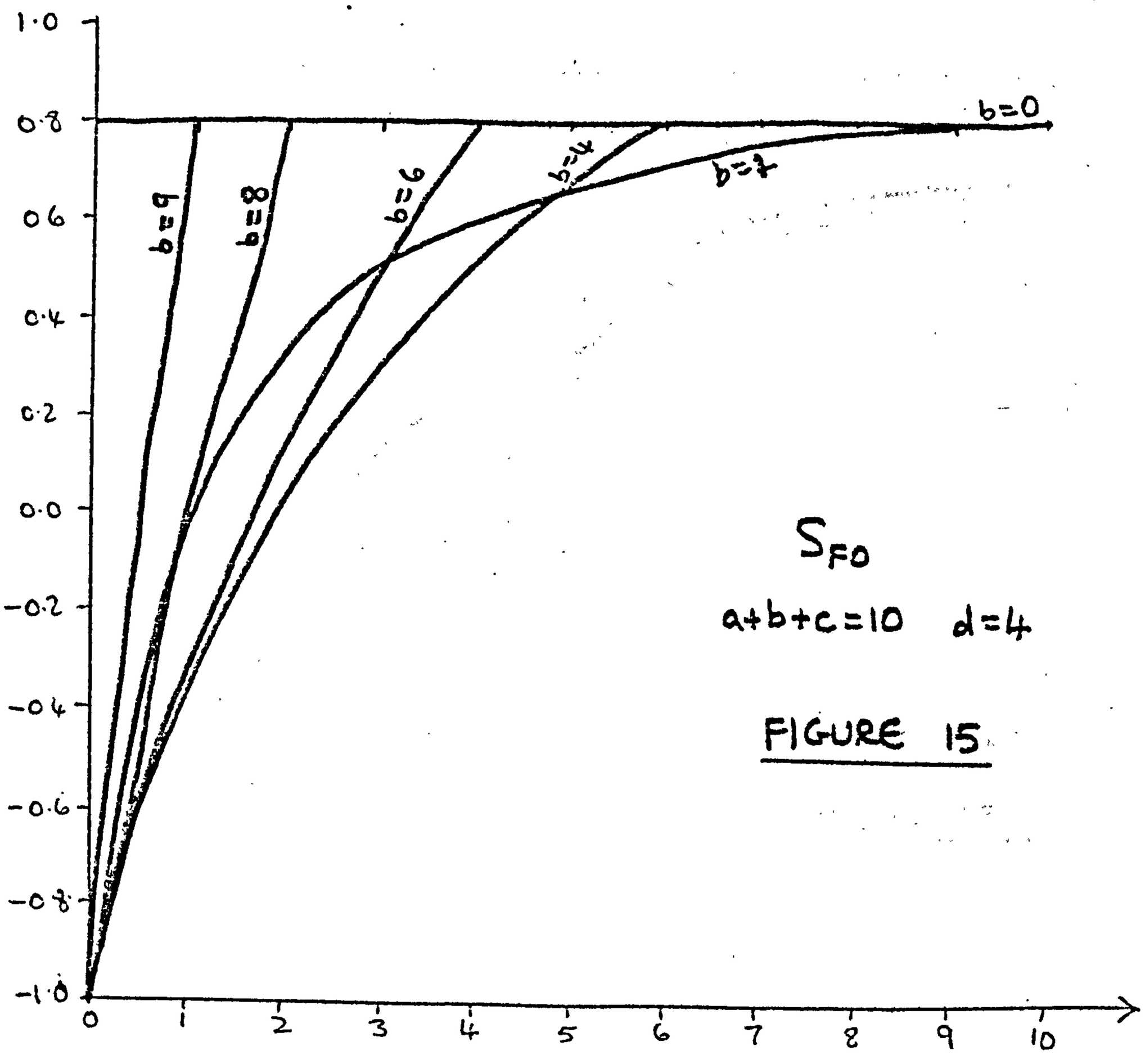


FIGURE 15

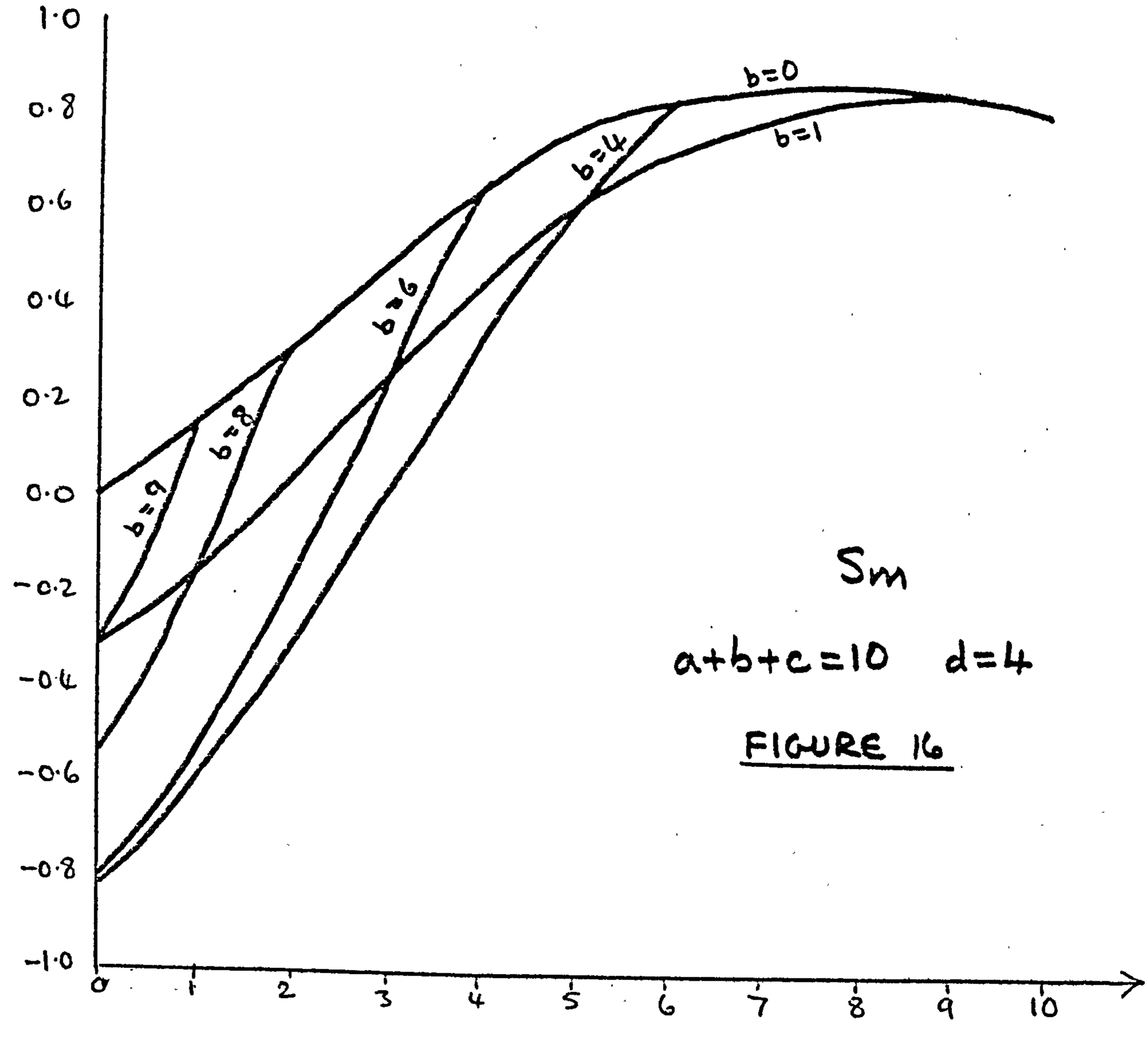


FIGURE 16

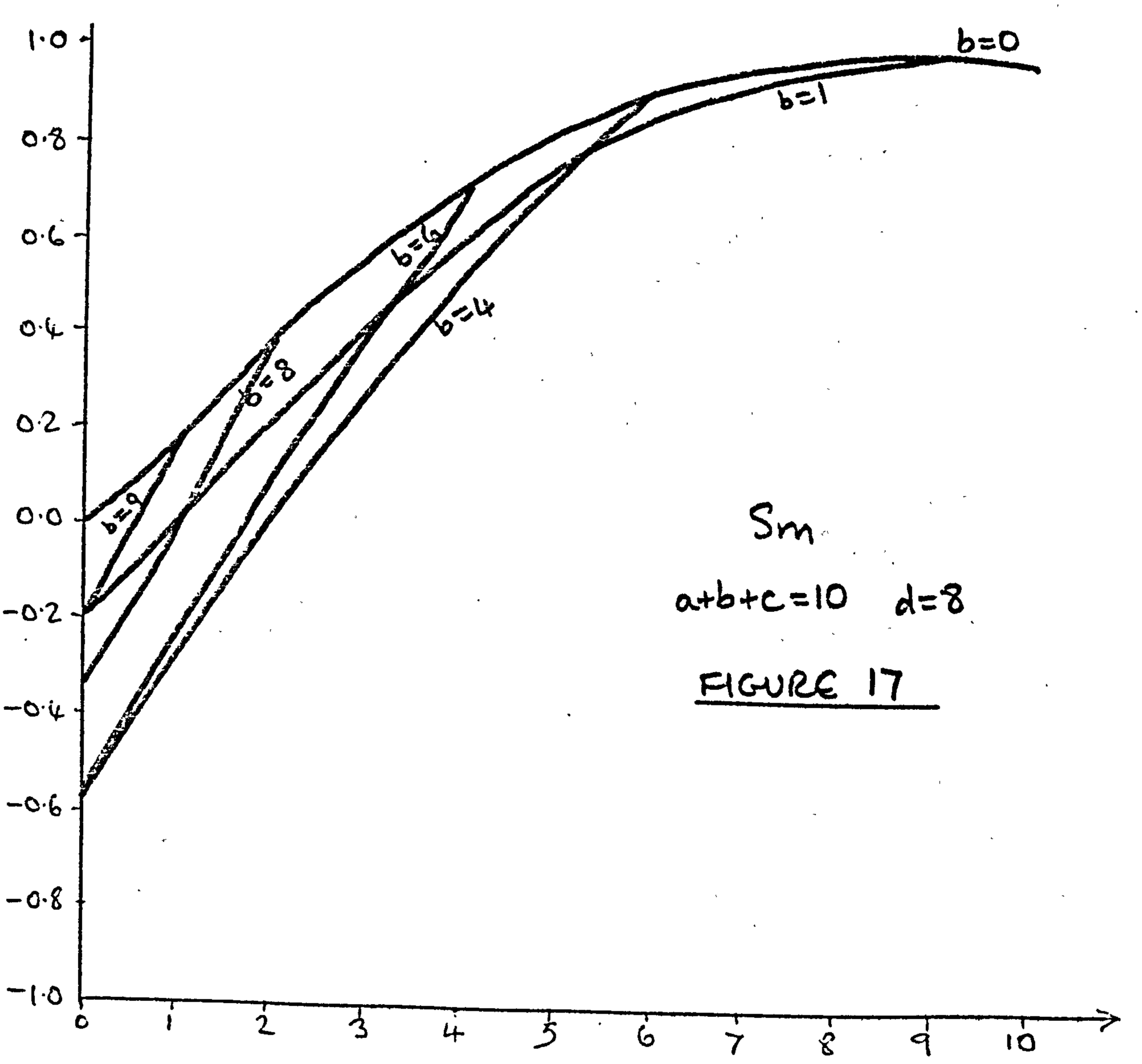


FIGURE 17

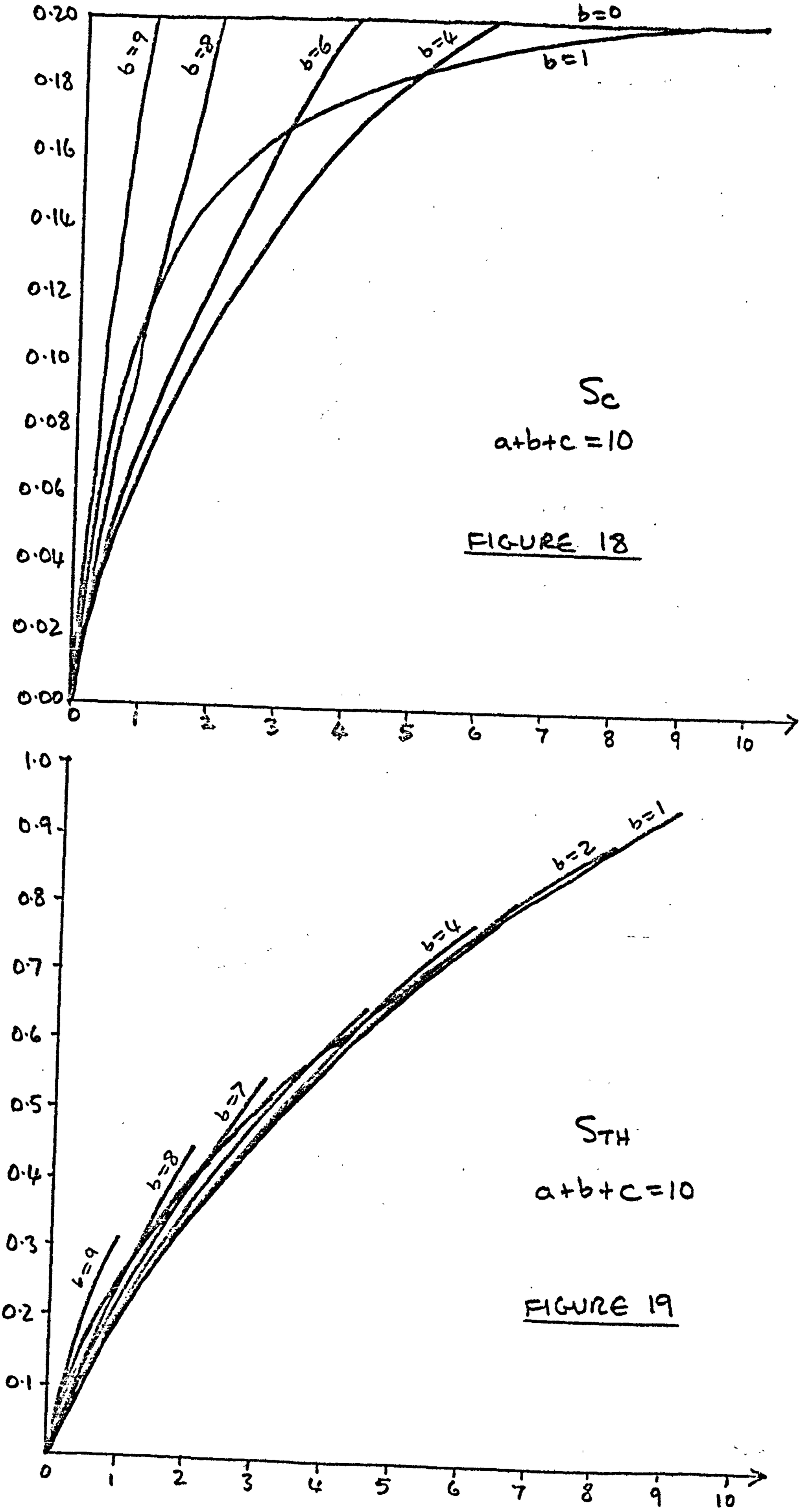
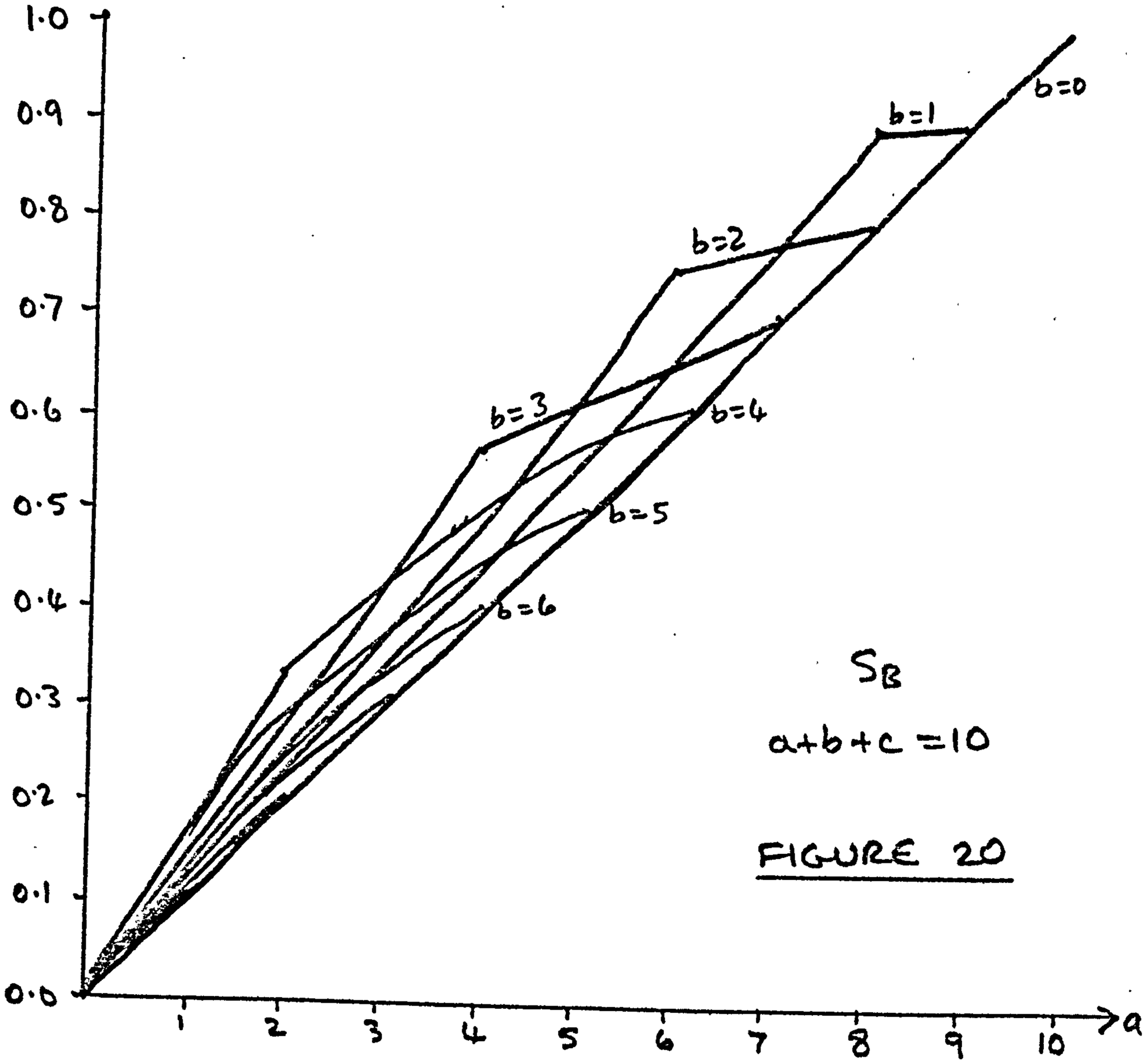


FIGURE 18

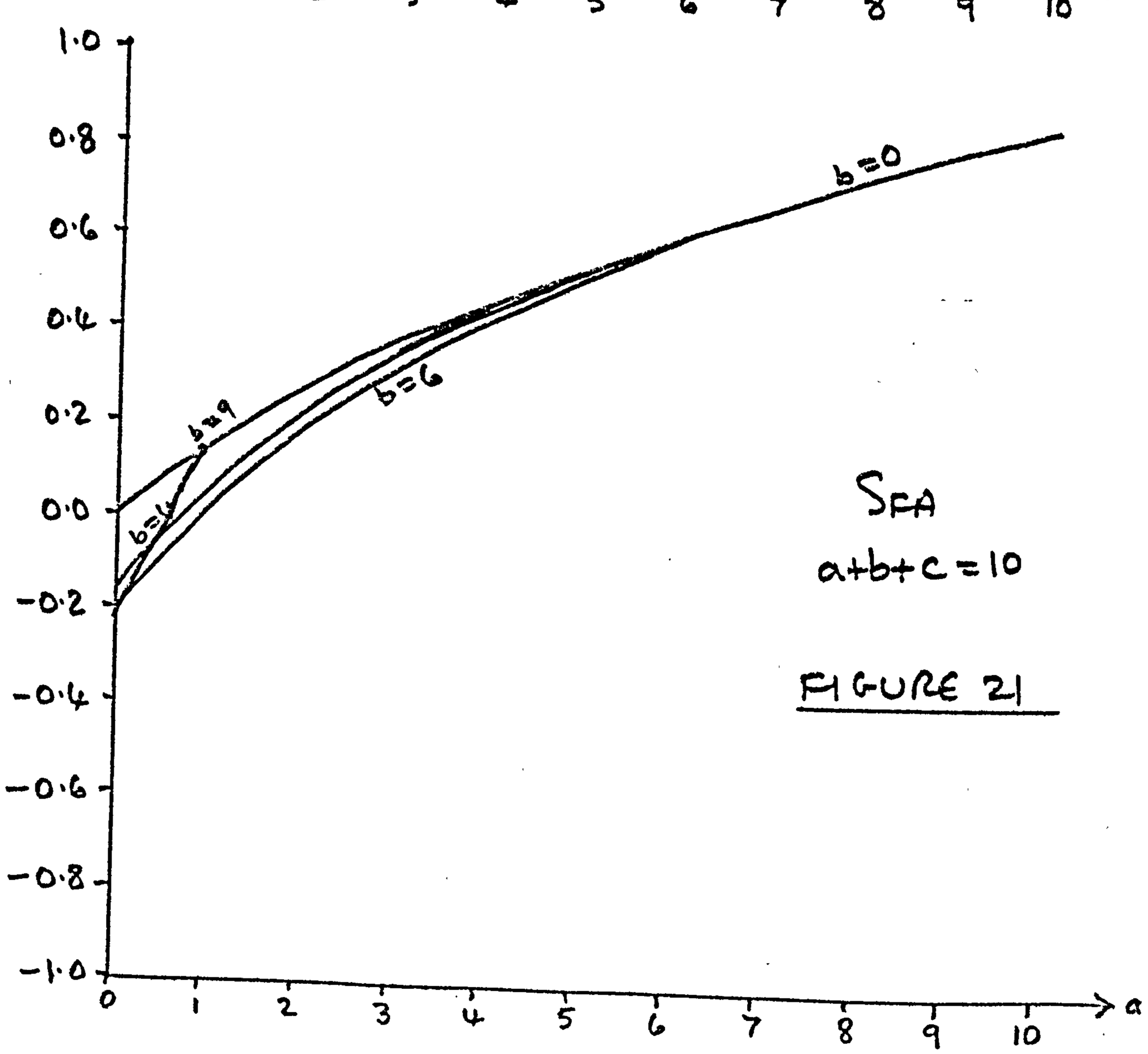
FIGURE 19



S_B

$a+b+c=10$

FIGURE 20



S_{FA}

$a+b+c=10$

FIGURE 21

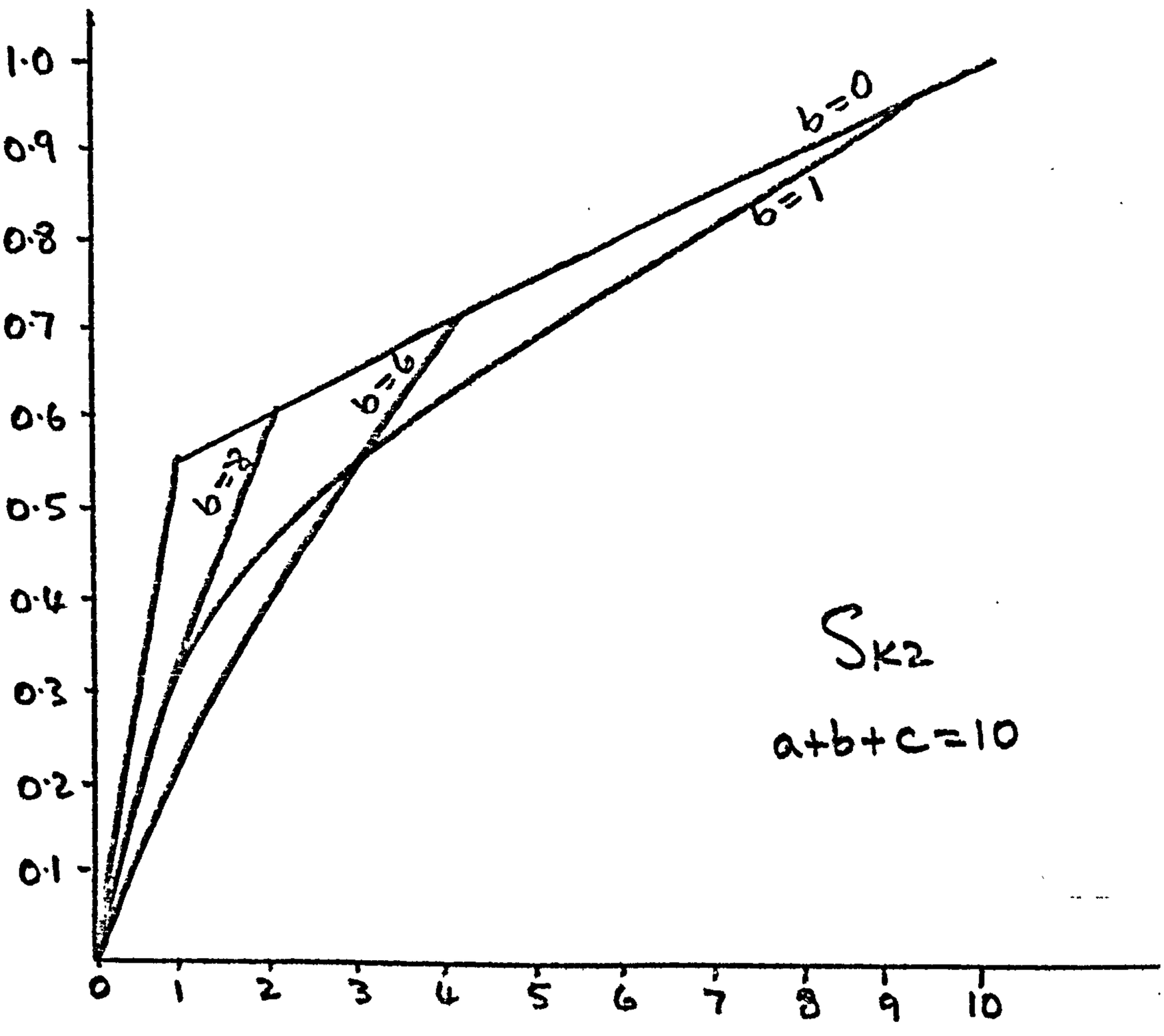


FIGURE 22

Consider the two cases:

a	x
0	d

A

a	x-y
y	d

B

the similarity measures can be split into two groups, depending on which of these examples they give the higher similarity. Only two measures give table B the higher similarity, these are S_B and S_{TH}.

Conclusions

The choice of similarity measure depends on the application. Important questions of the relative importance of a and d, for example, are a useful approach. We tabulate below the methods under such questions.

1. Can you reverse variables?
   YES - S_{SM}, S_T

2. Has d the same importance as a?
   YES - S_{SM}, S_T, ϕ, Q, S_{DA}, S_M, C_8

3. Can you move 1 element from b to c without affecting similarity?
   YES - S_{SM}, S_{RR}, S_J, S_D, S_T

4. Can you use the measure in our ecology case?
   YES - ϕ, C_8, S_W, Z

5. Is the method based on χ^2?
   YES - ϕ, C_8

6. Is the similarity affected by a new variable on which both objects score 0?

   NO - S_J, S_D, S_C, S_{TH}, S_W, S_B, S_{FA}, Z

7. Does the similarity always decrease if a is decreased by 1 and b increased by 1?

   NO - S_M

These type of questions should limit the choice of measure to a few, which will probably from the graphs shown turn out to behave similarly, so that the final choice is less important. Another consideration in the choice of measure is the range in which the values of a, b, c and d lie - for example if b and c are roughly the same in a study then questions such as those which occurred in our ecology example are of less importance.

In the discussion we have suggested that measures based on χ^2 may not be useful measures of association. We have also given examples where the behaviour of S_M, S_B, and S_{TH} appears not to be in agreement with what is intuitively acceptable.

It should be noted that while we have dismissed measures which are functions of others from our discussion, it may be that for cluster analysis or ordination a simple function of a measure may perform better than the original measure itself.

2. Unorderable Multi-State Measures

We have discussed data types in Section B.2 and the problems of this particular type have already been considered. If we are unable to 'improve' the data by measuring similar attributes which are orderable, or by forcing an ordering on the data, then we may proceed in several ways.

We could score objects on two attributes by 1 for a match and 0 for a mismatch. For example Rogers and Tanimoto use this type of approach and use the expression -

$$\frac{\sum x_{it}y_{it}}{\sum x_{it} + \sum y_{it} - \sum x_{it}y_{it}}$$

(where $x_{it} = 1$ if object x possesses state t of attribute i).

However matches are rarer than in the binary case, and it may thus seem logical to weight matches accordingly, also a match on a six-state variable would be in general rarer than a match on a three-state variable. Thus we could weight a match on a particular variable by the number of possible states. We would then have the similarity measure -

$$\frac{\sum_i S_i \sum_{t=1}^{S_i} x_{it}\, y_{it}}{\sum_i S_i}$$

(where S_i is the number of states in the i^{th} attribute).

One of the earliest measures due to Smirnov is even more complex, weighting states by their varity. The logic behind the method is discussed by Sokal and Sneath (1963), and

Smirnov (1968); the formula may be expressed by -

$$\frac{n}{\sum S_i}\left[\sum_i \sum_t \frac{x_{it}y_{it}}{\sum_{all\ Z} Z_{it}} - \sum_i \sum_t \frac{(x_{it-1})(y_{it-1})}{\sum_{all\ Z}(Z_{it-1})}\right] - 1$$

(where $Z_{it} = 1$ if object Z possesses attribute i in state t, otherwise $Z_{it} = 0$).

As we do not have one particular state which may dominate a in presence/absence data, we may use any of the binary measures previously mentioned which include a and d, by scoring a match as 1. As in the above two examples of possible measures, the weighting of attributes and states is one of the chief difficulties.

3. Ordered Multi-State Measures

With ordered multi-state data one could simply ignore the ordering and analyse as if it were unordered, this would of course reduce the information content of the data. Another approach would be to estimate the interval between states and analyse as if the data were continuous, or one could repeat the analysis with different between state intervals. However this introduces assumptions which may not be valid.

A more rigorous approach would be to use non-metric scaling methods to produce a scaling of the data which 'best' approximates to continuous data. Since ordered data is largely found in psychology, the methods of psychological scaling are those which are most used in practice.

4. Continuous Data

The two measures of similarity which are most commonly used are the correlation coefficient -

$$c_{xy} = \frac{\sum(\bar{x} - x_i)(\bar{y} - y_i)}{\sqrt{\sum(\bar{x} - x_i)^2 (\bar{y} - y_i)^2}}$$

and Euclidean distance, which in n dimensions is given by an extension of Pythagoras' theorem -

$$d_{xy} = (\sum(x_{ij} - x_{ik})^2)^{\frac{1}{2}}$$

There has been much discussion on the relative merits of these measures in numerical taxonomy. Eades (1965) has stated:

> "the correlation coefficient is inappropriate as a measure of taxonomic resemblance",

whereas Boyce (1969) in a study involving five measures of resemblance including the two above, states:

> "Of the five coefficients, the correlation coefficient reacts to the components of size and shape in a way most similar to that which the taxonomist usually adopts".

The usual argument against correlation is that it ignores the magnitude of variables, and the argument against Euclidean distance is that it ignores the 'slope' of the vector of variables for each object.

Euclidean distance has difficulties where correlations occur between variables. In the example given by Boyce (1969) measurements were taken of a number of primates and

with Euclidean distance since females tend to be much smaller than males, they were often misclassified. This is caused by all variables being dependent on the overall size of the animal, thus in using correlation good results were found. Alternatively one could reduce this size factor to only one variable by either principal components analysis or by simply dividing all lengths by the animals overall height, or by normalizing the variables vector for each animal, and then Euclidean distance could be used (this is in fact an angular distance measure).

One difficulty with Euclidean distance is that as shown above it depends a great deal on the scaling of the data. Also a large difference on one variable can cause the distance to be large.

Both measures have their uses and the choice depends almost entirely on the application.

Other distance measures have been proposed for continuous data, for example the average Euclidean distance -

$$\frac{1}{n} \sum |x_{ij} - x_{ik}|$$

There is also the set of measures called the Minkowski metrics -

$$\left(|x_{ij} - x_{ik}|^{\frac{1}{\lambda}} \right)^{\lambda}$$

where $\lambda=1$ gives the so-called City-block metric, and $\lambda=2$ gives Euclidean distance.

An often used alternative to simple Euclidean distance is the square of this. The squaring emphasizes the large distances, but the space then becomes non-metric.

Penrose (1954) has suggested the splitting of Euclidean distance squared into two components, one which measures 'size' and the other 'shape'. Thus we have

$$\frac{1}{n}\sum d_i^2 = \underbrace{(\frac{1}{n}\sum d_i)^2}_{\text{SIZE}} + \underbrace{(\frac{1}{n}\sum d_i^2 - (\frac{1}{n}\sum d_i)^2)}_{\text{SHAPE}}$$

The first term is the square of the average distance and the second term is proportional to a coefficient used by Zarapkin (1934). However as pointed out by Rohlf and Sokal (1965) the shape coefficient ignores additive size difference but not proportional size difference.

Pearson's coefficient of racial likeness (1926) is of similar form to Euclidean distance. It is

$$\left\{\frac{1}{n}\sum\left(\frac{\bar{x}_{ij} - \bar{x}_{ik}}{\frac{s_{ij}^2}{n_j} - \frac{s_{ik}^2}{n_k}}\right)\right\}^{\frac{1}{2}} - \frac{2}{n}$$

which is a measure of the difference between two groups. This measure has come in for criticism from several authors. Seltzer (1937) has discounted the measure since it ignores intercorrelations, and varies with the number of characters and the size of the sample, he describes the measure as:

"nothing more than a test of significance".

Cronbach and Gleser (1953) state that the coefficient of

racial likeness

> "proved unsatisfactory in the anthropological research for which it was developed".

One of the problems with distance measures such as these is that if the original variables have intercorrelations then overweighting can occur. This can be mathematically 'extracted' by using the Mahalanobis D^2 statistic, or by a preliminary principal components analysis, but a large amount of the correlation may come from clusters present in the data (see Gower 1969).

Other correlation measures have been little used, such as tetrachoric or rank correlation. The main difficulty with rank correlation, although it has useful distribution-free properties and has been recommended for use in discriminant analysis by Kendall (1966), is that it can reduce the distance between clusters and thus make them less distinguishable.

The cosine measure (sometimes referred to as angular separation) has been used. It is independent of proportional differences in size, but not additional differences. It is

$$\frac{\sum x_{ik}x_{jk}}{(\sum x_{iv}^2 \sum x_{jv}^2)^{\frac{1}{2}}}$$

χ^2 measures can be used directly with continuous data. Measures such as χ^2 with Yates correction, $\sqrt{x^2/n}$ and the Coefficient of Contingency $C = \sqrt{\frac{x^2/n}{1+x^2/n}}$ have all been used.

Also based on χ^2 is the pattern similarity coefficient of Cattell, Coulter and Tsujioka (see Cattell 1949) -

$$r_p = \frac{E_i - \sum d^2}{E_i + \sum d^2}$$

where E_i = twice the median χ^2 for i d.f.

A heuristic similarity measure proposed by Bray and Curtis (1957) and used in several American papers (see Hole and Hironaka 1960) is

$$S_{ij} = \frac{2 \sum_k \min (x_{ik}, x_{jk})}{\sum_k x_{ik} + \sum_k x_{jk}}$$

this has been used as a distance measure by the transformation $d_{ij} = 1-S_{ij}$, this gives

$$d_{ij} = \frac{|\sum x_{ik} - \sum x_{jk}}{\sum x_{ik} + \sum x_{jk}}$$

This is of similar form to the Canberra metric

$$\sum \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$$

which is based on Kulczynski's $\frac{2a}{2a + b + c}$

Mixed Data Types

The problems when data is of mixed types (e.g. some binary and some continuous) has been briefly considered in Section B.2.

The problem can be solved by either moving data up a category and assuming a property which it may not possess, or by moving down a category and losing information.

Another possibility is to evaluate several similarities, using one data category for each and to employ a weighted average to determine overall similarity.

If unorderable multi-state and binary data are mixed then one can proceed by using the measures of the former, since binary data is a special case of this. Similarly, binary data is a special case of ordered multi-state data. Also suitably scaled binary data can be considered along with continuous data.

Information Measures

The use of information theory in classification is new, only having been introduced in the last ten years, but it is now to some extent fashionable. Perhaps the first suggestion of the use of information measures comes in Rescigno and Maccacaro (1961), but these were not used until much later by Williams and Lambert (1966), Lance and Williams (1967).

In information theory we are concerned with the receipt of particular messages from a source. We call p_i the probability of a particular message i being received. The probability is lower, the more unexpected the message and the more information received. Based on this and other considerations we can define (Coombs et al 1970) information as $I_i = -K \log p_i$ and for convenience one takes $K = 1$.

The <u>expected information</u> conveyed by a set X of several possible messages is called the <u>uncertainty</u> or <u>entropy</u> of the message set.

$U(X) = -\sum p_i \log p_i$ from the definition of expectation.

This can be used as a measurement in numerical taxonomy. If we consider a set of binary measures on n individuals, then the two states have probabilities p_i and $1-p_i$ and so -

$$U(X) = -\sum_i \left(p_i \log p_i + (1-p_i)\log(1-p_i) \right)$$

If a_i individuals possess the i^{th} attribute then the best estimate of p_i is a_i/n. Thus we define the <u>information content</u> of the set X

$$I = n.U(X) = \sum n \log n - \sum \left(a_i \log a_i + (n-a_i)\log(n-a_i) \right)$$

This expression becomes zero if all objects are identical, and increases as points become more dispersed.

Information content is additive so we can obtain the gain in information content if two sets are combined by simply adding the information content of each set.

However entropy is a measure of disorder and both a clustered distribution and a regularly spaced distribution have low values of uncertainty (Sneath 1969). Also the use of the method with centroid sorting (the information-analysis method of Williams and Lambert 1966) has a tendency for equal sized clusters to be found.

The theory of information is covered in Theil (1967) and Coombs et al (1970). The relation to classification is discussed in Boulton and Wallace (1970) and Wallace and Boulton (1968), and it is shown how continuous variables may be incorporated.

Multi-Level Data

The classification of objects sometimes occurs where for each variable we have a string of values, and thus we have a two-dimensional matrix for each object. For example we might have a set of companies and measures on these such as turnover, profit, liquidity, etc., each of which is measured on several years. The difficulty can be partially resolved by considering the n by m matrix as a $n*m$ vector and proceeding as in the normal case.

A related problem occurs in soil classification, where for particular geographical points soil samples are taken at various depths. The soil sample is analysed for chemical composition and thus for each point there exists a matrix of depth against composition. The problem is that the depths at one point may not correspond to the depths at another - because of land-slipping, faults, etc., extra strata may exist at one point, or be entirely missing. Rayner (1966) has suggested forming a similarity matrix for each pair of points and taking the levels with highest similarity thus forming a set of data which can be analysed in the normal way. As Lance and Williams (1967) have pointed out, if this were used in our companies example then one firm whose

profits etc. were declining, and one with increasing variables would be measured as very similar.

The problem of not being able to match soil samples at the same level is a very similar problem to that of homology in biology where two species of animal which derive from a common ancester have several characteristics the same although the size, relative position of organs, etc., may be quite different. These animals are said to be homologous. Jardine (1967) has discussed homology and its relation to taxonomy.

Comparisons of Measures

Comparison studies of measures on data are as yet very few. One major difficulty is that the results of an analysis are heavily dependent on the method used.

Boyce (1964, 1969) has compared five similarity measures on one set of data - correlation coefficient, cosine, Euclidean squared distance, the shape coefficient of Penrose and the mean character difference. The data was normalized for each variable. He ranked the five measures in order of success in depicting the true relationship between the objects; they were correlation coefficient (best), cosine, shape coefficient, mean character difference and squared Euclidean distance. The cluster method used was group average.

Rohlf and Sokal (1965) have compared the correlation coefficient and Euclidean distance coefficients giving a bivariate frequency chart of the values of r and the

corresponding Euclidean distance. This study confirmed the size and shape effects and stated that if objects varied greatly in size then the correlation coefficient should be used and only if this size effect could be extracted can Euclidean distance be employed. They conclude that both measures should be used and results compared.

Green and Rao (1969) compared eight measures on one set of data, by comparing the results at an eight-cluster level in the solution. The results were portrayed by using multi-dimensional scaling, the similarity between results being measured by the fraction of points classified in the same groups. The eight measures were Euclidean distance squared, Mahanobis D^2, Gower's log distance measure, Kendall's rank distance, City block distance squared, cosine, correlation and covariance. The results showed most measures giving similar results (covariance, Gower's and City block being identical) except for the correlation coefficient and Kendall's measure which had very different results.

C

CLUSTER ANALYSIS METHODS

1. General Discussion
2. Explanation and Discussion of Methods
3. Comparisons of Other Researchers
4. Choice of Methods for Study
5. Comparison of Methods
6. Conclusions

C.1 GENERAL DISCUSSION

Cluster methods can be divided into two types - monothetic and polythetic. Monothetic models are not strictly multivariate methods, they are divisive in operation, dividing into groups successively on single variables, deciding at each stage which variable is 'best' to split on. Basically they divide into two groups on the basis of that variable which has the most discriminating power. They have been designed mainly for binary data - with continuous or multi-state data we would have the problem of where the best cut-off point on that variable was, and also a particular variable may give a good division into more than two groups. Polythetic methods are the most common type, and are those with which we are primarily concerned. Polythetic methods do not single out particular variables but use them all simultaneously.

An advantage of monothetic methods is that the groups they produce have unique characteristics, and hence division of new items into the appropriate groups is simple.

Cluster methods sometimes have the property that an object can belong to two or more groups at the same time. These overlapping cluster methods are fairly uncommon, but the property of multi-group membership is useful in some cases. Monothetic methods cannot produce overlapping groups. We can thus divide methods into three distinct groups:

1. Polythetic, disjoint
2. Polythetic, overlapping
3. Monothetic

Further classification of methods into types is difficult because of the wide differences in approach to the problem by various authors in different fields. Sometimes methods are classified according to the type of data input - e.g. we might talk of binary methods - but many can be used or adapted for all types of data. The class of polythetic, disjoint clustering is by far the largest (and often only these are called cluster methods), and we can perhaps distinguish two special sub-types, so we divide into three classes:

(a) hierarchic methods

(b) relocation methods

(c) other types

Hierarchic methods are those which form a dendrogram. They have the advantage that the structure of the data can be displayed easily. They can also show sub-clusters and outliers more easily than most other methods, and in fact the tree diagram may be a requirement in itself. Hierarchic methods are normally performed by agglomerative or divisive algorithms, beginning with each object in a single cluster and gradually merging until all objects are in one cluster, or vice versa. These algorithms although optimal in some sense at each fusion or division stage, do not generally produce overall optima, in the sense of obeying Axiom 2' on page 17.

Relocation methods are those in which the membership of clusters varies by attempted improvements of the clusters. Objects are allowed to join other groups if this gives a

better value of the objective function being used. These cannot produce hierarchies, and are generally slower than the hierarchical methods. They produce an optimum at each stage, but these can be local optima.

These two categories do not overlap, but methods exist which it is difficult to determine if they come into these categories or not - for example, methods may produce partial hierarchies, or tree forms similar to dendrograms.

Of the methods which are not of the hierarchic or relocation type, some are heuristic approaches which were originally designed for hand computation, others involve partial enumeration of possible groupings and mathematical programming has been used - the approaches are numerous.

Of the five types of method we have outlined each has evolved in its own way. Monothetic methods were founded and are still mainly used in ecology studies where the concept of 'indicator species' (i.e. that some species of plant are much more important in determining vegetation types), which has been in existence for many years, lead to this type of approach. The majority of the work on overlapping groups has been carried out in information retrieval where the information required by two people will often overlap. Hierarchic methods are used by biologists, botanists, etc., where division into hierarchies of genus, species and sub-species is a requirement. Relocation methods are a more recent introduction and cannot be identified with a particular science (except perhaps mathematics). Sciences

such as pattern recognition and sociology have lead to special types of method which are used at present exclusively in those fields.

In the following section we shall examine examples of five types of method, concentrating on those methods which have been widely used, or have historical significance. New and promising methods are also included. Terminology has been introduced and explained progressively and some original examples are included for illustration.

C.2 EXPLANATION AND DISCUSSION OF METHODS

EARLY METHODS

1. FACTOR ANALYSIS
2. B-COEFFICIENT SEARCH
3. TRYON'S METHOD
4. MATRIX SHADING
5. RAMIFYING LINKAGE
6. CATTELL'S METHOD

1(a) HIERARCHICAL

7. NEAREST NEIGHBOUR
8. FURTHEST NEIGHBOUR
9. WEIGHTED AVERAGE
10. GROUP AVERAGE
11. CENTROID
12. MEDIAN
13. FLEXIBLE

13A. EXTENSION OF FLEXIBLE

14. WARD'S METHOD
15. SINGLE LINK ON K-LINK CRITERIA
16. NUCLEUS METHOD
17. DISSIMILARITY ANALYSIS
18. PROFILE CLUSTERING

OTHER HIERARCHICAL METHODS

1(b) ITERATIVE RELOCATION

19. BEALE'S METHOD
20. GROUP AVERAGE RELOCATION
21. NEIGHBOURHOOD METHOD

OTHER ITERATIVE RELOCATION METHODS

1(c) MISCELLANEOUS

22. MODE ANALYSIS
23. CONDENSATION MODEL

OTHER MISCELLANEOUS METHODS

2. OVERLAPPING METHODS

3. MONOTHETIC METHODS

EARLY METHODS OF CLUSTER ANALYSIS:

1. Factor Analysis

Cluster analysis began as a result of the upsurge in the use of factor analysis in the 1930's. Many authors who experimented with the then new technique of factor analysis were using it as a method for clustering variables into 'like' sets, by dividing the variables into groups according to their factor loadings and not using the loadings simply as a method of determining underlying dimensions.

The method has however many difficulties; unless oblique axes are used, the method relies on there existing groups of orthogonal factors; there is the problem of where to make the cut on each of the new factors; the fact that different rotations of the data will give different results; and that there is no measure of the 'goodness of clustering'.

Another difficulty is the possibility of having high positive and negative loadings on the same factor - the method normally proceeds simply to ignore the high negative loadings, (although this could be improved by taking absolute values, or if more appropriate by considering all high negative values as a separate cluster). A further problem is the fact that the number of factors which are obtainable from a data set of n observations and m variables depends on the number of variables - the number of factors cannot exceed m-1.

The method may often produce overlapping clusters, as an object may have high loadings on more than one factor; this may or may not be a disadvantage. The method is described in Sokal and Sneath (1963), and is apparently, despite its problems, still in use today. Recent users include Harman (1966) and Ford (1970).

Other workers have used variations of this approach. Fleming (1935) divides the space into eight regions formed by the first three factor axes, and hence obtains 'clusters'. Hopkins (1967) has used a similar method of dividing the first four principal axes into half at the average point, and reallocates points near the origin to their nearest neighbour. Noy-Meir (1973) successively divides on the principal components beginning with the largest. Aaker (1971) suggests visual clustering from a low-dimensioned factor representation - however this may well distort true interpoint distances.

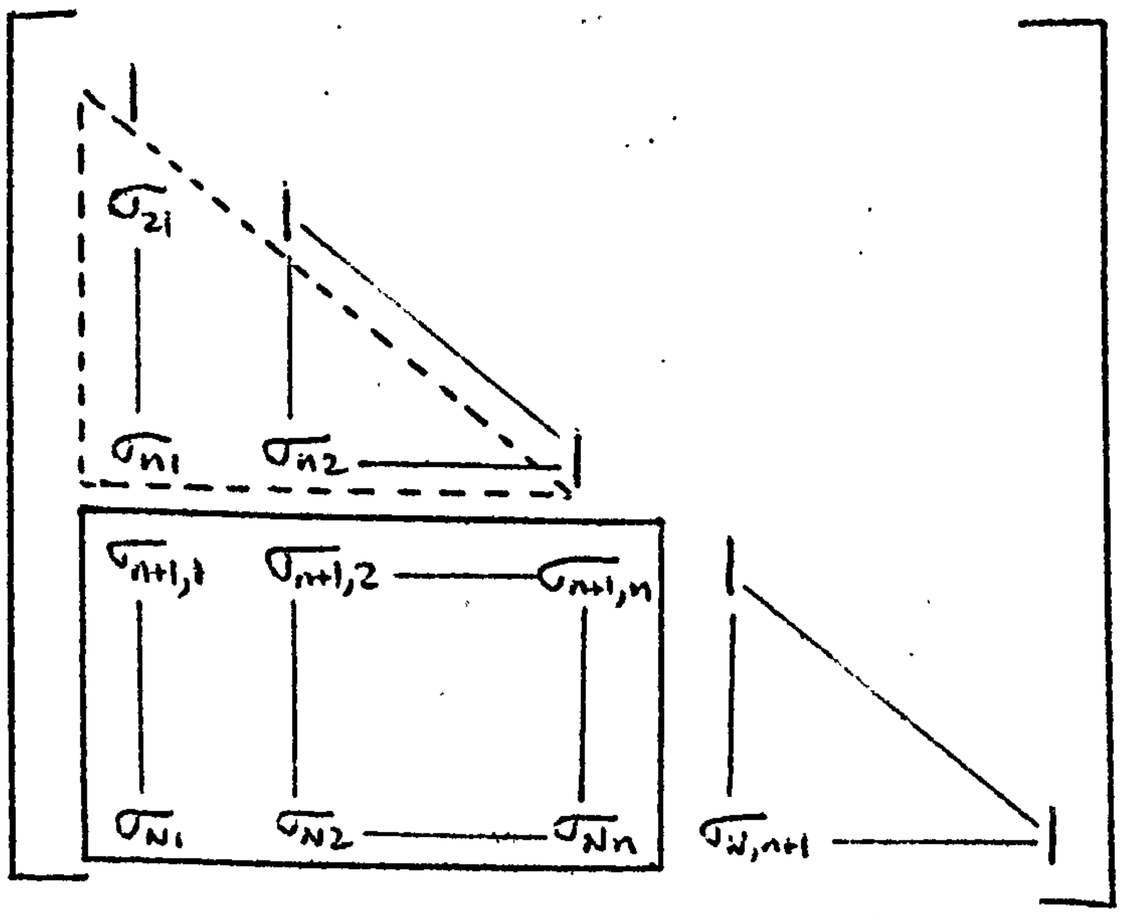
2. <u>B-Coefficient Search</u>

This method used first by Holtzinger and Harman (1941) (although Kahl and Davis 1955 attribute it to Tryon 1939) is a heuristic search method for finding clusters, which can be performed fairly simply by manual methods. It centres around a 'goodness of cluster' coefficient B. This coefficient which we shall call the B-coefficient is defined as the ratio between the average within-group correlation and the average correlation of the group members with all points outside the group.

$$\text{Thus} \quad B = \frac{\frac{1}{n(n-1)} \sum_{i,j \in G} \sigma_{ij}}{\frac{1}{n(N-n)} \sum_{i \in G, j \notin G} \sigma_{ij}} = \frac{N-n}{n-1} \frac{\sum_{i,j \in G} \sigma_{ij}}{\sum_{i \in G, j \notin G} \sigma_{ij}}$$

where n is the number of objects in the cluster G and N is the population size. Note that the coefficient is not defined for groups of 1.

We can represent the ratio on the correlation matrix by considering the matrix elements to be rearranged so that the n group members occur first:



The B-coefficient is then the ratio of the average element in the dotted triangle to the average number in the boxed-in rectangle.

If the B-coefficient had a high value then this would indicate that there was a higher cohesion between members in the group than with those outside the group under consideration. If the B-coefficient was unity then this would indicate that the group were not particularly cohesive or non-cohesive. If one could calculate the B-coefficient

for every possible cluster then one could simply select those with the highest coefficients. However this is computationally infeasible even on a computer for more than 10-12 objects, and thus a search method is used.

The method is described in Harman (1966) and in Clements (1954). Clusters are built up until the B-coefficient drops below the arbitrary level of 1.3. From the correlation matrix the pair with the highest correlation are selected and their B-coefficient calculated (this will normally give a coefficient well in excess of 1.3). Then the object most highly correlated with the previous pair is considered for a member of the group, and the B-coefficient of the 3 objects is calculated, and if it is larger than 1.3 the object joins the group. This process is continued until an object is considered which causes the coefficient to fall below 1.3, in which case this object is rejected from the group and of the remaining non-clustered objects the pair with the highest correlation are selected and the process continued as below. The procedure is continued until all points have been placed in groups, or have been found not to fit in any group.

One of the difficulties of the method is that the procedure is not well defined, Harman excludes members from groups if a drop in the B-coefficient "seems to be too great", and uses pre-judgment of the grouping expected - "Test 11 is retained, although it causes a drop of 47 points in B, because it is of the same general nature as Tests 10 and 12". Clements has used pre-judgment as shown later.

Another major difficulty is in the B-coefficient itself - the difficulty occurs when negative correlations exist, Clements has stated that "some individuals have expressed doubt whether the B-coefficient method can handle situations involving negative coefficients of correlation. Such doubt is groundless", and employs the absolute magnitude of the coefficient in cases where it becomes negative. Consider the following correlation matrix:

	X	Y	Z
X	1	-0.5	0.1
Y	-0.5	1	0.3
Z	0.1	0.3	1

We obtain -

$$B(X,Y) = \frac{-0.5}{0.2} = -2.5$$

$$B(Y,Z) = \frac{0.3}{-0.3} = -1.0$$

$$B(X,Z) = \frac{0.1}{-0.1} = -1.0$$

Thus X and Y give the highest absolute value in B and thus would be selected as a group, despite the negative correlation between them.

The difficulty occurs when either the average correlation within the group or the average between group correlation is negative.

This may be rectified by the addition of 1 to each element of the correlation matrix, and division of each element by 2, to convert all the matrix elements to be in the range 0-1,

i.e. we substitute $\sigma^1 = \frac{\sigma + 1}{2}$

Even with this improvement the method is not too accurate, with an arbitrary cut-off point at B = 1.3, and a not too well defined procedure for cluster searching, and the method, although still apparently in use today, is not recommended.

Fortier and Solomon (1966) have used random sampling of groups, and calculating the B-coefficients of these groups to try and find an approximate resolution of objects into clusters. However they state that their results were disappointing and that quicker more direct methods gave better solutions. Their conclusion was that the distribution of the B-coefficient is very skewed with the best B values in the tail.

3. Tryon's Method

This method was first published in 1939 in a monograph by Tryon. Unfortunately this book, probably the first on cluster analysis, is now almost impossible to obtain due to the limited number originally printed, however according to Bailey (personal communication) all the material of the original book is covered by Tryon and Bailey (1966) and Tryon's method is also discussed by Cattell (1944).

The method is a second-order process, based on the belief that similar objects will have similar similarities with other objects. The method begins by selecting a first element as a nucleus for a cluster, this is chosen as the element i with the highest variance of the squared correlations in each column of the similarity matrix. The

second element is chosen according to the size of an 'index of proportionality' - P_{ij}^2 given by:

$$P_{ij}^2 = \frac{\sum_k r_{ik} r_{jk}}{\sum_k r^2_{ik} \sum_k r_{jk}^2}$$

The element with the largest P_{ij}^2 joins the first element if $P_{ij}^2 > .81$. Other elements j are selected to join the first cluster on conditions -

1. The mean within group $P_{kj}^2 > .81$
2. No within group $P_{kj}^2 < .40$

When no elements are able to join this cluster a new cluster is begun.

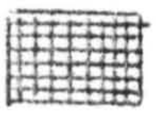
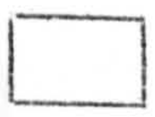
The method is similar to obtaining the second-order correlation matrix from the columns (or rows) of the original matrix and performing a simple clustering of the objects using this matrix.

The method is somewhat archaic and relies on 2 arbitrary levels .81 and .40 (which may be changed by the user), but the method, revitalized by Tryon and Bailey's recent book, and built into a computer package BC-TRY, is still used today. Articles using the method have appeared largely in marketing, for example Myers and Nicosia (1968).

4. Matrix Shading

This simple method of clustering can be traced as far back as Burt's famous book 'The Factors of the Mind' which appeared in 1940. The method is based entirely on the

similarity matrix between objects, the range of the values of which is dissected into about 4 or 5 bands so that the elements are divided into that number of roughly equal groups. Thus for example the elements of a correlation matrix may be divided into four groups such as 1.0 to 0.6, 0.6 to 0.4, 0.4 to 0.2, 0.2 to -1.0. Each element is then replaced in the matrix by a shaded square according to the group it belongs to, the shading being of a certain type for each group, and the intensity of the shading increasing with the similarity. Thus we may replace the elements of a correlation matrix as follows:

Correlation	1.0	to	0.6	 filled
	0.6	to	0.4	 squared
	0.4	to	0.2	 lined
	0.2	to	-1.0	 blank

The order of the objects in the matrix is then rearranged by inspection so as to try and place the darkest shaded squares as near as possible to the diagonal of the matrix - this brings similar objects near to one another in the order they appear in the matrix. The objects in the matrix are rearranged until there appear to be no more rearrangements which will improve the position. The improvements involve a certain 'trade-off' between multiple objective - it may be better to move a very high similarity away from the diagonal, so that two or three moderately high similarities will be nearer the diagonal. Once the matrix has been finally rearranged, it is inspected - if marked clusters are present,

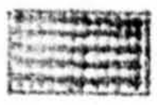
they will show as squares of dark shading on the diagonal of the matrix, surrounded by lighter shading.

Consider the hypothetical similarity matrix:

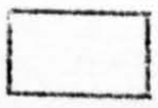
	1	2	3	4	5	6	7	8
1	100	10	64	-12	81	04	09	19
2	10	100	-24	54	25	14	46	95
3	64	-13	100	-12	74	01	08	15
4	-12	54	-12	100	35	16	26	25
5	81	25	74	34	100	-03	00	29
6	04	14	01	16	-03	100	41	16
7	09	46	08	26	00	41	100	-01
8	19	95	15	25	29	16	-01	100

We divide the similarities into 4 groups:

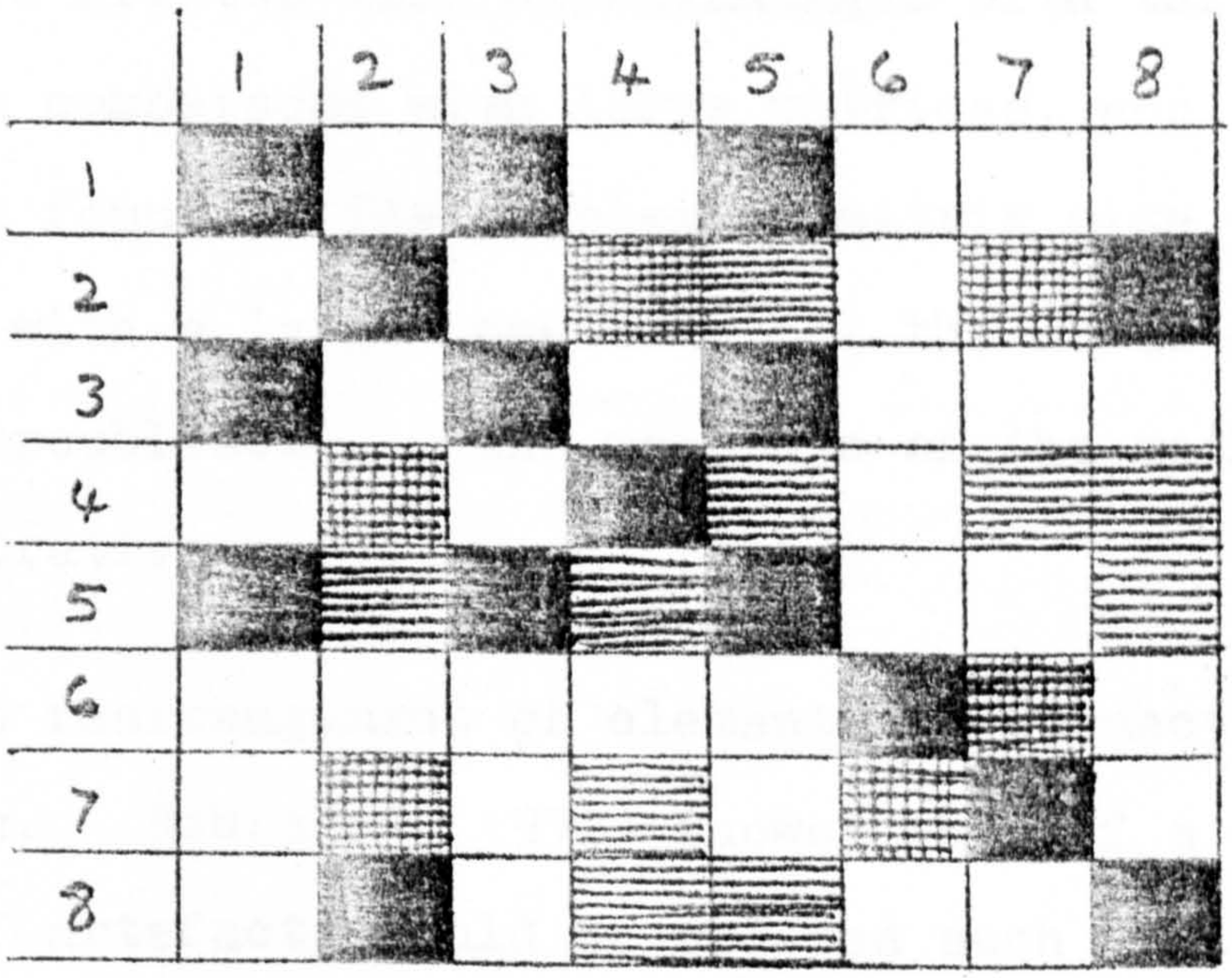
100 to 60 

60 to 40 

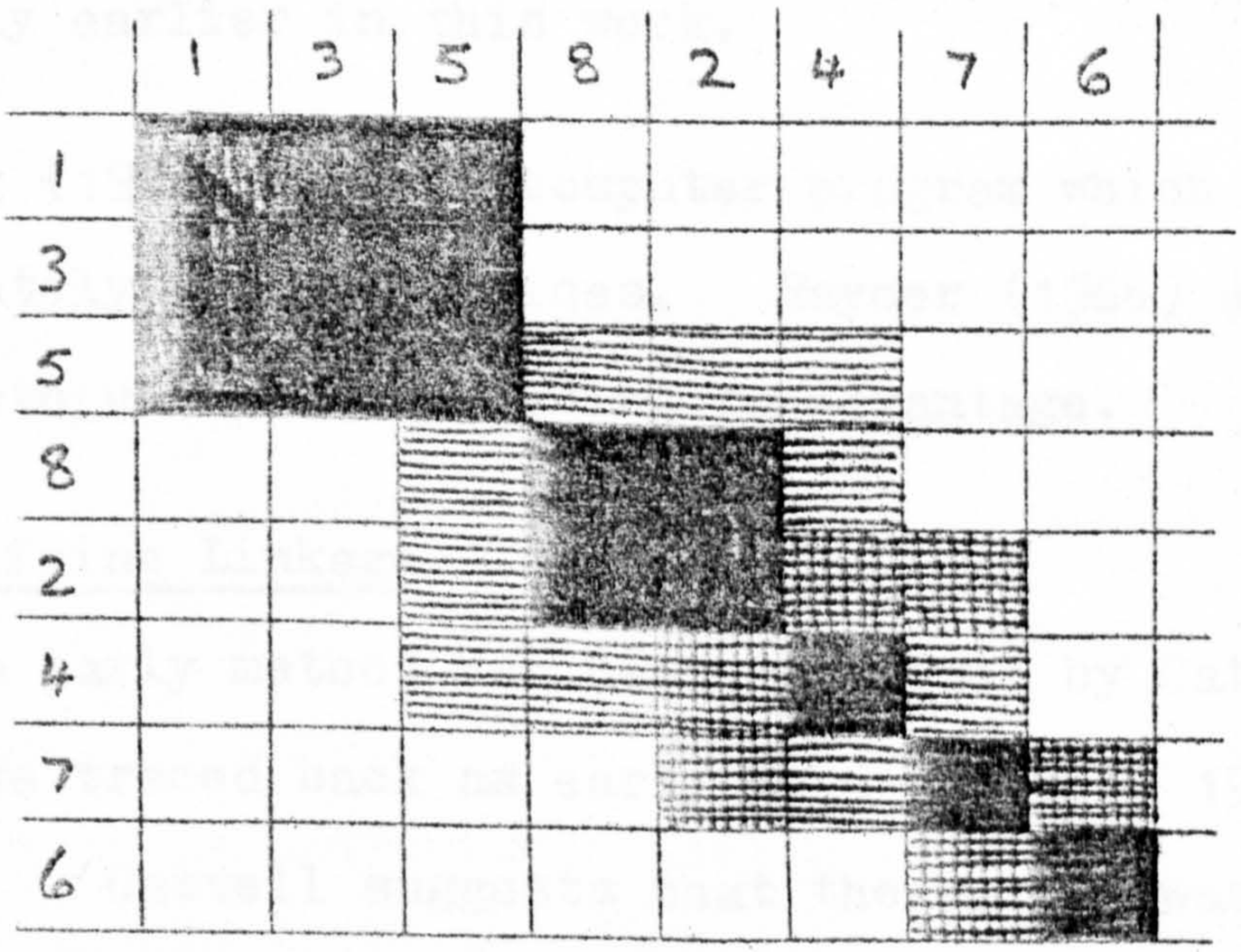
40 to 20 

20 to -25 

By replacing the values in the matrix by their associate shaded square, we obtain:



Several improvements can be seen by visual inspection, simply by transposing adjacent elements. For example, exchanging 6 and 7 brings two shaded squares nearer the diagonal and moves only blank squares further from the diagonal. The matrix becomes, after several exchanges, the following:



From this rearrangement a definite group of elements 1, 3 and 5 can be seen, a pair group of 8 and 2 with an associate element 4 and another weaker pair group 6 and 7.

There are two main disadvantages with this method; the method is cumbersome with large matrices, and local optima are often found. The problem of matrix size can be partly overcome with a large computer, but the local optima problem is more troublesome. The main use of the method today is in an illustrative role.

This rearrangement of elements is connected to seriation. Robinson (1952) showed that if a similarities matrix of artefacts could be ordered such that all similarities increase as one approaches the diagonal of the matrix i.e. such that for all i,

$$S(i, j) \not> s(i, j-1) \qquad j \leq i$$

$$S(i, j) \not> S(i, j+1) \qquad j \leq i$$

then the order of the objects in the matrix should correspond to their order in time. Seriation methods are considered more fully earlier in this work.

Ling (1971) gives a computer program which prints appropriately shaded matrices. Rayner (1966) and Clarke (1962) both use matrix shading to advantage.

5. Ramifying Linkage

This early method has been discussed by Cattell (1944) and can be traced back as early as a paper in 1942 by Sandford. Cattell suggests that the method was at that time used in most cluster studies. The method begins with a threshold level in the similarity matrix, above which pairs of elements are said to be linked. By a systematic method,

the groups of objects which are all linked to the other members of the group, are found. These form the resultant clusters.

The method is a forerunner of the more modern complete-link method. The speed of the method depends on the number of similarities above the threshold, but Cattell estimates that with 200 variables and a threshold level such that a tenth of the similarities form links, as many as 60,000 linkages must be inspected.

The method has the problem of deciding a best threshold and thus has no advantages over the complete linkage method performed by computer.

6. Cattell's Method

This has been proposed in Cattell's 1944 paper as a faster method (and as such, more suitable for hand calculation) than the ramifying linkage method. The method begins by setting a threshold, above which similarities are considered significant. Objects which have similarities above this threshold are said to be linked. Pairs of objects which are themselves linked, and both have linkages with the same two or more other objects become members of the same group. As ramifying linkage was a simpler version of the complete link method, Cattell's method is a version of the Klink method (with K = 3) explained later, and has no advantages over this method.

1(a) HIERARCHICAL METHODS

7. Nearest Neighbour

The method of nearest neighbour clustering, sometimes called the single link method, was one of the first clustering methods to be used, because it was easily suited to hand calculation. The method found its use mainly in the field of numerical taxonomy and is normally attributed to Sneath (1957), although the technique was independently introduced by McQuitty (1957) in the same year as Elementary Linkage Analysis, and was in use even earlier (Florek et al 1951). The method is the same as the Minimum method of Johnson (1967), and the method of Jizba (1964), and very similar to one of the methods of Cattell and Coulter (1966). Examples of the use of the method are in Holloway and Jardine (1968), Lessel and Holt (1970) and Muir et al (1970).

The method is perhaps best understood by describing the way in which it is performed by an agglomerative algorithm. Initially each observation is considered to be in a cluster on its own and then the two 'nearest' observations (i.e. that pair with the highest similarity) are clustered (linked) into one group. The algorithm proceeds by successively linking the pair with the next highest similarity (if one of the pair is already a member of a group, then the other observation joins that group; if both observations are already in groups, then the groups join), until all the points are in one group. From this a dendrogram can be constructed. The method may be used with a 'cut-off point' - groups only being formed up to a certain similarity level. The method can

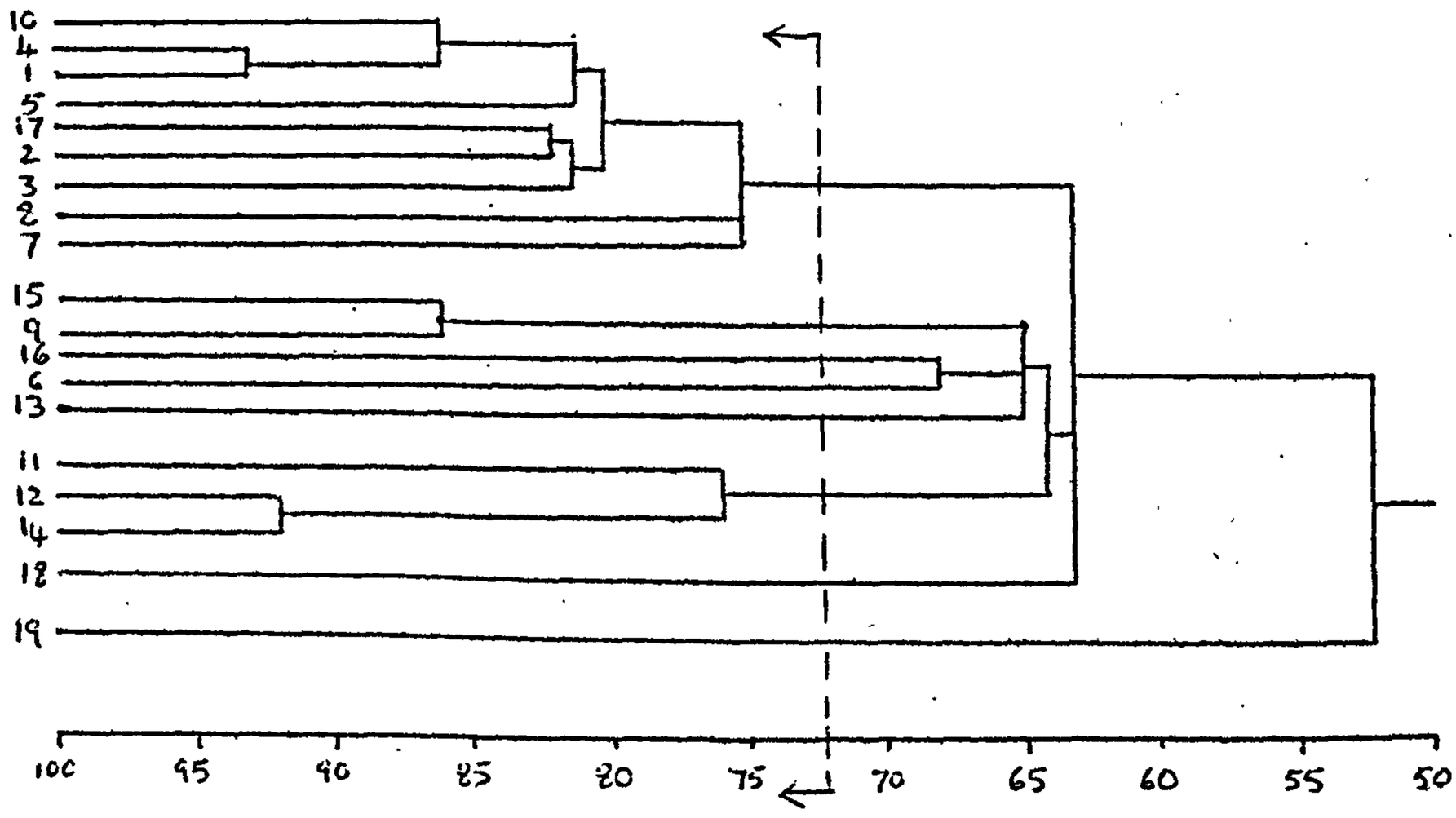
also be performed by gradually completing one group at a time - this is the <u>sequential</u> approach, as opposed to the <u>global</u> approach.

As the method is suitable for hand computation we can explain further by a simple example.

The following correlation matrix (decimal points omitted) is taken from Kahl and Davis (1955) who were attempting to investigate the relationship between various measures of socio-economic status.

1																	
2	80																
3	77	81															
4	93	70	70														
5	81	70	65	50													
6	57	63	52	59	49												
7	75	69	53	74	63	34											
8	73	69	75	71	63	59	60										
9	48	57	60	39	55	65	41	53									
10	78	59	65	86	54	62	62	60	39								
11	53	36	41	47	46	64	50	43	38	41							
12	54	48	45	43	51	48	50	32	30	34	76						
13	54	60	53	53	48	65	44	41	57	30	33	28					
14	49	46	43	39	36	50	38	30	35	45	62	92	49				
15	37	48	54	36	46	48	47	40	86	23	45	37	61	40			
16	43	39	45	40	44	68	43	47	49	35	43	62	54	46	29		
17	49	82	59	50	43	45	40	40	39	47	39	20	25	37	29	16	
18	51	34	36	41	52	34	56	34	39	34	63	35	30	22	21	35	23
19	29	45	41	22	29	48	35	32	48	20	36	27	52	14	52	44	20	13
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

As correlation is a similarity measure, our first step is to pick out the largest element in the matrix - this is 0.93 between the indices 1 and 4. The next highest element is found - 0.92 between indices 14 and 12. At 0.86, 15 joins 9 and also 10 joins 4 (which is already clustered with 1). This procedure continues, selecting next highest elements in the matrix, until all points are joined. At some stages, elements may be selected which join observations which are already in the same group - these are ignored. In the above example the resultant dendrogram is as follows:



A vertical cut through any point on the dendrogram sections the observations into clusters. From the dendrogram one normally chooses to cut the dendrogram in an area where few or no fusions of observations take place - this concept is related to that of within-group similarity and between-group dissimilarity - thus in the above dendrogram we have chosen to cut at the 0.70 level. This gives the following clusters:

A	1	OCCUPATIONAL CATEGORY (WARNER DEFINITION)
	2	OCCUPATION OF FRIENDS
	3	SUBJECTS EDUCATION
	4	OCCUPATIONAL CATEGORY (CENSUS)
	5	OCCUPATIONAL CATEGORY (NORTH-HATT DEFINITION)
	7	INTERVIEWERS RATING OF CLASS
	8	SELF-IDENTIFICATION OF CLASS
	10	SOURCE OF INCOME
	17	SUBJECTS WIFE'S OCCUPATION
B	9	SUBJECTS MOTHER'S EDUCATION
	15	SUBJECTS FATHER'S EDUCATION
C	11	CENSUS TRACT AVERAGE RENT
	12	INTERVIEWERS AREA RATING
	14	INTERVIEWERS HOUSE RATING

and five outliers 6, 13, 16, 18, 19

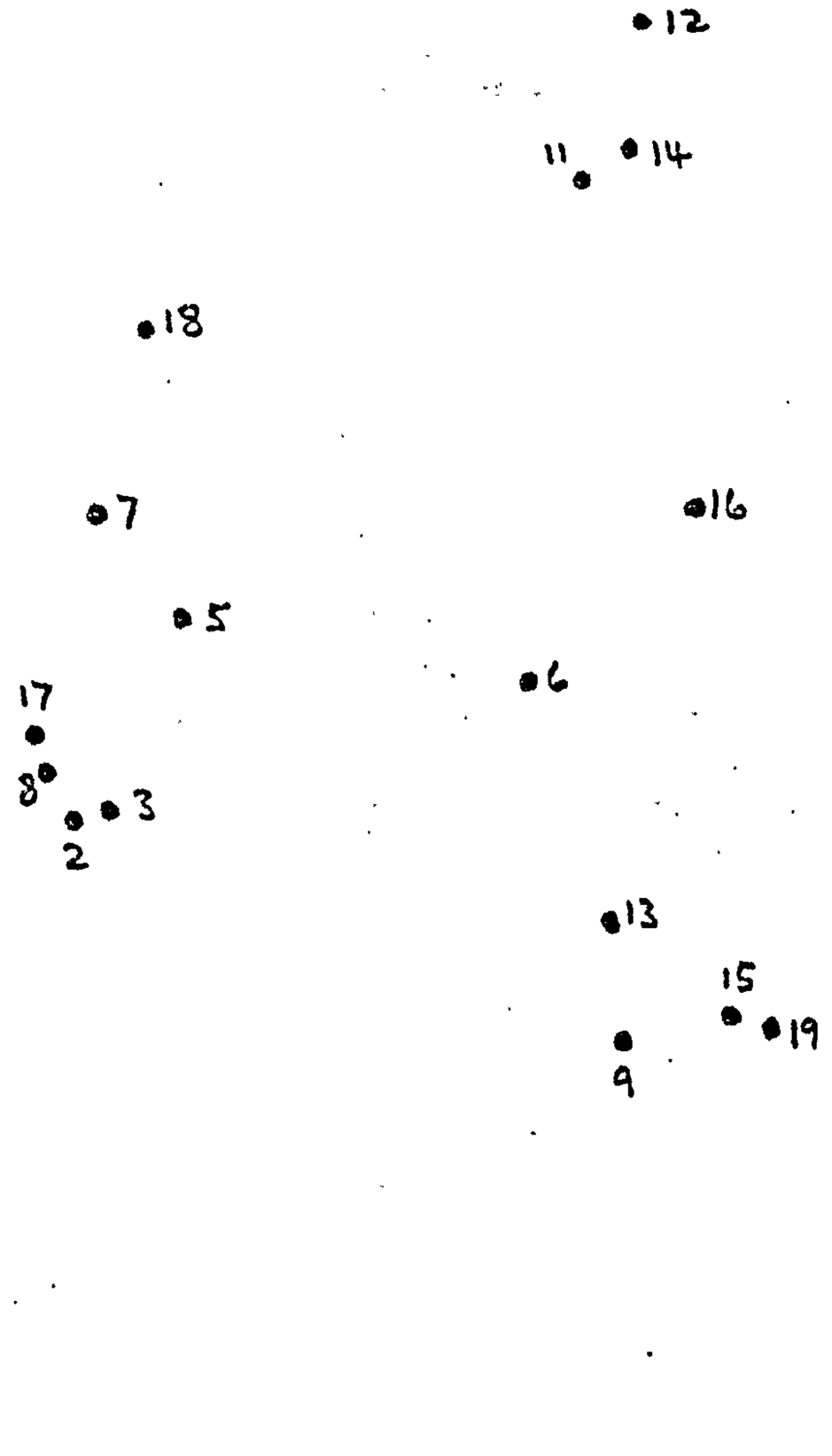
This grouping seems meaningful - group C is related to the type of houses in the district and the two items in group B seem quite well related together and not with the other measures. Group A seems to be a mixture of class and occupation measures which clearly are related.

This particular data set was used because it has been subjected to other analyses by various authors. Kahl and Davis used the method of search using the B-coefficient of Holtzinger to try to find clusters. Fortier and Solomon (1966) have used their own search method on the data, and more recently Solomon (1971) has subjected the data to factor analysis. King (1967) has also used the data as an illustration of his stepwise clustering method (another name

for the weighted average method - see later). The results of the cluster analyses can be shown as follows:

Nearest Neighbour	(12	14	11)	18	(10	4	1	7	5	8	3	17	2)	(9	15)	13	19	16	6
Kahl & Davis	(12	14)	(11	18)	(10	4	1)	7	5	(8	3)	(17	2)	(9	15)	(13	19)	(16	6)
Fortier & Solomon	(12	14)	11	18	(10	4	1)	7	5	(8	3)	(17	2)	(9	15)	13	19	16	6
King	(12	14)	11	18	(10	4	1	7	5	8	3)	(17	2)	(9	15)	13	19	16	6

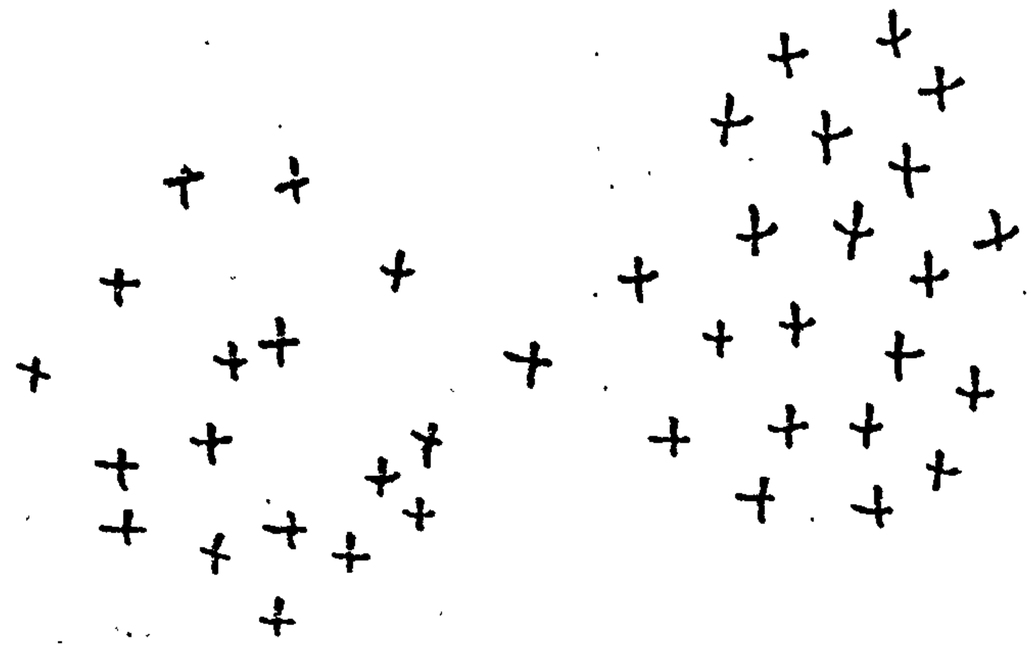
King's method is similar to nearest neighbour in that a subjective decision must be made to decide at what level to 'cut' the dendrogram into clusters, and in his dendrogram there is no clear indication of where to cut. At a later level, the clusters are identical with those found by nearest neighbour. The 2-factor analysis performed by Solomon of the data gives the following representation:



Solomon concluded that of the 3 methods (the above, except for nearest neighbour) King's method gave the best solution. By inspection of the 2-factor solution it can be seen that nearest neighbour gives an even better solution by joining 2 and 17 to the main cluster. Thus single link has produced a solution as good, if not better, than the other more complex and more time-consuming methods.

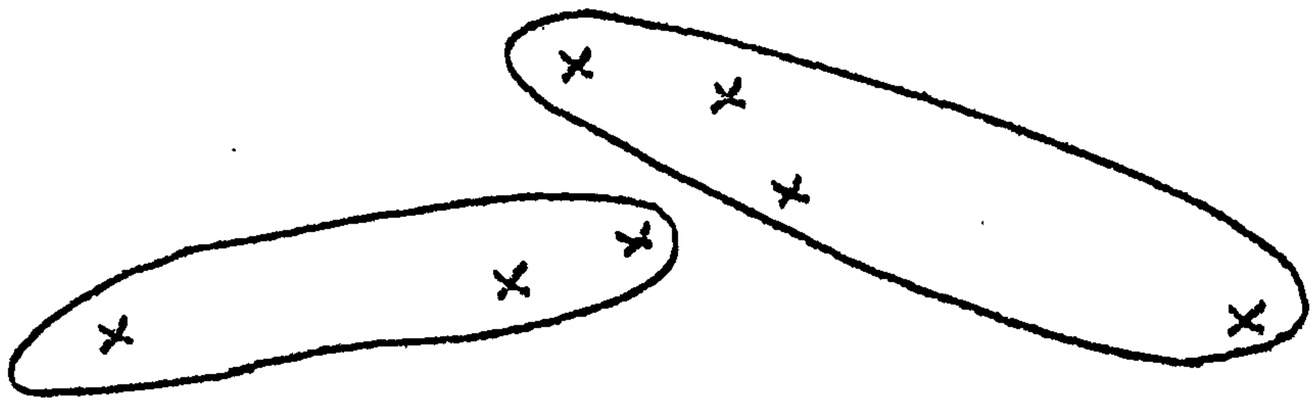
The nearest neighbour method is connected to the minimal spanning tree in graph theory. The minimal spanning tree problem can be stated as to find the branches in the network that have the shortest total length whilst a route exists between each pair of nodes. This is in effect the nearest neighbour sorting method - the nodes are the observations and the length of the branches are the dissimilarities between pairs of observations. Thus the single link method may be executed by a minimal spanning tree algorithm. See Bettman (1971), Wirth et al (1966), Gower and Ross (1969), Prim (1957), Roger (1971), Rohlf (1973).

The disadvantage of the nearest neighbour method is its 'chaining' effect as explained by several authors, Wishart (1969), Jardine and Sibson (1968). The chaining effect is that if data is subjected to error (or noise) an abberant point which exists between clusters will cause the clusters to fuse together too soon as the agglomerative algorithm proceeds, e.g.



Jardine and Sibson have defended the method by suggesting that the defects of nearest neighbour are the defects of hierarchic classification itself and that the solution to the problem is not to use hierarchic methods to classify.

Useful algorithms are given in Van Rijsbergen (1970), Jardine (1970), and Gower and Ross (1969). The method of Wirth et al (1966) (see also Estabrook 1966), is identical to nearest neighbour, but uses the number of 'links' in each cluster as a measure of connectedness. Gyllenberg (1963) and Carmichael et al (1968) also produce methods which are essentially nearest neighbour with a cut-off value. The method of Kamen (1970) joins each point to its nearest neighbour - this is not the same as the nearest neighbour method, and often produces very poor groupings. For example the solution to the clustering of these seven points is as follows:



8. Furthest Neighbour

The furthest neighbour method (sometimes called the complete link method) is similar to that of nearest neighbour, and is also one of the early methods. The method can be traced back as far as 1948 in a paper by Sorenson, and has been formally proposed by McQuitty (1960) as Hierarchical Syndrome Analysis, Johnson (1967) discusses it as the Maximum method, it is also identical to Constantinescu's (1965, 1967) method, and Bonners' (1964) program II, and is related to Cattell's (1944) Ramifying Linkage method. Gengerelli (1963) has a similar method which employs a cut-off value.

In order to explain the method and to show its relationship to the nearest neighbour method, the agglomerative algorithm is perhaps the best for descriptive purposes. As before, the observations are first considered to be separate clusters, and the first fusing is of the pair of objects with the highest similarity. The method proceeds to join points successively, but points may only join groups if the point has a similarity of above a certain level with all the members of the group. Similarly groups may only merge if all members of one group have a similarity above the threshold level with all the members of the other group. Thus in furthest neighbour the fusions depend on all the between-cluster similarities, whilst in nearest neighbour the fusions depend on only the largest between cluster similarity.

In order to explain a way in which the method can be executed simply by computer, we introduce the following data matrix from Clements (1954) which he obtained from Kroeber (1939). The similarity measure is Yule's coefficient of association $Q = \frac{ad - bc}{ad + bc}$. The objects in the study are various American Indian tribes from the north west of California.

Tribal Codes:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		77	68	64	65	73	70	72	66	59	49	59	45	00	01
2	77		93	89	79	90	88	73	70	44	32	35	19	-31	-31
3	68	93		90	81	90	89	71	62	72	26	32	09	-39	-30
4	64	89	90		92	89	89	69	58	57	54	61	14	-29	-24
5	65	79	81	92		74	84	77	64	56	61	45	24	-29	-18
6	73	90	90	89	74		99	86	58	61	48	46	13	-43	-24
7	70	88	89	89	84	99		92	66	66	47	49	22	-29	-18
8	72	73	71	69	77	86	92		59	76	54	55	30	-09	23
9	66	70	62	58	64	58	66	59		38	55	53	59	-07	01
10	59	44	72	57	56	61	66	76	38		67	79	61	21	41
11	49	32	26	54	61	48	47	54	55	67		74	60	38	37
12	59	35	32	61	45	46	49	55	53	79	74		64	32	45
13	45	19	09	14	24	13	22	30	59	61	60	64		42	43
14	00	-31	-39	-29	-29	-43	-29	-09	-07	21	38	32	42		86
15	01	-31	-30	-24	-18	-24	-18	23	01	41	37	45	43	86	

The furthest neighbour method proceeds as follows: first we select the highest element in the matrix, which is .99 between tribes 6 and 7. From these a new single variable 16 is constructed with the values -

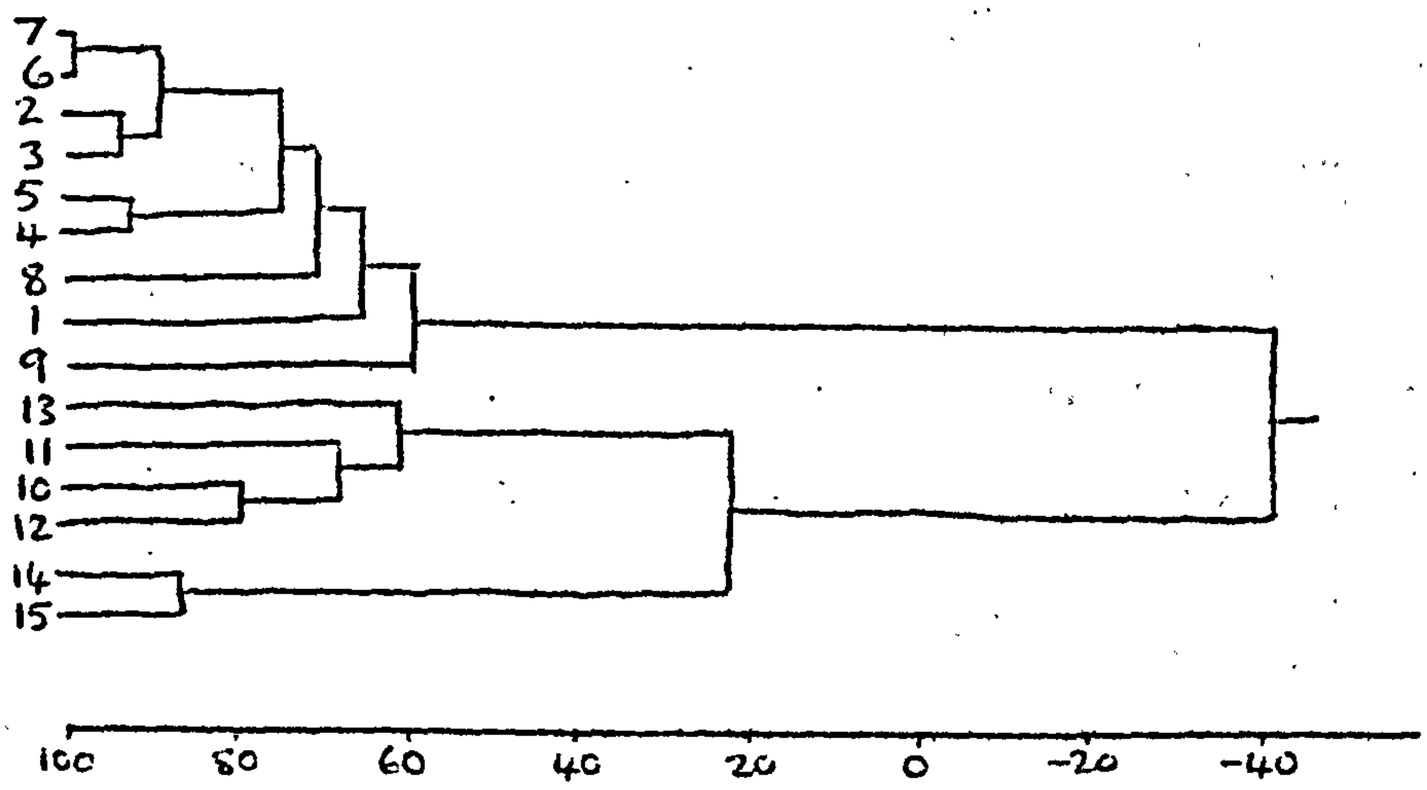
similarity (16,I) = Minimum (similarity (6,I), similarity (7,I))

(Note that if the maximum function is used instead of the minimum function then this would produce the single link method.)

Then the new matrix is scanned for the highest element which is .93 between tribes 2 and 3 and a new variable is constructed as before and the procedure continues, 5 and 4 merging next at level .92. When the new variable replacing 5 and 4 is formed the matrix is as follows:

Tribal Codes:	1	(2+3) 17	(4+5) 18	(6+7) 16	8	9	10	11	12	13	14	15
1	(77)	68	64	70	72	66	59	49	59	45	00	01
17 (2+3)	68	(88)	79	88	71	62	44	26	32	09	-39	-31
18 (4+5)	64	79	(79)	74	69	58	56	54	39	14	-29	-18
16 (6+7)	70	88	74	(88)	86	58	61	47	46	13	-43	-24
8	72	71	69	86	(92)	59	76	54	55	30	-09	23
9	66	62	58	58	59	(70)	38	55	53	59	-07	01
10	59	44	56	61	76	38	(79)	67	79	61	21	41
11	49	26	54	47	54	55	67	(74)	74	60	38	37
12	59	32	39	46	55	53	79	74	(79)	64	32	45
13	45	09	14	13	30	59	61	60	64	(64)	42	43
14	00	-39	-29	-43	-09	-07	21	38	32	42	(86)	86
15	01	-31	-18	-24	23	01	41	37	45	43	86	(86)

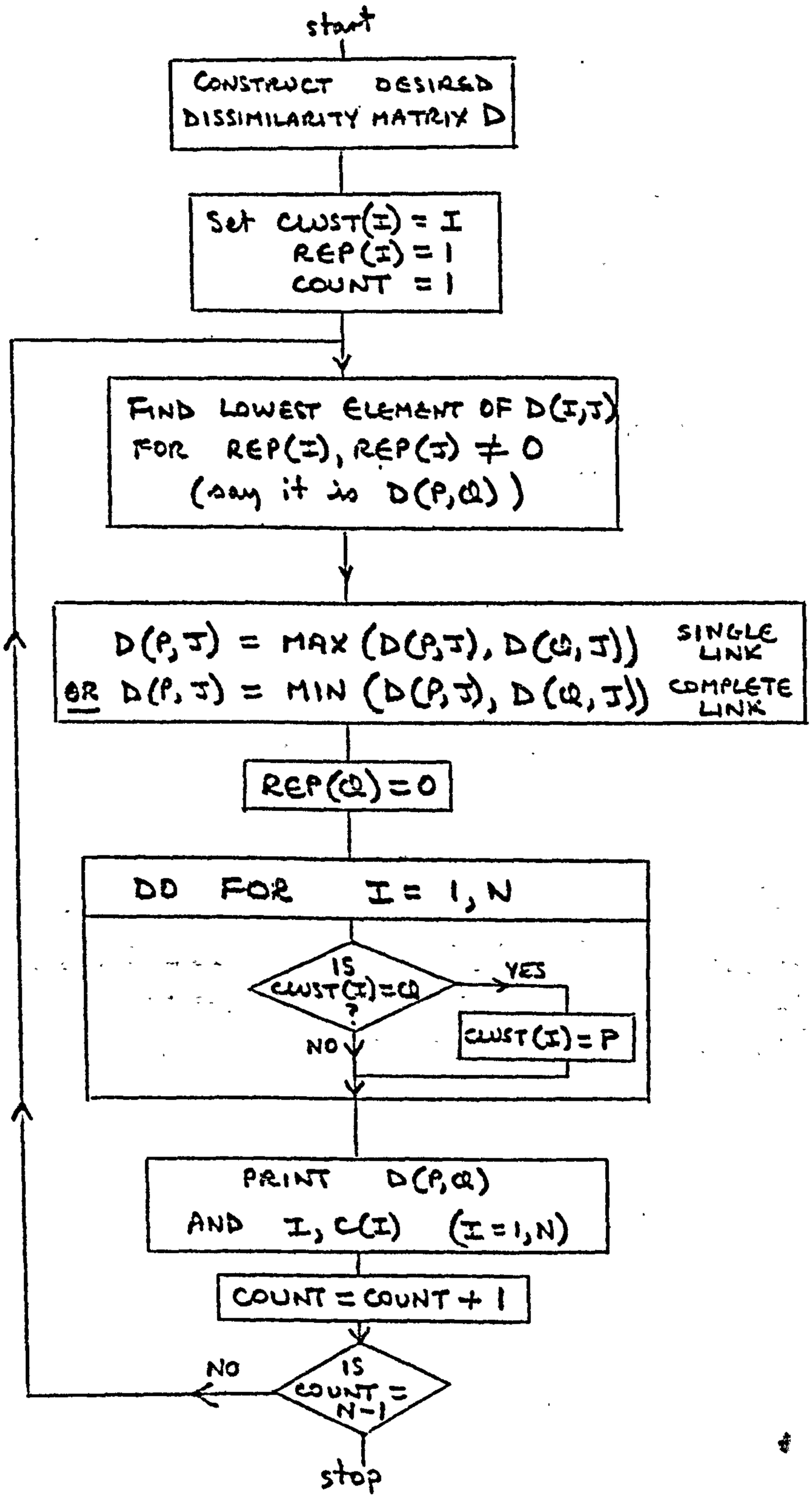
Again the largest element is selected, it is .88 between variables 16 and 17. The procedure continues until all the objects are in one cluster. In our current example the resultant dendrogram is:



As with the other hierarchical methods the dendrogram must be 'cut' to produce clusters, and in the single link example we have mentioned that this cut should be taken in a region where no fusions take place to produce within cluster similarity and between cluster dissimilarity, and here we may explain another criterion. With a similarity function such as correlation a cut taken at a negative value may allow sub-clusters to merge into a single cluster which have a very low similarity, thus the cut-off point must always be above zero and preferably much above this level, indeed some authors state a fixed level always to be used. Thus in the above dendrogram we choose not to cut the tribes into 2 clusters at about the .1 level, but into 3 clusters at about the .5 level. This gives the groups 1-9, 10-13, 14 & 15. In Clements' paper he states that these were the three groups found by Kroeber (1939) in his appendix to

Driver's paper by the method of matrix shading. Clements then proceeds to show the -coefficient method arriving at the same solution. However at the point before the last tribe was included (tribe 10), the groups were 1-9, 11-13, 14 & 15, and Clements places tribe 10 with the 11-13 group giving "so slight a drop that it may be ignored" in the -coefficient, without trying it with the two other groups. If he were to place tribe 10 with the 1-9 group the -coefficient rise from 3.287 to 3.647. Also it can be shown that the best solution using this method is the two group solution 1-13, 14 & 15 which gives the -coefficients -43.1 and -60.4, compared to 3.2, 1.5 and -60 for the three group solution.

The algorithm used in this example can be easily implemented on a computer and is suitable for both single and complete link methods. The flowchart is as shown in Figure 23.



FLOWCHART FOR SINGLE AND COMPLETE LINK

FIGURE 23

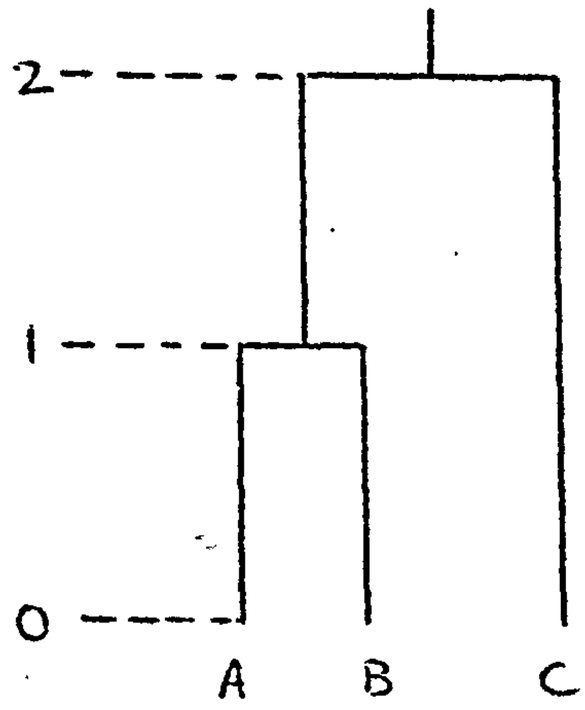
The advantage of the complete link method is that it successfully overcomes the undesirable chaining effects of single link. However the method has been criticized by Jardine and Sibson (1968) on two counts, that it is ill-defined and not continuous. They explain the term ill-defined as meaning that a unique result should be obtained from given data and that this is not always true with furthest neighbour. The problem is caused when there are ties in the similarity matrix.

E.g. suppose we have the dissimilarity matrix:

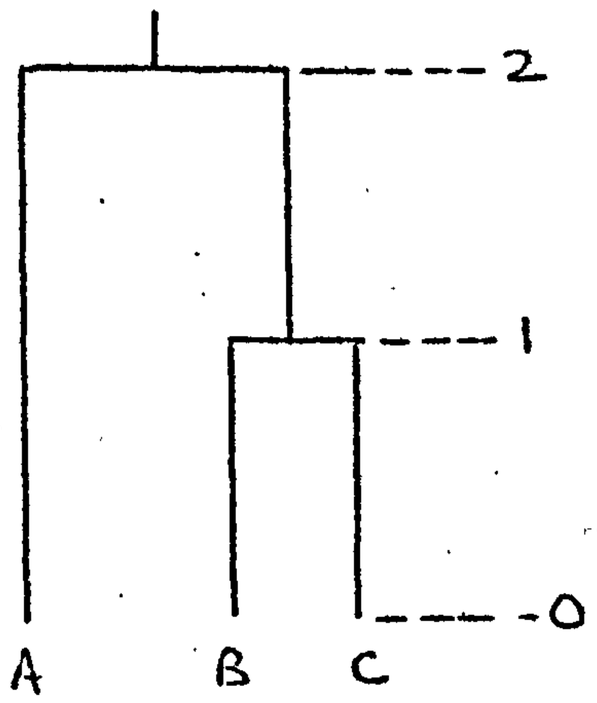
A	0		
B	1	0	
C	2	1	0

A B C

Then if A and B are fused first we obtain the dendrogram:



whilst if B and C are fused first we obtain:



These are clearly very different dendrograms, but it is expected in practical situations that ties will rarely occur, (see Williams et al 1971), and even so the computer program

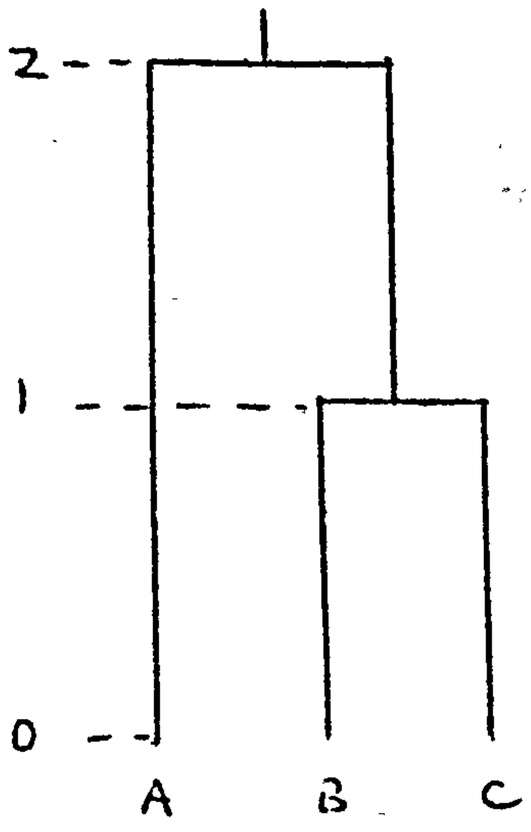
can be modified so that when ties occur all the mergers concerned take place at the same time.

The second of the criticisms of Jardine and Sibson is that of non-continuity. They define continuity as that small changes in the data create small changes in the dendrogram.

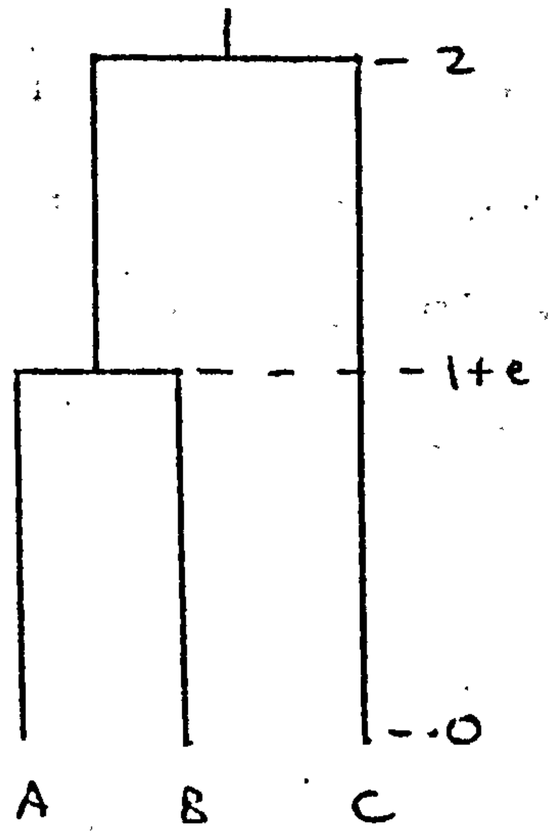
E.g. consider the dissimilarity matrix

A	0		
B	1 + e	0	
C	2	1	0
	A	B	C

If e 0 we obtain:



whilst if e 0



Again this seems to be a major change, but consider when this effect is likely to occur, and when the distortion is likely to be greatest. This is when -

$$D(A,B) >> D(A,C)$$
$$D(A,B) >> D(C,B)$$
and $$D(A,C) \simeq D(C,B)$$

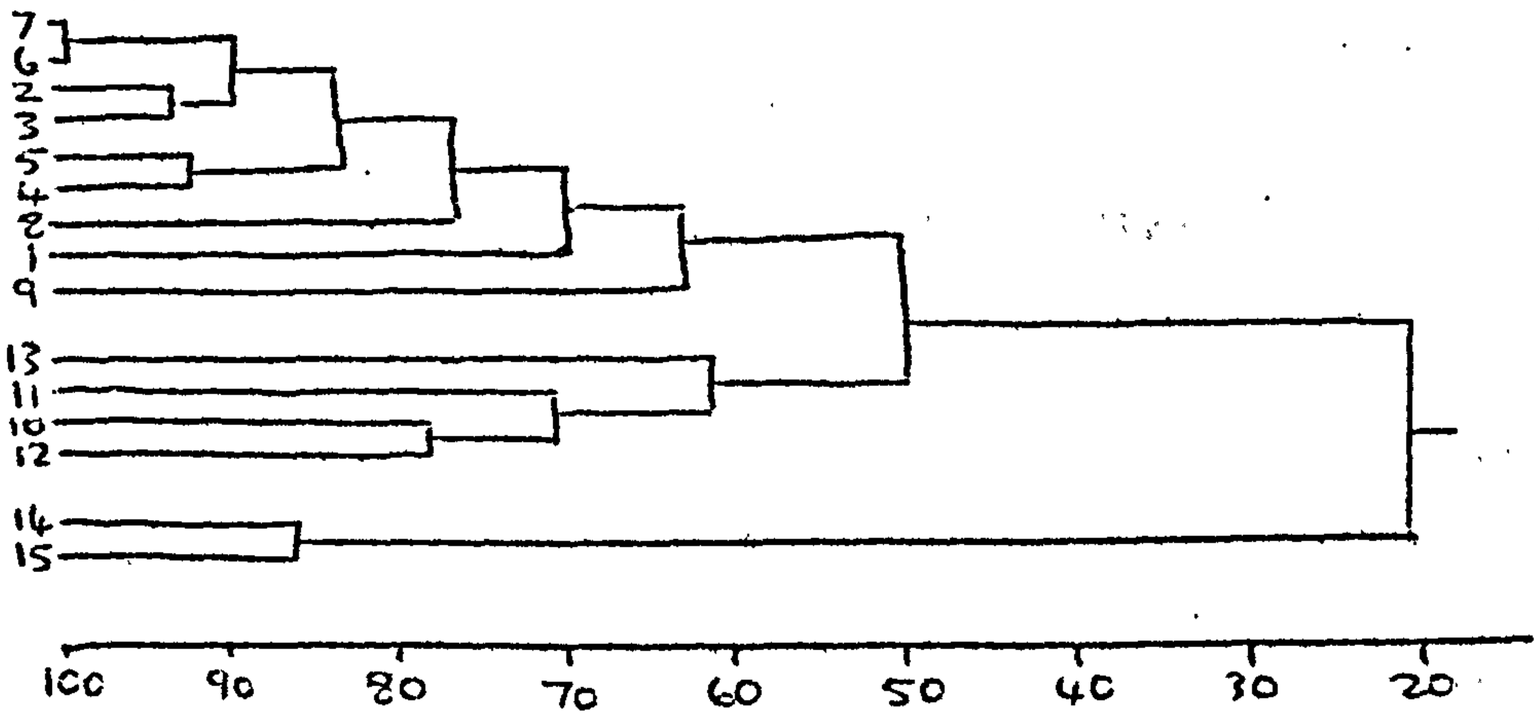
i.e. when points (or groups) are fairly well strung out and nearly equally spaced. Thus the prime effect of the lack of continuity is simply to reduce any chaining effects that may occur.

9. Weighted Average

The weighted average method is a natural extension of the single and complete link methods. As with other simple cluster methods, the fields of use are so numerous and widespread that the method has been independently introduced by several authors including Sokal and Michener (1958), workers in numerical taxonomy who called the method weighted average link, McQuitty (1966) in psychology who called the method Similarity Analysis by Reciprocal Pairs and King (1966, 1967) in investment analysis who gave the name Stepwise Clustering to the method, it is also similar to Fisher's (1969) method. The method has proved to be quite popular and has been used in Fry (1964), Grigal and Arneman (1969), Mello and Buzas (1968), Boyce (1964) and Valentine and Peddicord (1967).

The method can be executed in a very similar manner to that described for complete linkage except that when new variables are formed from the fusion of old variables, the new variable is created by calculating the arithmetic average of the previous two. Thus clusters fuse at a lower level of similarity than in single link, but at a higher level than that of complete link.

Using the Californian tribes data, the weighted average method produces the following dendrogram:



Comparing this with the furthest neighbour result it can be seen that with one or two exceptions the tribes conglomerate in the same order but at slightly higher similarity levels.

The main difference between the dendrograms is that in the complete link case the cluster of tribes 10-13 join tribes 14 and 15 before the group 1-9, and with weighted average the larger two groups join first. Also with weighted average it is difficult to pick out the 3 group solution in preference to the 2 group solution. By consideration of the similarity matrix it can be seen how the 2 interpretations of the matrix have been produced - the very low, mainly negative correlations between tribes 1-9 and 14 and 15 show that these 2 groups are well separated, and the correlations between groups 1-9 and 10-13 range from .09 to .76 and between 10-13 and 14 and 15 range from .21 to .45.

Thus the tribes can be considered as being in 3 groups - the groups 1-9 and 14 and 15 being either side of the 10-13 group.

This illustrates the difficulty in picking out clusters from dendrograms, and the usefulness of using more than one method to analyse data.

The weighted average method is difficult to calculate quickly manually, and is well suited for the computer, where it can be executed rapidly. Mather (1969) and Bonham-Carter (1967) have published programs for the method, but the method can easily be incorporated in the given algorithm by a simple change. The instruction

$$D(P,J) = MAX(D(P,J),D(Q,J))$$

becomes

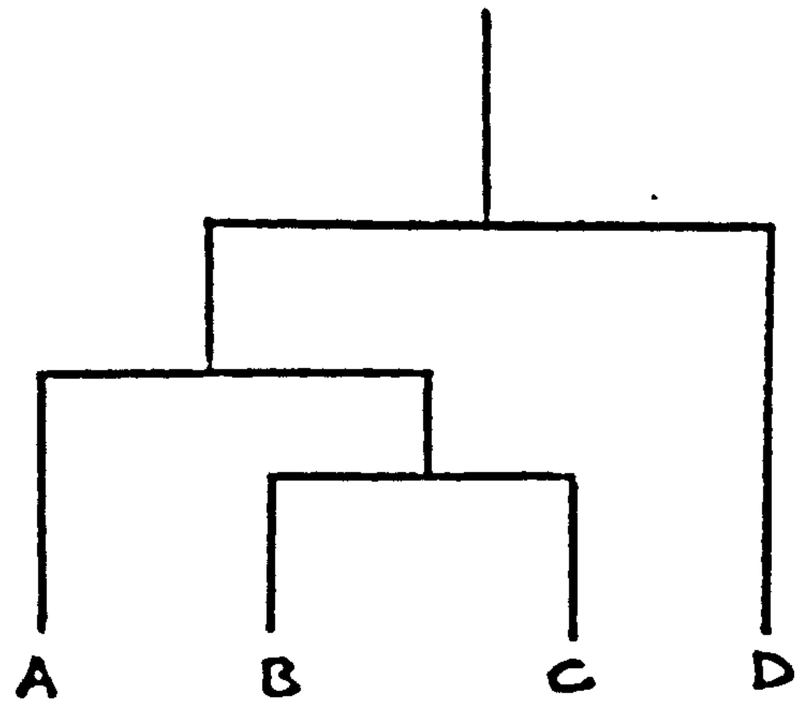
$$D(P,J) = \tfrac{1}{2}(D(P,J) + D(Q,J))$$

The criticisms of Jardine and Sibson to complete linkage also apply to this method and may be countered by the same argument as previously given. Careful use of the similarity coefficient to be used with the weighted average method is required as a meaningful average is required.

Lorr et al (1963) introduce a similar method but incorporating a cut-off point (see also Lorr and Radhakrishnan 1967, Lorr, Bishop and McNair 1965 and Stone 1960).

10. Group Average

This method was originally referred to as the unweighted average method because of its relation to the previous method. Consider the dendrogram below:



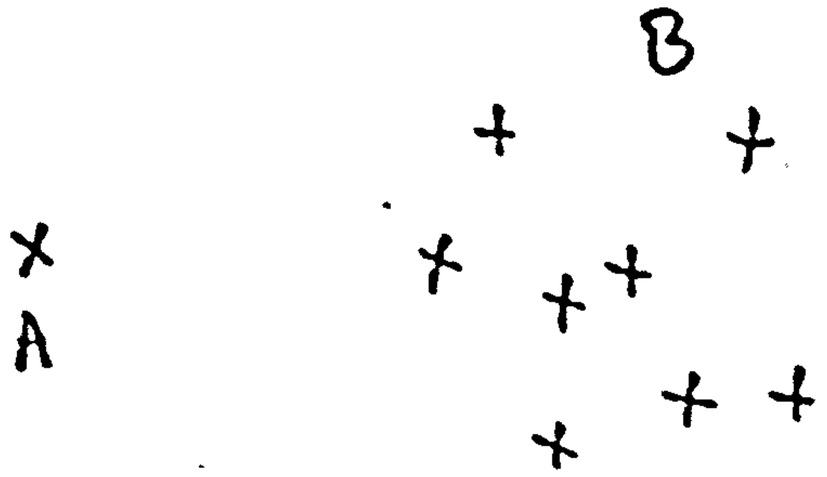
In the weighted average method, at the stage when A joins the group BC, because a simple arithmetic average is made between A and the average of B and C, B and C are effectively weighted by a factor of a half of that of A. Further, when the group ABC merges with D, then D has four times the weight of B.

With the group average method the new variable associated with the group ABCD when D joins ABC is calculated by averaging the dissimilarities of each of the original four variables, hence the term unweighted average method. This means effectively that when the group ABC is fused with D, the vector corresponding to the new object ABC is given a weight of three times that of D. This apparently paradoxical situation of using weighting with the unweighted average method has lead to the new term of group average.

The method can largely be attributed to Sokal and Michener (1958), although Lance and Williams (1967) have

found that they only used the method for element/group and not group/group fusion. The merits of using the weighted against unweighted methods are discussed in Sokal and Michener and also in Sokal and Sneath (1963). The method has also been proposed by Berry (1961), McQuitty (1966) and Rao (1952) after Tocher. Overall's method (1963) is a heuristic simulation of group average, the method of Hope (1969a, a, 1970) and that of Kendall (1966) are also very similar. Elton and Gruber (1969) give a sequential variation and Lorr's (1966) method is also sequential (as used by Bartko et al 1971 and Bass et al 1969), but leaves a large number of items unclassified.

The basic problem is that with the weighted method, too much weight may be given to the 'atypicalness' of late arrivals, whilst with the unweighted method the effect of a smaller group may be swamped when merged with a larger group. The question is partly dependent on the nature of the data and the experiment. This may be illustrated by a hypothetic example. Consider the points below as some of the objects in a cluster analysis:



At the stage when the group B on the right has been fused to one cluster, the next fusion must be with A (all the points not shown are further away from the group than a). The group B consists of various specimens of chimpanzees. Now if A is a gorilla then before fusions with other non-ape objects are to be considered then it seems that equal weight must be given to A as to the whole of B, the new centre point thus lying halfway between A and the centre of B. Thus the weighted average method should be used. However if A is simply a mutant chimp then he deserves less weighting in the analysis and the group average method should be used.

The group average method has been quite popular in taxonomy studies, and is often used in parallel with the weighted method. Examples are Hall (1965), Kaesler and McElroy (1966) and Boyce (1969). The program given by Bonham-Carter (1967) contains an option for either of the average methods.

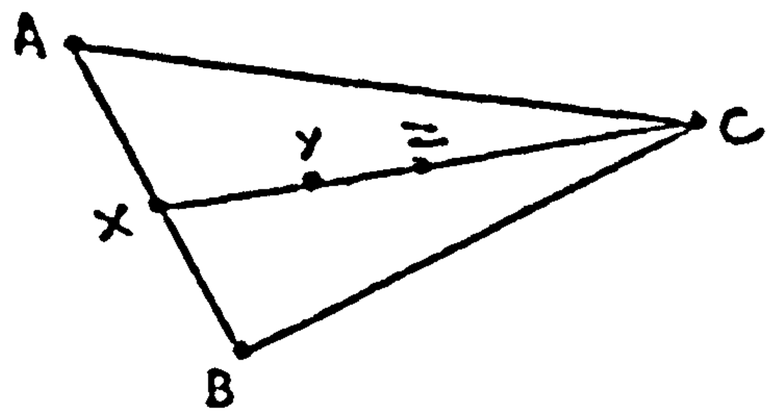
Our flowchart can be adapted for the group average method by an extension of the use of the array N, which previously was 0 for a redundant object and 1 otherwise. N is now added at each fusion and thus contains the number of objects in the group and so the statement N(P) = N(Q) + N(P) is included immediately before the statement N(Q) = 0. Also in the previous stage the group average fusions are given by

$$D(P,J) = \frac{N(P)*D(P,J) + N(Q)*D(Q,J)}{N(P) + N(Q)}$$

11. Centroid

The centroid is another hierarchical method of a similar type to the group and weighted average methods and is attributed to Sokal and Michener (1958). The algorithm proceeds in a like manner to the previous methods, except that as successive objects merge the new object associated with newly formed group is taken as the centroid of all the objects in the group.

This can be explained by a simple geometric analogue in two-dimensional Euclidean space. Consider the 3 points A, B, C.



When A and B merge their centroid is at the point X midway between A and B, and when C joins the group the group is represented by the point Y a third of the way along XC, the centroid of the triangle ABC. This is different from the group average method since when A and B merge, their similarity is the average of the similarities with other points, which is not equivalent to replacing A and B by a point X. With the weighted average method the 3 points are represented by the point Z at the midpoint of XC.

In order to form the new vector when A and B have merged to X we must consider the length CX; it can be shown that if

we use Euclidean distance squared as our dissimilarity measure, by use of the cosine rule in triangles ACX and ACB, that:

$$D(C,X) = \frac{D(A,C) + D(B,C)}{2} - \frac{D(A,B)}{4}$$

and further that

$$D(C,KL) = \frac{n_k D(C,K) + n_e D(C,L)}{n_k + n_e} - \frac{n_k n_e D(K,L)}{(n_k + n_e)^2}$$

And the method may be used with other dissimilarity or similarity measures.

The above expression can be incorporated into our hierarchical algorithm as an alternative to the previous methods.

The geometrical properties of the method are shown in Proctor (1966) and Gower (1967). The method has been used in Boyce (1964), Campbell et al (1970), Watson et al (1966) and Williams et al (1969). The sorting strategy is used with an information statistic, forming the Information Analysis of Williams et al (1966) which has been used by several authors, especially in ecology.

12. Median

This method was developed by Gower (1967) as an unweighted version of the centroid method. If we set $n_i = n_j$ in the centroid expression we obtain:

$$D(P',J) = \frac{D(P,J) + D(Q,J)}{2} - \frac{D(P,Q)}{4}$$

The fact that this and the previous methods could be performed by very similar algorithms was discovered by Lance and Williams (1966, 1967) (they omitted weighted average, but it is clearly able to be performed by the method). The simple flowchart for the method is given in Wishart (1969).

Lance and Williams gave the following linear relationship, which, by different choice of parameters could be used for the hierarchical methods explained so far:

$$D(P',J) = \alpha_p D(P,J) + \alpha_q D(Q,J) + \beta D(P,Q) + \gamma D(P,J) - D(Q,J)$$

The values of α_p, α_q, β and γ which give the methods are as follows:

		α_p	α_q	β	γ
1	Nearest neighbour	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
2	Furthest neighbour	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
3	Weighted average	$\frac{1}{2}$	$\frac{1}{2}$	0	0
4	Group average	n_p/n_k	n_q/n_k	0	0
5	Centroid	n_p/n_k	n_q/n_k	$-\alpha_p\alpha_q$	0
6	Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0

The median method is the most recent of the six methods, and has not been widely used. It has, however, been used in a recent pedology paper - Campbell, Mulcahy and McArthur (1970).

13. Flexible

The linear function shown above lead Lance and Williams to consider the effect of using other values for the parameters, and their possible use as cluster methods. They

constrained the possibilities by including the constraint that the measures used should be monotonic (this means that as the fusions take place the value of the dissimilarity function should increase continually, or if similarities are used, decrease continually). This can be stated, with the notation previously used, as the requirement that -

$$D(P',J) \geq D(P,Q) \quad \text{for all } J$$ (where P' is the group formed by the junction of P and Q)

This lead them to the sufficient monotonicity requirements -

$$\alpha_p + \alpha_q + \beta \geq 1, \quad \gamma = 0$$

(This can be shown to be true quite simply -

$$D(P',J) = \alpha_p D(P,J) + \alpha_q D(Q,J) + \beta D(P,Q)$$

$$\text{or} \quad \frac{D(P',J)}{D(P,Q)} = \alpha_p \frac{D(P,J)}{D(P,Q)} + \alpha_q \frac{D(Q,J)}{D(P,Q)} + \beta$$

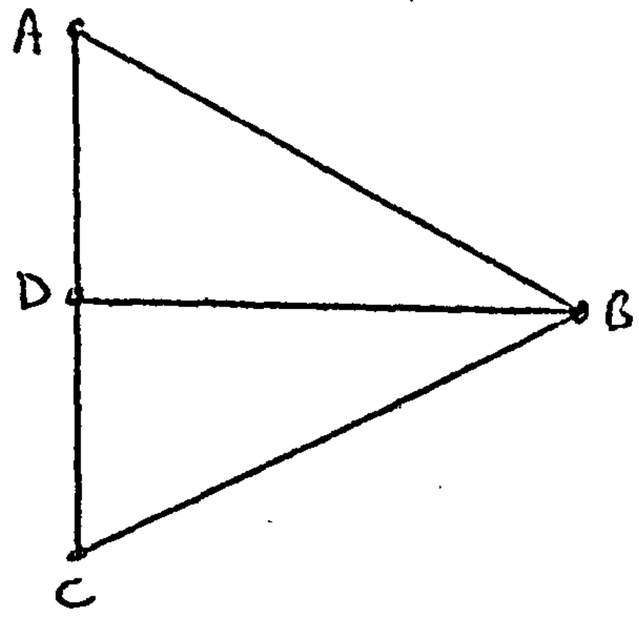
But $D(P,J) \geq D(P,Q)$ and $D(Q,J) \geq D(P,Q)$ because otherwise P and J or Q and J would have fused before P and Q.

$$\text{Hence} \quad \frac{D(P',J)}{D(P,Q)} \geq \alpha_p + \alpha_q + \beta$$

But the r.h.s. is 1, hence D is monotonic.)

The fact that the single link and complete link methods are monotonic and have $\gamma \neq 0$ shows that these are not necessary constraints.

The centroid and median methods are not always monotonic, this can be shown by a simple example. Consider the three points:



AC = 10
AB = 11
BC = 11
BD = 9.8

Of the three points A and C are the nearest (10 apart) and link first, and in the centroid and median methods are replaced by a point at D, halfway along the line AC. The next link to occur is between B and D which are only 9.8 units apart.

Non-monotonicity is however not as serious a problem as it is regarded by Lance and Williams who restrict their flexible strategy only to monotonic cases. The non-monotonicity of the centroid and median methods is useful in that it helps pick out clusters, for example in the above diagram the drop in the objective function (called a 'reversal' by Lance and Williams) can be interpreted as meaning that a split of the three points into two groups is not meaningful. In contrast to Williams, Lambert and Lance (1966), who state that "the presence of reversals may confuse the stratification at certain points" the present author believes that they aid the cluster analyst in selecting clusters.

Having derived the constraint $\alpha_p + \alpha_q + \beta \geqslant 1$, Lance and Williams make this an equality condition and further to the constraint $\gamma = 0$ they include the constraint $\alpha_p = \alpha_q$

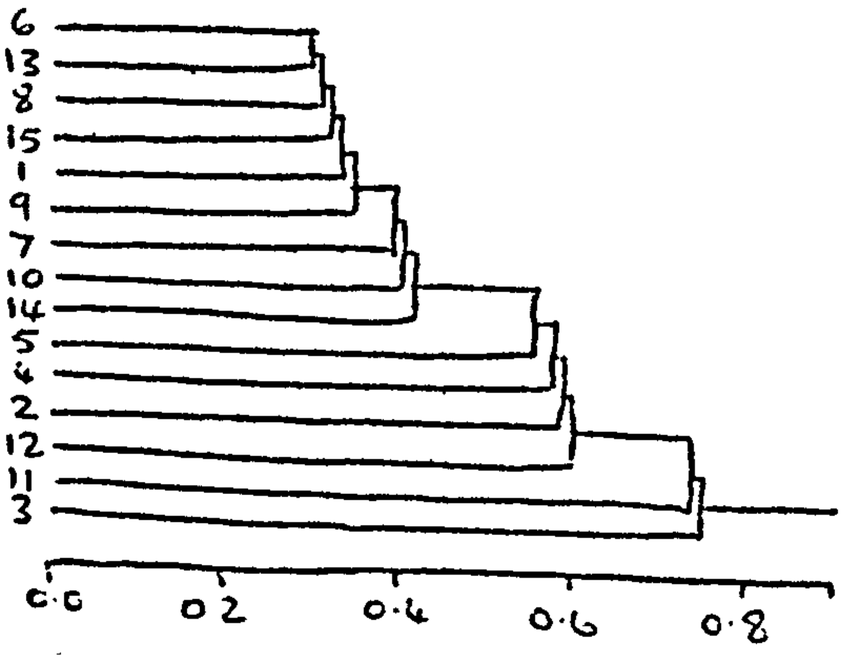
which is a requirement for a weighted method and also the constraint $\beta < 1$ (which constrains α_p and α_q to be strictly positive - a reasonable requirement) and so the flexible method considers cases under the quadruple constraints.

$$\alpha_p + \alpha_q + \beta = 1, \quad \alpha_p = \alpha_q, \quad \beta < 1, \quad \gamma = 0$$

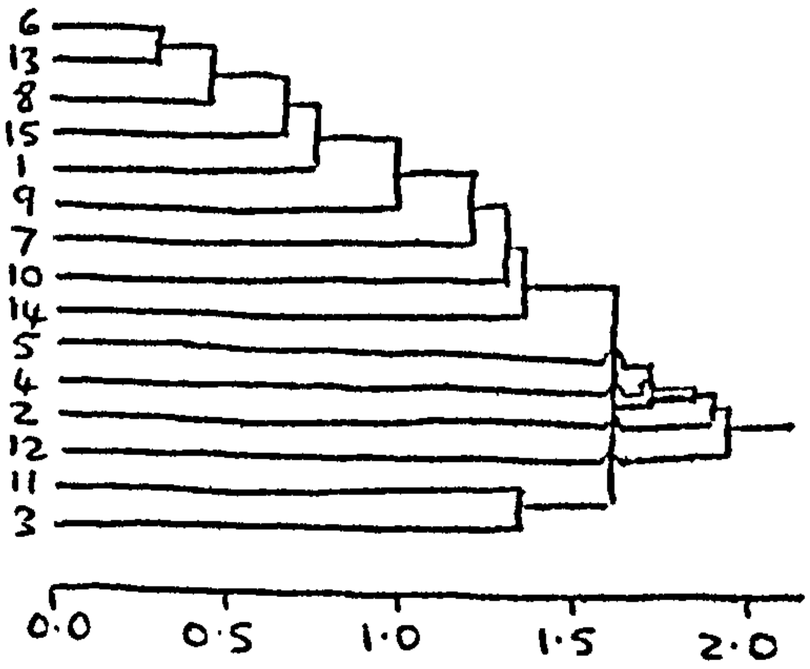
Thus by varying β subject to these conditions, a whole range of strategies can be used.

For a positive β as groups form, the original space in which the objects lie will be distorted in such a way that a newly formed group will appear to move 'nearer' some or all of the remaining elements (this is termed space-contraction), which causes chaining to occur. As β becomes smaller the method becomes more 'space-dilating' and as groups form they move away from other elements. Thus Lance and Williams "do not expect any requirement for the flexible strategy with positive " and have suggested that the point at which space-contraction and space-dilation are matched (space-conservation) is at $\beta = -0.25$. (Note that the flexible strategy with $\beta = 0$ is the weighted average method.)

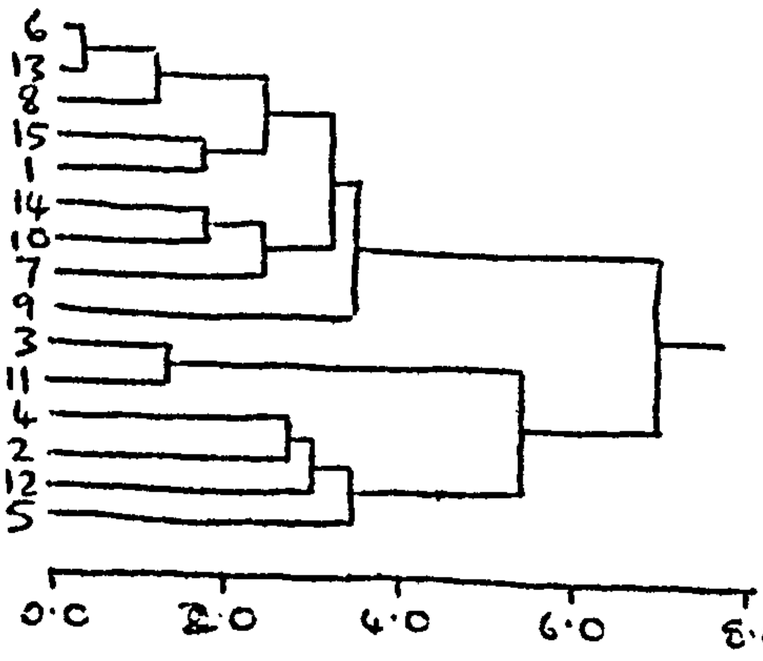
The effect of changes in β on a set of data can be illustrated by the following dendrograms, of the clustering of 15 random points unformly distributed in a unit circle.



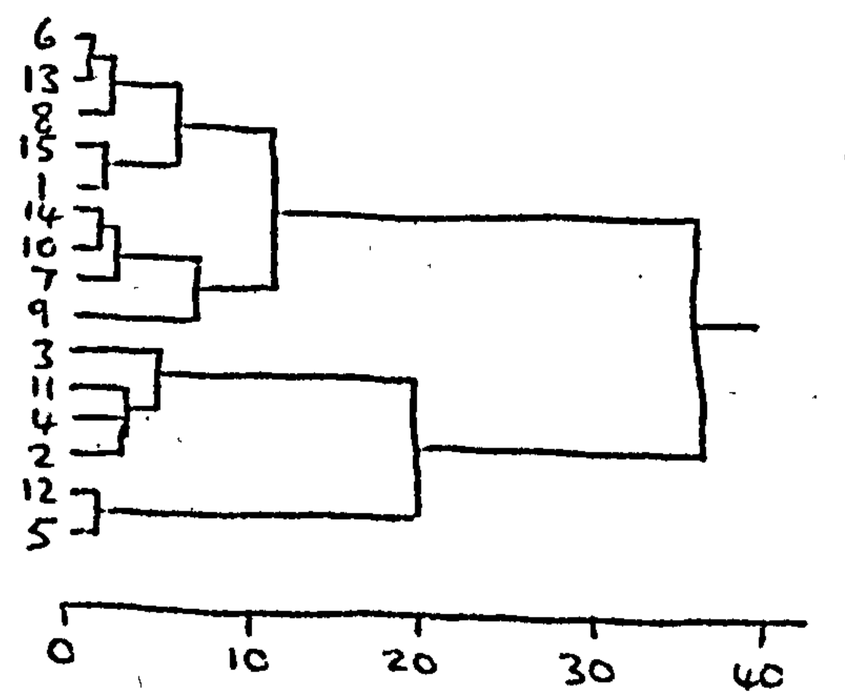
$\beta = 0.99$



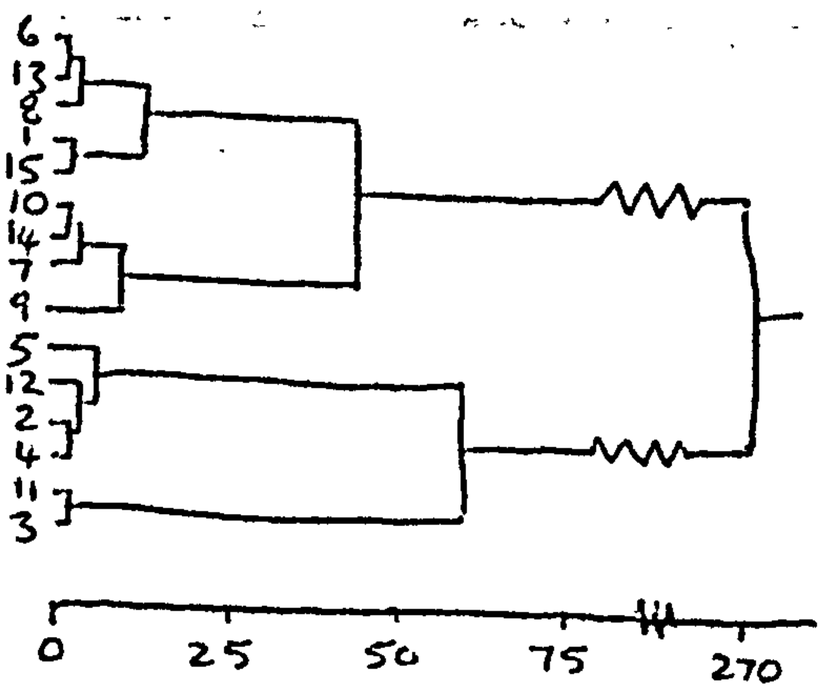
$\beta = 0.90$



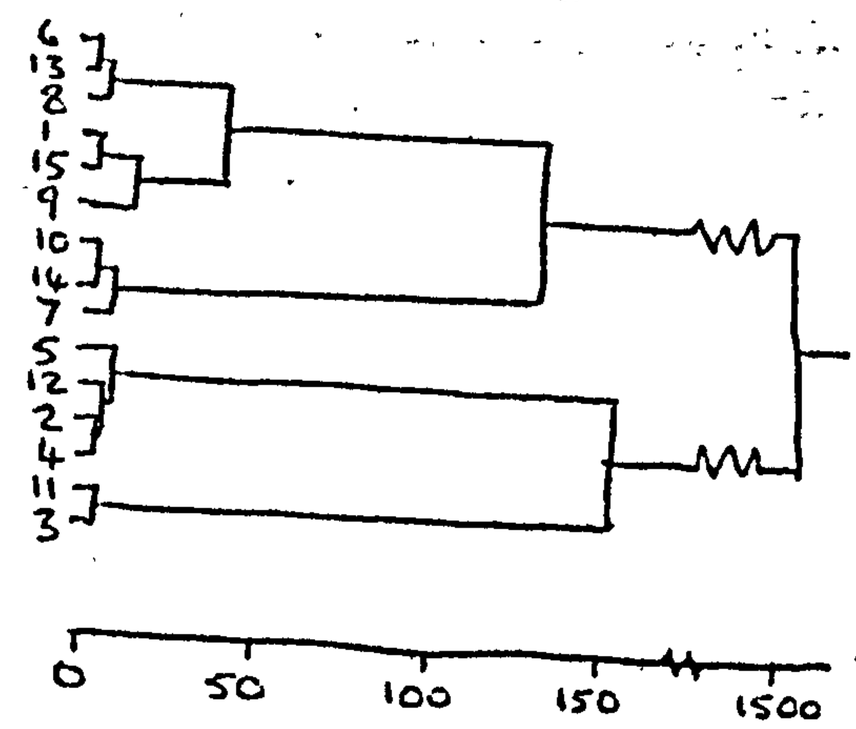
$\beta = 0.50$



$\beta = 0$



$\beta = -0.50$



$\beta = -1.0$

It can be seen that with $\beta = 0.9$ and $\beta = 0.99$ that chaining is present to a high degree and in fact the data can be made to chain completely by taking β sufficiently close to 1. This is caused by the fact that the method is extremely space conserving with high β; once the first two elements merge they 'move' so much closer to all other points, and thus the next fusion is almost certain to involve the first pair.

With a low value of β (-1.0 or less) the method becomes space contracting to such an extent that once groups join they 'move' so far away from the other objects that the next fusion excludes them, this is the opposite effect to chaining, which we shall call pairing. Small groups (which have hence had less fusions) are more likely to fuse than larger groups - in the extreme case elements will first join in pairs and then pairs of these pairs will combine and so on - hence the term pairing. Pairing causes resultant clusters to be of similar sizes.

However, in the above dendrograms it can be seen that for a wide range of β around zero the dendrograms are very similar apart from a stretching of the similarity axis.

13A. Extension of Flexible

With Lance and Williams' linear relationship

$$D(P',J) = \alpha_p D(P,J) + \alpha_q D(Q,J) + \beta D(P,Q) + \gamma \left| D(P,Q) - D(Q,J) \right|$$

we could consider any values of the four parameters. However we have concentrated on the case where group weights

are not used and hence $\alpha_p = \alpha_q$. Also since γ weights members of a group differently we have set $\gamma = 0$. Thus we have two parameters to vary with one another, α_p and β. Lance and Williams constrain these by the relationship $2\alpha_p + \beta = 1$ which eliminates reversals. However the only valid criticism of reversals is that they make the dendrogram difficult to draw and analyse. Reversals can however be of use, because they can emphasize that points are in the same group - for example in the illustration on page 159 we could not consider A and C as a single group apart from point B. Thus the extended flexible can be used as a cluster method, which does not necessarily produce a dendrogram.

14. Ward's Method

This hierarchical method was proposed by Ward (1963) (also see Ward and Hook 1963) in the field of personnel research, and possibly because of this has not had the more widespread use as the methods previously described which originated in numerical taxonomy. The method has also been independently put forward by Orloci (1967). The method proceeds agglomeratively by uniting elements or groups so as to minimize the overall within-cluster variance at each fusion stage. Another reason why the method has not had wide use is that for some time it remained a more time-consuming method than the other hierarchic methods. However it has recently been shown by Anderson (1966, 1971a), Wishart (1969c) and Burr (1970) that the method can be incorporated into the framework of Lance and Williams' linear relationship, by use of the values -

$$\alpha_p = \frac{n_i+n_k}{n_i+n_j+n_k}, \quad \alpha_q = \frac{n_j+n_k}{n_i+n_j+n_k}, \quad \beta = 1-\alpha_p-\alpha_q, \quad \gamma = 0$$

Note that these values give monotonicity, and that a weighted version of the method would give $\alpha_p = \alpha_q = 2/3$, which is the flexible method with $\beta = -1/3$.

The method has a tendency for small groups to link rather than large, because they cause less increase in the total within cluster variance. This can be shown as follows -

$$\beta = 1 - \frac{n_i+2n_k+n_j}{n_i+n_j+n_k}$$

which replacing n_i+n_j by n_e, the size of the newly formed group, becomes -

$$\beta = \frac{-n_k}{n_e+n_k}$$

and as $\frac{n_k}{n_e} \to 0$ $\beta \to 0$

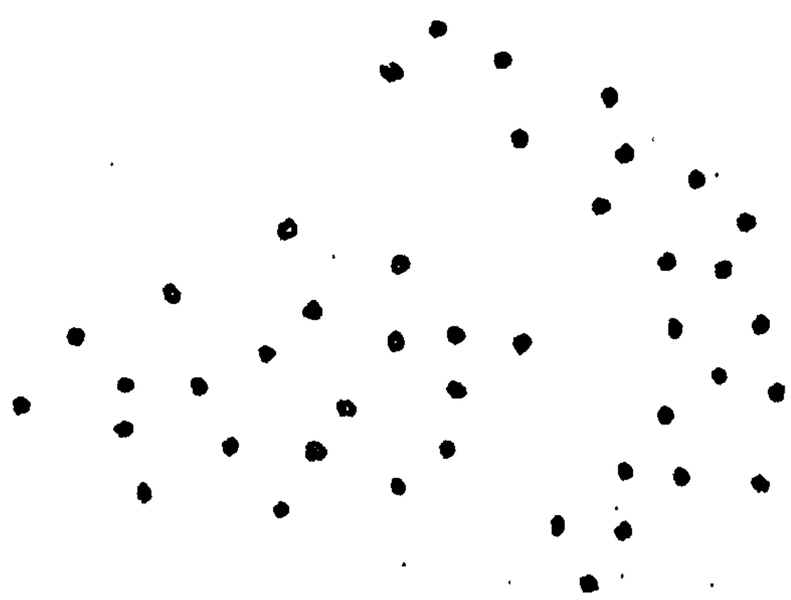
and as $\frac{n_k}{n_e} \to \infty$ $\beta \to -1$

Thus a new group becomes further from larger groups and nearer smaller groups.

The method has been used by Goddard (1970) in land use studies, by King (1969) in geography, and discussed by Ling (1971c). The eight methods examined so far are included in a published computer program by Wishart (1969b). Our program is given in Appendix 1.

15. Single-Link on K-link Criteria

Some cluster methods are designed to search for groups which have the property that all points in a group are more similar to their cluster centre than to any other cluster centre (these methods have been termed by Wishart 1969d minimum variance methods). However, reasonable as this requirement may seem, consider the following groups:



Here, two definite clusters are present, but the requirement clearly does not hold. These types of group have been called 'natural' groupings, which are simply dense swarms of points which may be of any shape. The two types of groups, round and natural have been considered by Cattell and Coulter (1966) and termed 'round' and 'straggly'. However the distinction is not always so definite as methods may allow varying degrees of 'straggliness'. The method of single linkage would be able to pick out the two groups in the configuration above without difficulty, but a single freak point between the two groups would be enough to make the method fail. The single link method is an example of the danger of allowing a large degree of straggliness - there is a tendency to produce the chaining effect.

In a particular study, if the variables are not weighted correctly or irrelevant variables are included, then any groups which were expected to be hyperspheres will become hyper-ellipsoids, and may not be found by minimum variance methods.

The fact that the single link method is able to find straggly groups so well, leads us to the possibility of adapting the method to constrain the chaining effects.

This proposed method, the single-link on K-link criteria method (Klink) is an attempt to do this. The method proceeds by an agglomerative algorithm in which a point is joined to its nearest neighbour if the average distance to its first K nearest neighbours is less than a particular threshold. This threshold is increased until all objects are in the same set. For K = 1 the method reduces to the single-link method.

The program for the method can be explained in more detail, as follows: the program stores for each point I, the average distance to the points first K nearest neighbours DK(I), and the point number of its nearest neighbour N(I). The point J with the least average distance to its neighbours DK(J) is found and is joined to its nearest neighbour N(J) = K say. For the point J, DK(J) and N(J) are recalculated, excluding the point K. Similarly for the point K, DK(K) and N(K) are recalculated excluding point J. The vector DK is then re-examined and the element with the smallest value is found, this point is joined to its nearest

neighbour and DK and N are recalculated for each particular point excluding points which belong to the same group as that point. This procedure continues until all points are in one cluster.

A slight difficulty is encountered when a group is formed which has less than K other points exterior to that group. An approximation may be made in these situations, to adjust for this, based on a randomly distributed population. This adjustment is based on the finding of Thompson (1956) that the distribution of distances to neighbours of all orders are related to the χ^2 distribution.

The flowchart for the Klink method is given in Figure 24.

All of the previous methods, with the exception of nearest and furthest neighbour, can be said to be based on distances between group centres, but this method is of the density type - a point is linked to others if it lies in the densest region.

16. Nucleus Method

This method is designed to find either overlapping or non-overlapping groups of any shape without excessive chaining. Objects can either belong to the nucleus of a single group or be associated with any number of groups. The method is hierarchical and similar to the Klink method in that it is based on density. The nucleus is formed by accumulating points which are near neighbours and which have the same K other points as near neighbours.

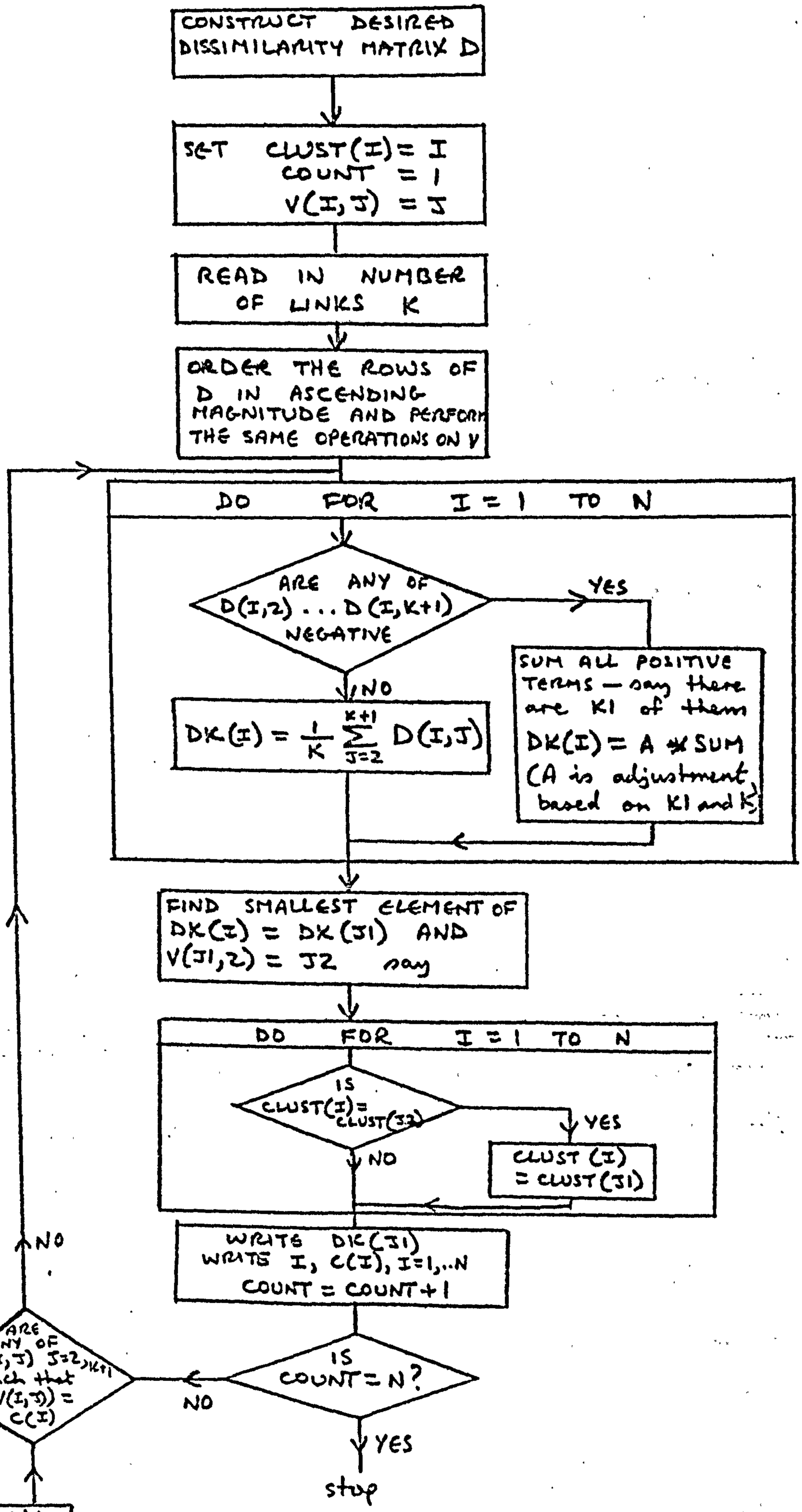

CONSTRUCT DESIRED DISSIMILARITY MATRIX D
SET CLUST(I) = I
COUNT = 1
V(I,J) = J
READ IN NUMBER OF LINKS K
ORDER THE ROWS OF D IN ASCENDING MAGNITUDE AND PERFORM THE SAME OPERATIONS ON V
DO FOR I = 1 TO N
ARE ANY OF D(I,2) ... D(I,K+1) NEGATIVE
YES
NO
SUM ALL POSITIVE TERMS — say there are K1 of them
DK(I) = A * SUM
(A is adjustment based on K1 and K)
$DK(I) = \frac{1}{K} \sum_{J=2}^{K+1} D(I,J)$
FIND SMALLEST ELEMENT OF DK(I) = DK(J1) AND V(J1,2) = J2 now
DO FOR I = 1 TO N
IS CLUST(I) = CLUST(J2)
YES
NO
CLUST(I) = CLUST(J1)
WRITE DK(J1)
WRITE I, C(I), I=1,..N
COUNT = COUNT + 1
IS COUNT = N?
YES
stop
NO
ARE ANY OF V(I,J) J=2,K+1 such that C(V(I,J)) = C(I)
YES
NO
eliminate this V(I,J) + move all following V(I,J) along and put last element = -1.0

FIGURE 24

The program begins by searching the dissimilarity matrix for the smallest element, say D(J1, J2). The pair J1 and J2 are now said to be 'linked'. A matrix LINK(I, J), (I=1),...N; J=1,...N-1) stores the linked points (- LINK begins with all entries zero and when J1 and J2 are linked, the first zero entry in row J1 is made equal to J2 and the first zero entry in row J2 is made equal to J1). Then the matrix LINK is searched to see if any pair, which are themselves linked, have K or more linked points in common, in which case they are clustered together and form a nucleus. Since in any one cycle there are only two new entries to LINK there are only 2(N-2) + 1 = 2N-3 pairs which can cluster together in a cycle, and thus the number of searches can be reduced. The procedure continues by linking the pair with the next highest similarity, the linkages are entered into LINK, which is then searched again to see if any points amalgamate. When all the points form one group the program terminates. The points which are clustered together form the nucleii at any stage, and the associate elements to the nucleii are found from the matrix LINK which may be printed out at any stage if desired.

The flowchart for the method is shown in Figure 25.

When non-overlapping groups, as points are clustered to form nucleii they are output and this produces a dendrogram. For overlapping groups, or simply for more information, the matrix LINK is printed each time points join nucleii. From

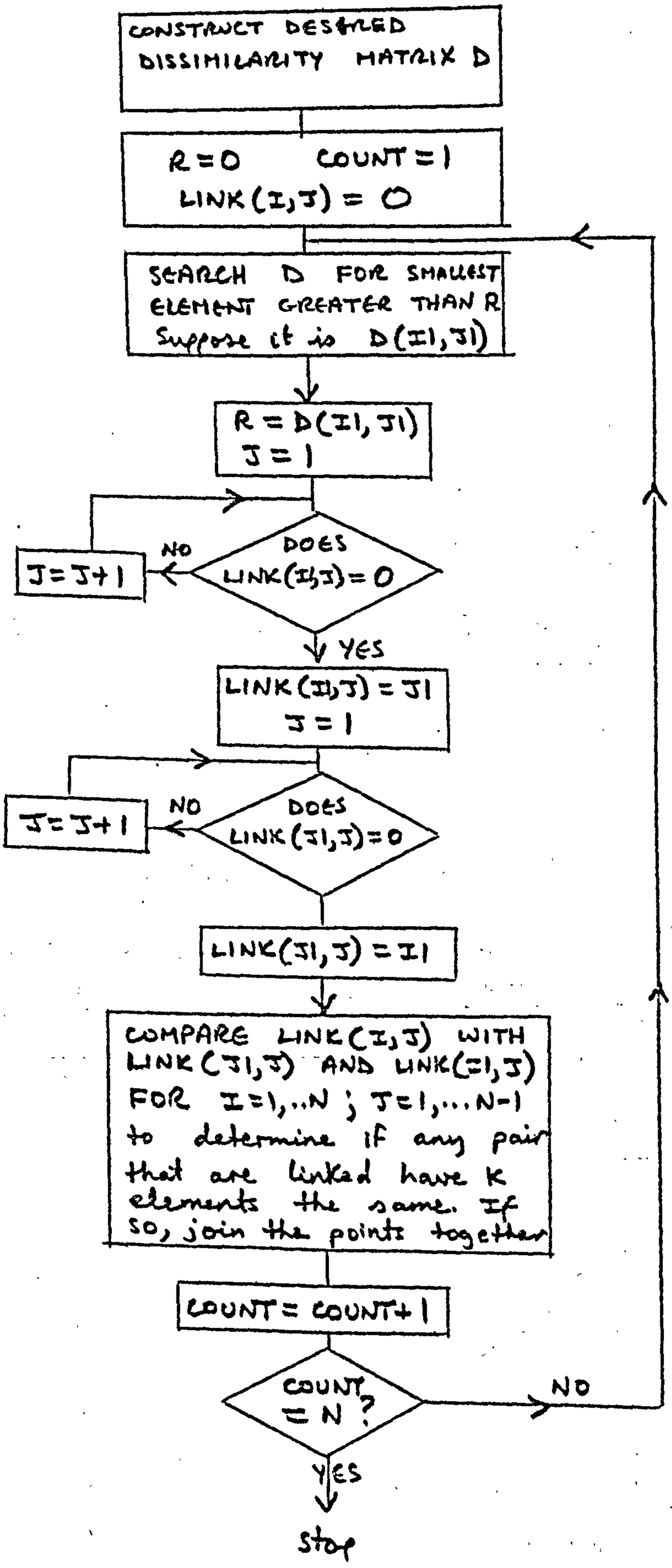
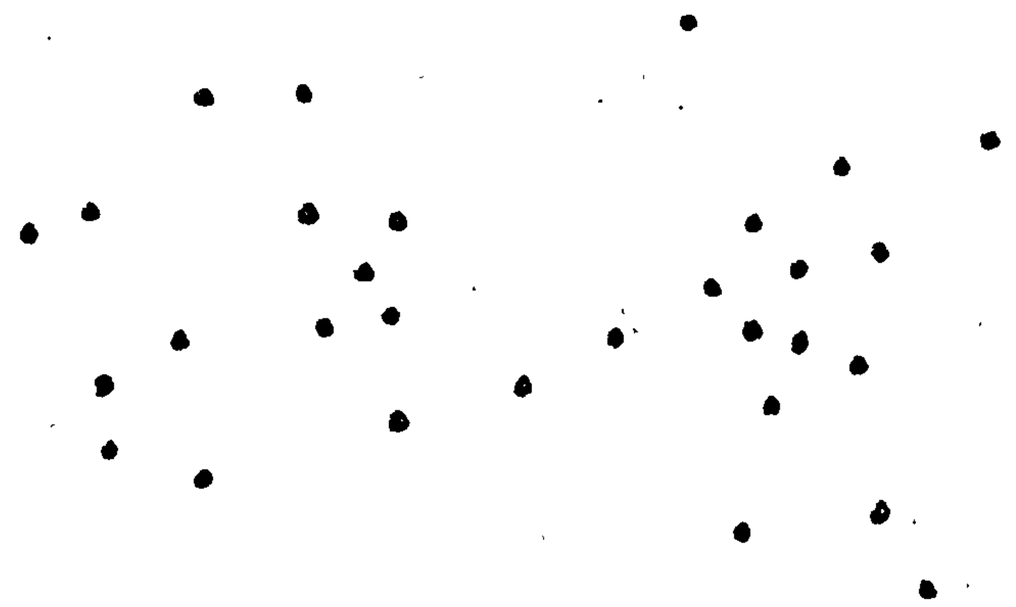

CONSTRUCT DESIRED DISSIMILARITY MATRIX D
R=0 COUNT=1
LINK(I,J) = 0
SEARCH D FOR SMALLEST ELEMENT GREATER THAN R Suppose it is D(I1,J1)
R = D(I1, J1)
J = 1
DOES LINK(I,J)=0
NO
J=J+1
YES
LINK(I,J) = J1
J = 1
DOES LINK(J1,J)=0
NO
J=J+1
LINK(J1,J) = I1
COMPARE LINK(I,J) WITH LINK(J1,J) AND LINK(I1,J) FOR I=1,..N ; J=1,...N-1 to determine if any pair that are linked have K elements the same. If so, join the points together
COUNT = COUNT+1
COUNT = N ?
NO
YES
Stop

FIGURE 25

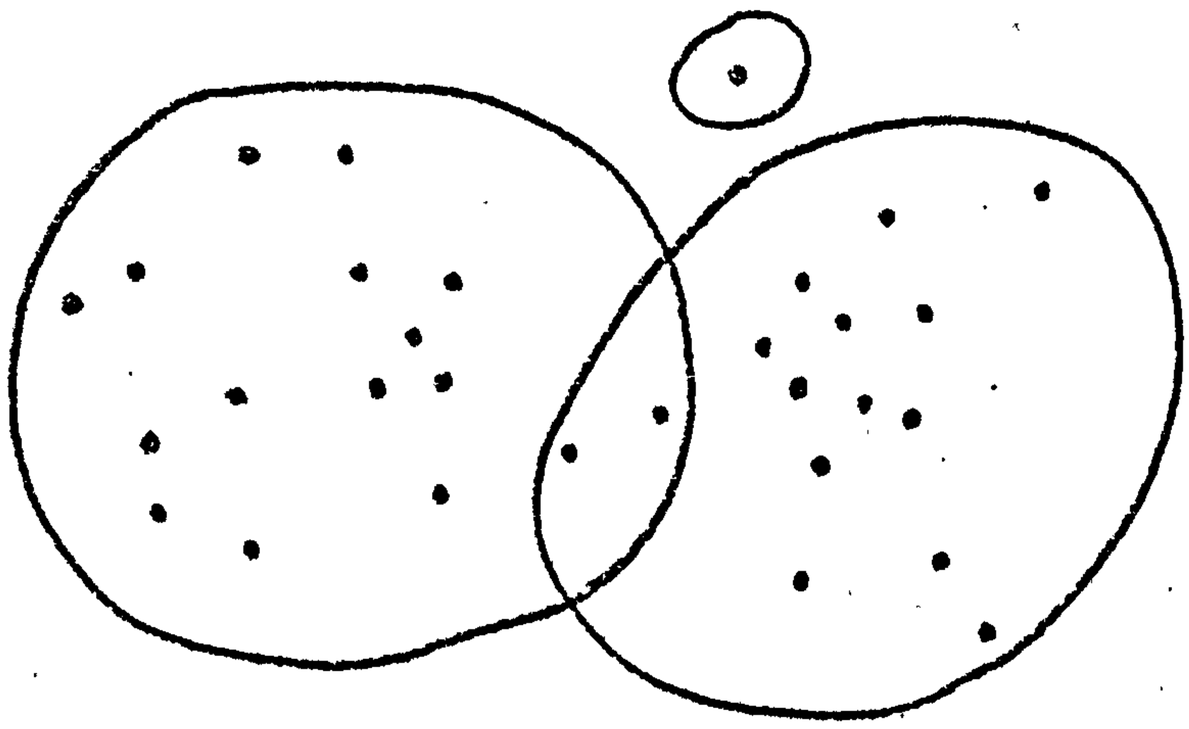
this the nucleus can be obtained together with points outside which are linked to nuclear members, and thus overlapping groups are formed.

Example

Consider the points:



There is clear evidence of two groups being present but the exact membership of the groups is uncertain. The nucleus method gives the following result:



The two groups are clearly depicted with two points in the centre which may belong to (n)either group and a single outlier.

Other overlapping methods are discussed later in this section.

17. Dissimilarity Analysis

All of the techniques discussed so far are hierarchical techniques which have been traditionally regarded as being performed by agglomerative algorithms, but as Jardine and Sibson (1971) have pointed out, the agglomerative nature of the algorithm is not a property of the method. However, some techniques exist which by their very nature require divisive algorithms.

Dissimilarity analysis is one such technique, suggested by McNaughton-Smith (1964, 1965), indeed Cormack (1972) has stated that it is "the only feasible suggestion for a polythetic (i.e. dependent on several attributes) divisive technique". The method has been little used in published studies.

Although the method is termed dissimilarity analysis, it can easily be adapted for use with similarities. The procedure is based on a measure of dissimilarity between groups. First the object is found which is most dissimilar to the group formed by the remaining objects. I.e. if we let the set of objects be X then for all $I \epsilon X$ we find MAX(D(I,X-I)) say this is A. Then for each $J \epsilon (X-A)$ we find MAX(D(J,X-A-J)-D(J,A)), and if this is positive then this object joins the set A, thus we have found a second point which is more similar to the first point than the remainder of the group. This procedure of finding MAX(D(J,X-A-J)-D(J,A)) is then repeated until it drops below

zero. Thus two groups have been formed. The analysis can thus be repeated for all sub-groups, until only one element exists in each group.

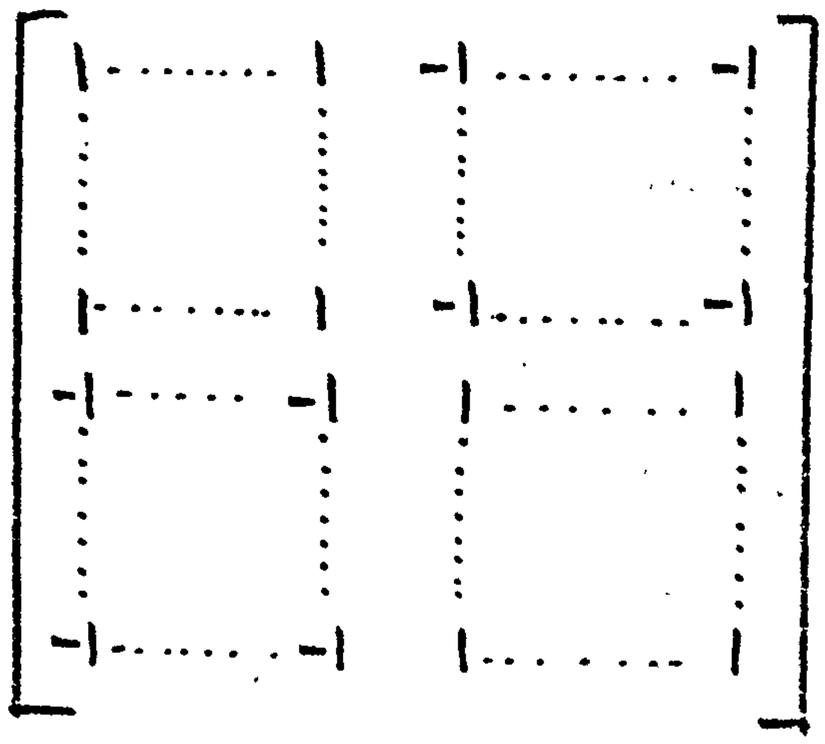
In the original method the dissimilarity measure used was 'objectively weighted squared Euclidean distance' $\sum((x_{aj}-x_{bj})^2 \sum x_{jk}^2)$. The paper also suggests a 'practical modification' - that of simply finding MIN(D(J,A)) provided that MAX(D(J,X-A-J)-D(J,A)) is positive.

In the present study the method has been used with the measure of Euclidean distance squared, and the group centre at any stage is taken to be the group average. A measure is also required of the cohesiveness of any group which may be formed in order to detect clusters, since the method will divide random data. The measure used in this study was the average value of the MAX(D(J,X-A-J)-D(J,A)) for all the points which are in the group.

18. Profile Clustering

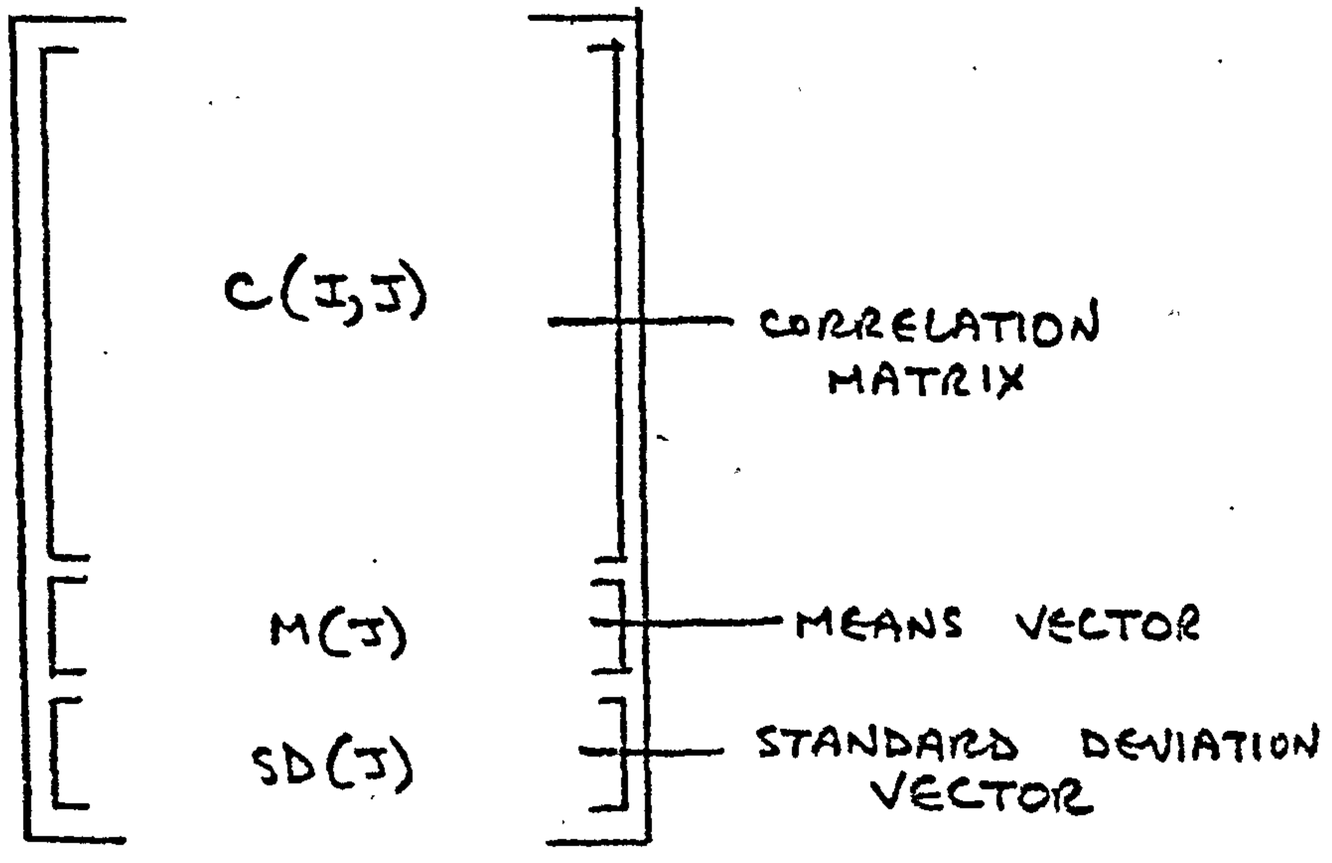
Another method which is, by nature, divisive is profile clustering. This little known method has been described by McQuitty (1967b) under the name of Intercolumnar Correlational Analysis, and is related to the 2nd order method of Tryon. (It has been used by Seymour et al 1973.) The original method is based entirely upon the correlation measure. Once an initial correlation matrix is produced the method proceeds by forming the correlation matrix between columns of the initial correlation matrix, and then using this as a new correlation matrix proceeds to find the

correlation matrix between columns of this matrix. This process is continued until a limit is reached. Although unproven, the limit in all cases is a matrix, all of whose elements are either ± 1, or 0. In all but extreme cases the limit is reached where all elements are ± 1 and the matrix can be rearranged to the form:



Thus the original observations have been split into two groups. The conclusion that this has split the population into two homogenous clusters is based on the idea that if 2 items are similar they will have similar relationships with other items (a notion similar to Tryon's - see page 126). The 2 subjects can further be divided by the same method. In practice the number of iterations taken for the matrix to have all elements $<-.99$ or $>.99$ is nearly always less than ten, but weaker conditions such as calculating until the signs of the elements of the matrix stabilize would cause less computation (but an increase in program complexity). (The method as programmed uses the limit of all elements $<-.9$ or $>.9$.)

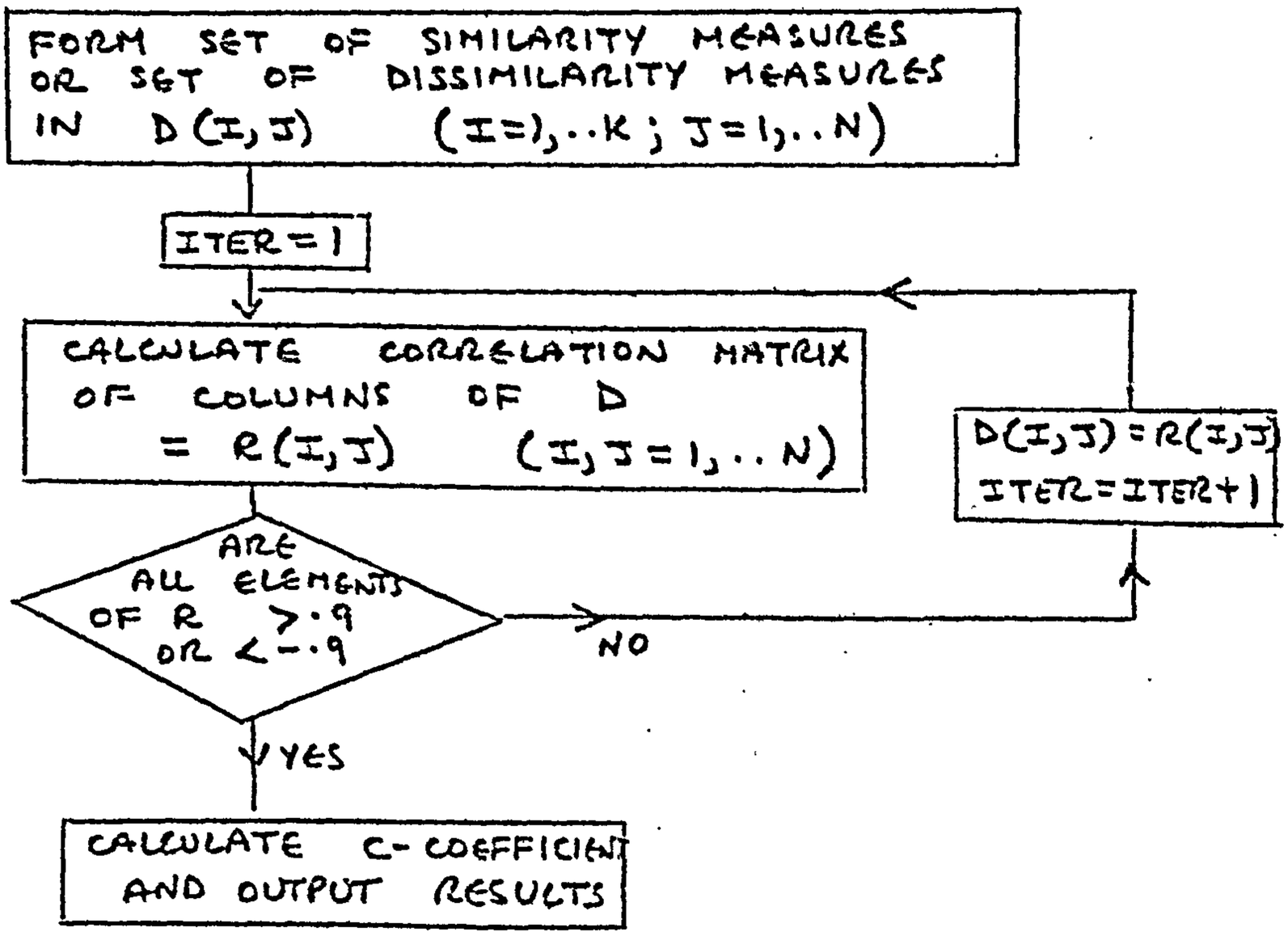
The method may be extended to any set of similarity measures by forming the matrix of correlations between rows of the similarity matrix. In fact more than one similarity matrices or vectors may be used at the same time. For example the following matrix may be used:



and weightings could be employed on the sub-matrices or any row or set of rows in the matrix. Also once a split has taken place the new initial matrix could be obtained by simply selecting the required columns of the original matrix, or by selecting certain rows of the required columns.

A further necessary improvement to the method is in order to facilitate determining the number of clusters present - the number of iterations required to reach a limit is an indication of how 'easy' a split is, but this does not vary over a very wide range (although it is useful for additional information). The index of group cohesion we propose is a variation of the Holtzinger B-coefficient ((see page 122) - the C-coefficient described on page 126) which is specifically designed for correlation matrices.

Flowchart



Other Hierarchical Methods

The previous twelve methods which have been described in detail are all (with the exception of the extended flexible method) designed to produce a dendrogram and are hence called hierarchical methods. The centroid and median methods can produce reversals which complicate the dendrogram, and in fact do not produce an ultrametric approximation to the original data. However these reversals are not common and do not involve large drops in the objective function. The extended flexible method can also produce reversals with certain parameters, and can give large drops in the value of the objective function. In these cases the methods cannot be considered as methods which produce a dendrogram, but are still cluster methods. The methods of dissimilarity analysis

and profile clustering do form hierarchies, but there is no simple measure by which one can scale the dendrogram.

The process of forming a hierarchy from a dissimilarity matrix can be considered as that of producing a 'best' approximation to the matrix in an ultrametric space. This has been shown by Jardine, Jardine and Sibson (1967) (also S. Johnson 1967 and Hartigan 1967).

A procedure based on gradually transforming the matrix into one obeying the ultrametric properties has been put forward by Roux (1969). The method proceeds to find the largest ultrametric matrix with all elements less than those of the distance matrix, by iterative reduction of the greatest side of the triangle formed by each set of three points. However he has stated (Roux 1972, personal communication) that the method "is now abandoned because its results are rather less accurate than those given by the classical average linkage method of Sokal and Sneath as an example". This method has also been suggested by Hartigan (1967) who uses a combination of heuristics for reassignment, etc. This is however very slow - taking 16 minutes to analyse 50 objects.

Perhaps the most certain method of finding clusters would be to consider all possible groups; however this becomes computationally infeasible for more than about 12 objects. The method of Edwards and Cavalli-Sforza (1965) proceeds by considering all possible divisions into 2 subsets and finding that division which maximizes the between set sum

of squares, then all possible divisions of these sets into two more each and so on. This method is very labourous - Gower (1967) gives computer times of 100 hours with 21 objects and 54,000 years with 41 objects. Scott and Symons (1971) show that not all combinations need be attempted if convex clusters only are required. Calinski (1969) has also tried to improve on Edwards and Cavalli-Sforza's method.

Fisher (1958) uses enumeration in the 1 dimensional case and because of the additive properties of within-group sum of squares shows that by a system of record keeping a lot of calculations need only be done once, and hence the computer time is short. This case is also simplified since groups must be contiguous.

Sheperd (1966) mentions that the most important part of a cluster is the densest part, and that this might not be central in the cluster. He suggests the use of single link to find dense centres and then group average to allocate other points.

However, as most agglomerative methods proceed initially very similarly to single link this seems unnecessary. Also there is danger of chaining if the single link part is not terminated early enough. It seems that differences from the results with group average alone would not be large.

Neely's method as used and described by Lankford (1969) is a type of single-link method which does not require complete storage of the distance matrix. Only simplicial

neighbours to each point are evaluated - these are formed by linking points sequentially by lines until no more can be linked without lines crossing each other - linked points are then called simplicial neighbours, (these are not unique and depend on the order of linkage, also the definition extended to more than two dimensions would seem to include all links). The method proceeds by the single-link method but only simplicial neighbours can be linked. The only advantage of this method over single link which is stated is that the storage requirements are smaller, however single link can be programmed with even less storage than Neely's method.

Burr (1970) gives a suggested agglomerative method which could be incorporated into Lance and Williams' algorithm without much difficulty. Burr's method minimizes the within-cluster variance at each fusion stage. This means we use the transformation

$$D_{Jk} = (p_{ik}D_{ik}+p_{jk}D_{jk}+p_{ij}D_{ij}-n_iS_i-n_jS_j-n_kS_k)/p_{Jk}$$

where $p_{ik} = (n_i+n_k)(n_i+n_k-1)$

and s_i = sum of squares within cluster i.

The Heterogeneity Analysis of Hall (1967a, b, 1969) groups objects hierarchically according to the lowest value of his heterogeneity measure which is the ratio of the dispersion within the group to the dispersion of a dummy group having maximum bimodality within the boundary of the present group.

Tanimoto's method (Rogers and Tanimoto 1960, Tanimoto 1960, Rogers and Fleming 1964) groups sequentially attempting to optimize entropy. This is reviewed by Ornstein (1965) who suggests improvements.

McQuitty has introduced several methods, some of which are very similar to the methods we have already discussed. His methods are best reviewed in R. Johnston (1968).

1(b) ITERATIVE RELOCATION

19. Beale's Method

One of the disadvantages of hierarchical methods as cluster methods is that once points are joined together in the same cluster then they cannot be disunited. A class of methods which allow points to change their 'parent' cluster at various stages in the algorithm is called iterative relocation. Perhaps the most well-known method of this type is Beale's method (Beale 1969, Scicon 1971), sometimes called iterative r-location or Euclidean cluster analysis.

In the original paper the algorithm to perform the method did not calculate the full distance matrix but calculated distances as required. This reduces the computer storage required (although principal components analysis is sometimes necessary to reduce the effective number of variables). However this is not essential to the method, which can proceed from a full distance matrix.

The method is based on hyper-spherical clusters, and it is very similar to Ward's method of clustering. It uses the sum of squares of deviations (Euclidean distance squared) as a distance measure. The method proceeds agglomeratively by combining the nearest two clusters, then reallocating points to the cluster centres, and if any reallocations occur then new centres are found and the procedure continues until stabilization occurs (or until a fixed number, say ten, reallocations have been performed, since the method is not guaranteed to stabilize), and then the nearest two remaining

clusters are joined and so on. Provision can be made (Beale 1971) for block moves, moving several points between clusters at once, to reduce the possibility of local optima.

As the method increases calculation time by reallocating, recalculating centres, etc., Beale recommends using the method over the range of clusters in which one is interested, beginning with a random allocation to at least three more centres than the maximum one is interested in, in order to let the clusters 'settle down'.

An obvious extension of the method would be to use Ward's method to produce an initial configuration of clusters (see Lance and Williams 1967, Wishart 1971). Note: as Beale's method is simply Ward's method with provision for reallocating points, this involves simply 'switching off' the reallocating proceed for some iterations. However Wishart (1971) suggests from clustering two sets of data that the 'worst possible' solution of allocating every p^{th} object to the p^{th} cluster as an initial clustering produces a faster convergence to an optimum p-group solution than p carefully selected 'good' points. This conclusion is based on somewhat conflicting results, of the two sets of data used by Wishart, only with one set is a faster convergence produced, and this seems a rather small sample to take evidence on. Beale (1971) points out that this type of investigation is against the original advice of beginning with at least three more groups than required, and so is somewhat irrelevant.

With iterative location three factors virtually determine the method. Firstly the way in which two clusters are selected to be united, secondly the way it is determined if points are reallocated or not, and thirdly the way in which the cluster centre is defined.

With Beale's method the three factors are:

C(1) That pair which when combined produce the least increase in the within group sum of squares i.e. the minimum of

$$\frac{n_k n_l d_{kl}^2}{n_k+n_l}$$

where n_k, n_l are the number of points in clusters k and l respectively and d_{kl} is the distance between the centres of clusters k and l.

C(2) A point in cluster k is reallocated to cluster l if

$$\frac{n_l}{n_l+1} d_l^2 < \frac{n_k}{n_k-1} d_k^2$$

where d_l is the distance of the point to the centre of cluster l which has n_l members (similarly d_k, n_k).

C(3) The arithmetic mean of the cluster members i.e.

$$C_{ik} = \frac{1}{n_k} \sum_{i \epsilon k} x_{ij}$$

The method has been used to group British towns by Andrews (1971) and in gap analysis by Morgan and Purnell (1969). An example of the method used with Fisher's Iris data is given in Scicon (1971) and Hitchin (1970).

An important part of Beale's method is that it includes a test to choose the number of clusters. This is based on an F-test to decide if a division into C_2 clusters is significantly better than a division into C_1 clusters. The resulting statistic

$$\frac{R(C_1)-R(C_2)}{R(C_2)} \Big/ \left\{ \left(\frac{N-C_1}{N-C_2}\right) \left(\frac{C_2}{C_1}\right)^{2/n} - 1 \right\}$$

is treated as an F-ratio with $(n(C_2-C_1),\ n(N-C_2))$ d.f.

where $R(C_1)$ is the residual sum of squares in the C_1 cluster formulation (similarly $R(C_2)$, N is the number of objects, n is the dimensionality of the space in which the points lie. The calculation of n would seem to require the original variables to be reduced by factor analysis, otherwise if $n = \min(N-1, m-1)$ were used this would give an inflated value of n.

A flowchart for the method is given in Figure 26.

20. Group Average Relocation

As Beale's method is an extension of Ward's method, so we can incorporate iterative relocation into many hierarchical methods. Thus group average can be used as an iterative relocation method. This method has not been formally proposed but has been used by Wishart (1971b). Group centres are defined as in Beale's method. The criterion which is to be optimized is the within groups average distance. We have:

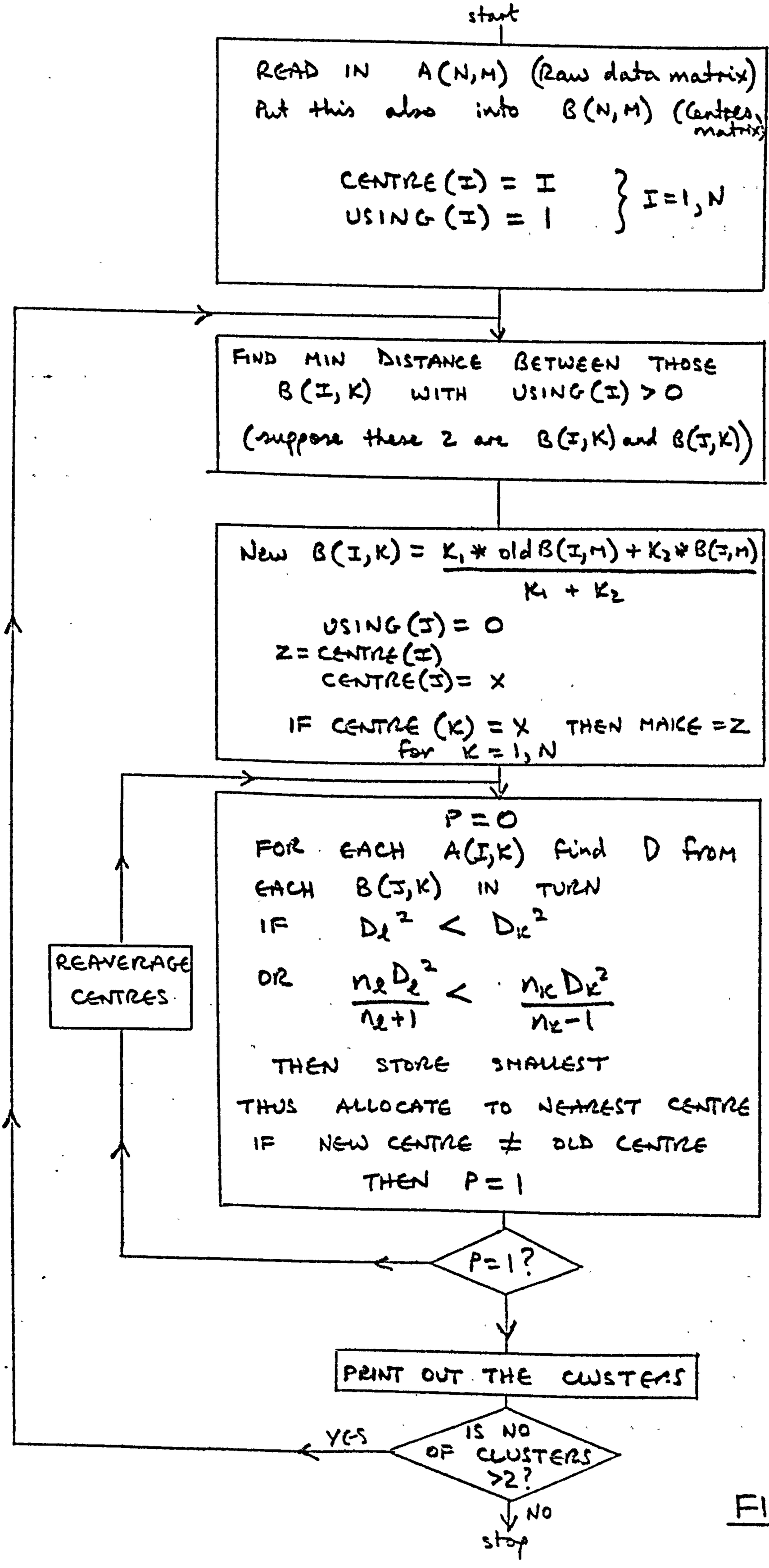

start
READ IN A(N,M) (Raw data matrix)
Put this also into B(N,M) (Centres matrix)
CENTRE(I) = I
USING(I) = 1
} I=1,N
FIND MIN DISTANCE BETWEEN THOSE B(I,K) WITH USING(I) > 0
(suppose these 2 are B(I,K) and B(J,K))
New B(I,K) = K1 * old B(I,M) + K2 * B(J,M) / K1 + K2
USING(J) = 0
Z = CENTRE(I)
CENTRE(J) = X
IF CENTRE(K) = X THEN MAKE = Z for K = 1,N
P = 0
FOR EACH A(I,K) find D FROM EACH B(J,K) IN TURN
IF D1² < DK²
OR n1 D1² / (n1+1) < nK DK² / (nK-1)
THEN STORE SMALLEST
THUS ALLOCATE TO NEAREST CENTRE
IF NEW CENTRE ≠ OLD CENTRE
THEN P = 1
REAVERAGE CENTRES
P=1?
PRINT OUT THE CLUSTERS
IS NO OF CLUSTERS >2?
YES
NO
stop

FIGURE 26

C(1) Those two groups which cause the least increase in total within groups average distance.

C(2) A point is reallocated to another group if its average distance to the other cluster's members is less than the average distance to its present cluster co-members, i.e. the point i is transferred from cluster k to cluster l if

$$\frac{\sum d_{ik}}{n_k - 1} > \frac{\sum d_{il}}{n_l}$$

In practice C(2) virtually determines the groups found, so this and Beale's method may be simply performed as options in the same program, which can be simplified by using the same C(1) criterion.

It would seem that any iterative relocation method based on a hierarchical method would be assured to perform no worse than the plain hierarchical method on any set of data, but this is not always the case since relocations which occur at the k cluster stage change the input for the k-1 cluster stage. Thus the iterative relocation methods can sometimes (rarely) move into a position from which they cannot escape. This risk is reduced by using block moves, but since all possible block moves cannot be considered, this can still occur.

21. Neighbourhood Method

In our discussion of hierarchical methods we have identified two types of cluster - round and straggly, and with these, two types of method according to which type of cluster they are designed to find. The previous two methods

and those to be discussed in the next sub-section are all to find round groups. The following proposed method is an iterative relocation method which attempts to discover underlying groups of any shape.

The method is based largely on C(2) - the way in which relocations occur. Here a point is relocated if its average distance to the nearest k points of another cluster is less than the average distance to the nearest k points in the cluster to which the point is currently allocated. If a cluster has less than k points then the average distance to all the members of that cluster is used, thus if k is set to a very high value then this procedure acts in the same manner as the group average relocation method. If k is set equal to unity then the method is reduced to single link. Thus k is a strictness factor which allows clusters to increase in 'straggliness' as k is increased.

The C(1) and C(3) criteria currently used with this program are as in Beale's method, but it is anticipated that criteria more consistent with C(2) would produce better results, but would increase computation time, and as the major factor in deciding group membership is C(2) we use the simpler criteria. As in other methods one may begin with a few more groups than one is interested in and use a random allocation to initiate the procedure.

A program to perform the method is given in Appendix 1.

Other Iterative Relocation Methods

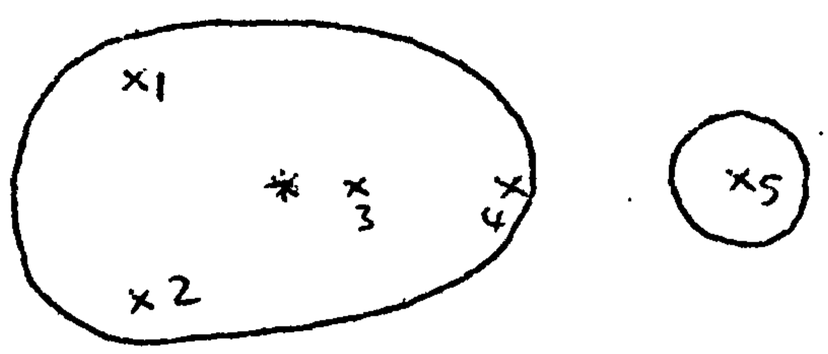
Iterative relocation methods can be divided into two general types - those which proceed hierarchically allowing for relocation at each stage, and the second type which produce a single solution, often by a process where clusters' centres can be created and other centres disappear, a sort of birth/death process. The three methods outlined so far are of the former type.

Probably the earliest use of this second type of methodis in Thorndike (1953) who proposed a method similar to that of group average relocation. He proceeds by doing a 2-group solution, then 3-group, 4-group and so on. The initial group centres are set up separately for each run, the centres for the 2-group run are the two points which are furthest apart, for the 3-group run a third point is added to these which is least near to either of the first two; similarly other starting points are produced. The method might be simply improved by adding a third point which was furthest from the two final clusters in the 2-group solution, for input to the 3-group run, and so on. The method seems sub-optimal to those already discussed.

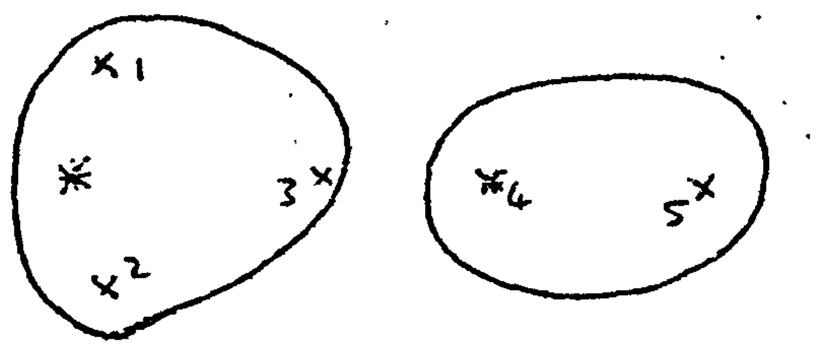
Perhaps the next method to be proposed was that of Jancey (1966) which is apparently (see Lance and Williams 1967) very similar to that of Forgy (1965). With this method points are reallocated to their nearest centre. This is done separately for 2, 3, 4, etc., centres, with, apparently a new random start each time. Whereas Forgy uses

a random split of the population, Jancey uses a random split in space. Jancey points out that the technique is sensitive to the number of individuals in a group and the initial clusters. The method incorporates an unusual attempt to reduce the problem of local optima when the cluster centre is to be moved, it is not moved to the centre of gravity but as far again in the same direction. However this can introduce such instability in the system that one can move away from the global optimum to a local optimum. This can be shown in the following simple example.

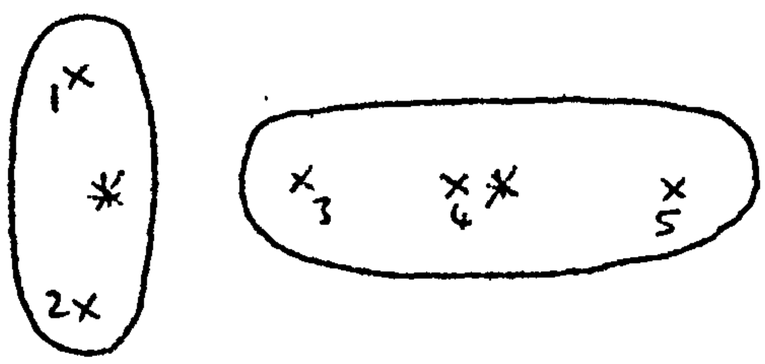
Suppose we have the five points:



with the current groupings and the centres shown. Point 4 must be relocated and if the centres are moved as in the method we arrive at the situation which produces the simplest



clustering of the five points, but here we must reallocate point 3 and our final solution becomes



This might be considered to be sub-optimal to the grouping 1-2-3, 4-5. The method has been used by Wishart and Leach (1970).

A method which uses the birth/death type of process mentioned earlier is that of MacQueen (1967), called the k-means method. The method employs two parameters: C for 'coarsening' and R for 'refinement'. The procedure begins with the first k points being centres, then each point is assigned to the nearest centre and the centre are recomputed. If two centres are within C of each other the groups are merged, having a new common centre. If any point is found to be further than R from the nearest centre then it becomes a new centre.

A similar method has been proposed by Ball and Hall (see Ball 1965, Ball and Hall 1966, 1967, 1968), their method being called ISODATA. This initiates groups by selecting a 'typical' set of points as centres, splits groups if their within-group variance exceeds a threshold θ_E, and combines groups whose centres are closer than a value θ_C.

Both methods require the setting of two parameters, and thus investigation is required of the output for different values, in order to determine the best solution. Both methods are internally inconsistent - the k-means method splitting groups on the basis of points and merging on the basis of centres, and the ISODATA procedure splitting on variance and merging on distance.

Nagy (1969) has extended the above method to allow overlapping clusters, although how this is done is not

explained. A possible way of facilitating this would be to allow a point to be a member of more than one cluster if the distances to two or more centres were similar. Nagy's method, by inspection of his schematic flow-chart, incorporates a third variable parameter - an overlap threshold. This would seem to complicate the use of the method to a high degree.

Rubin's hill-climbing method (Rubin 1967) reallocates a point i from cluster k to cluster l if

$$\frac{1}{n_k-1}\sum_{j\epsilon k-i} S_{ij} < \frac{1}{n_l}\sum_{j\epsilon l} S_{il}$$

The method attempts to maximize the summed 'stability' over all objects. Stability for a point i belonging to cluster k is defined by

$$\frac{t}{n_k-1}\sum_{j\epsilon k-i} S_{ij} - (1-t)\max_{l\neq k}\left\{\frac{1}{n_l}\sum_{p\epsilon l} S_{ip}\right\}$$

where t is a parameter which is a measure of the fineness of the clustering. Each object is given a similarity t with a group with no members, and in this way new clusters can form.

Friedman and Rubin (1967, 1968) attempt to minimize the within-cluster scatter by iterative relocation. As this varies with the number of groups a solution must be found for various numbers of groups. This has been shown by Scott and Symon (1971) to have a tendency to give equal sized groups. Improvements have been suggested by Marriott (1971) to overcome this.

Both of the above methods reallocate points by considering moving each point in turn to all groups and seeing if an improvement in the objective function can be found. They also employ a set of heuristic strategies to try and avoid local optima. These include:

(a) Forcing passes - which gradually eliminate a group by reallocating the member points one at a time to their next best group. The objective function is calculated after each point move, and if any improvement is found this grouping is retained. This procedure is repeated for each group until no further improvement can be found.

(b) Reassignment passes - which assign each object to the group with nearest centre of gravity and then calculates the objective function to see if there is any improvement.

The Friedman and Rubin method has been used by Boggis and Held (1971) in grouping electricity consumers according to their pattern of daily usage. Hodson (1970, 1971) has discussed the ISODATA method and used it on a set of British handaxes with some success.

The main disadvantage of iterative relocation methods of a hierarchic nature is the additional computational time necessary to consider the relocation of each point several times. The difficulty of the birth/death methods, although they are fairly fast, is the presetting of various parameters to appropriate values.

The BCTRY computer package of Tryon and Bailey (Tryon and Bailey 1970, Bailey undated) contains various types of

factor analysis often called cluster methods by the authors. They do however include a method which falls into the more accepted form of cluster analysis. This is their OTYPE program which uses iterative relocation to try and minimize the distance between points and their cluster centres. If two clusters are closer than a user-set parameter they merge, and the process is repeated. (The method has been used by Crovello 1968, 1969, Harman 1970, Myers 1968, Sethi 1971.) This is very similar to Bonner's Method III (1964, 1966) which begins with an arbitrary element and finds all those within a preset distance d of this point. The centroid is calculated and further points may join if they are within d of this point, also points can be discarded. Relocation takes place until stability is reached, then the process is repeated for another point, excluding those already clustered.

Boulton and Wallace (1970) give a birth/death type method which is basically a set of tactics in order to optimize their information measure (Wallace and Boulton 1970). They include reallocating, splitting, merging and adding part of one cluster to another cluster. Boulton and Wallace (1973) propose a hierarchic version, where the two items are merged which lead to the least decrease in 'information'.

1(c) MISCELLANEOUS METHODS

22. Mode Analysis

This method proposed by Wishart (1969a, d) is a density method which seeks 'natural' classes of any shape. The algorithm for the method proceeds agglomeratively but as proposed initially did not produce a dendrogram. The method attempts to identify regions of high density by the following procedure.

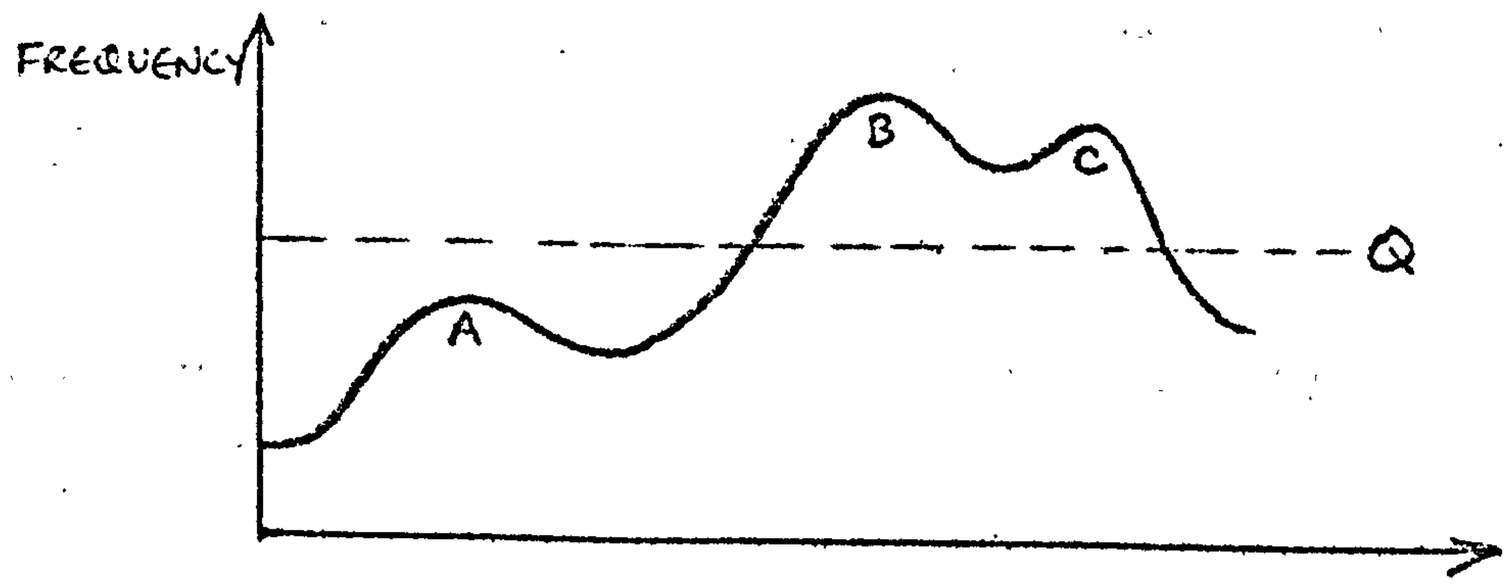
For each point i, an associate density measure k_i is calculated - the original proposal for this measure was the distance to the k^{th} nearest neighbour. As the method proceeds each point becomes 'dense' in descending order of the associate value k_i. Thus the point in the densest region becomes 'dense' first and forms a mode. The point with the next highest k_i becomes dense next and if it lies within k_i of the first dense point then it becomes a member of that mode, otherwise it forms a new mode. As points gradually become dense then either it lies within the current density threshold of one dense point in which case it joins that point, or it lies within this distance of more than one existing dense points, in which case all these modes merge, or it initiates a new mode. Other modes can also merge if the distance between any pair of dense points from different modes becomes less than the density threshold.

If k=1 then the method reduces to the nearest neighbour method. The density measure has recently been improved (Wishart 1971a) by using the average distance to the nearest

2k neighbours instead of the distance to k^{th} nearest neighbour. (Note: in Wishart (1972b) 2k+1 is used.) From Wishart's publications it seems that he restricts to k being integral, but in fact any multiple of a half could be used, but the results are apparently insensitive to k over the recommended range of 3-6.

The output from the method is at two levels - a list of dense points which belong to each mode, which form the nucleii of the clusters, and then a list of associate non-dense points for each model which have been allocated by the nearest neighbour method. Wishart (1969d) advocates restricting the output to the clusters found just before modes are to merge.

A major difficulty of this method is that dense modes which are close together will merge before other less dense and more isolated modes form. In one dimension this can be illustrated:



At density level Q two modes have joined, and mode A has not yet been initiated. In order to eliminate this defect hierarchical mode analysis has been proposed (Wishart 1971a) which does not actually merge clusters as they join, but

outputs the level at which they would have amalgamated. Once all points have become dense, the dendrogram may be drawn.

Another drawback with the method is that it does not always find small clusters of size less than 2k because they tend to be masked by the density measure. The allocation of the non-dense points to modes can also introduce misclassification, and outliers are forced into groups.

A flowchart for the method is given in Figure 27.

23. Condensation Model

This method is based on the analogy of considering the objects under study as points in space, which exert attracting forces on each other as if they had mass and were subjected to the gravitational pulls of the other points. Thus the points gradually condense to one. A similar method has apparently been proposed by Forgy (1963), in an unfortunately unpublished work, but according to Lyerly (1968) and Wishart (1969) the method was found to be not entirely satisfactory. Sneath (1965) also has considered a gravitational method to determine curves from noisy data. A method given by Butler (1969) is similar but is less sophisticated.

The method can be described by its analogy of points in Euclidean space attracting each other by a modified inverse square law, but the method can be used with any dissimilarity measure, or easily modified for use with similarities.

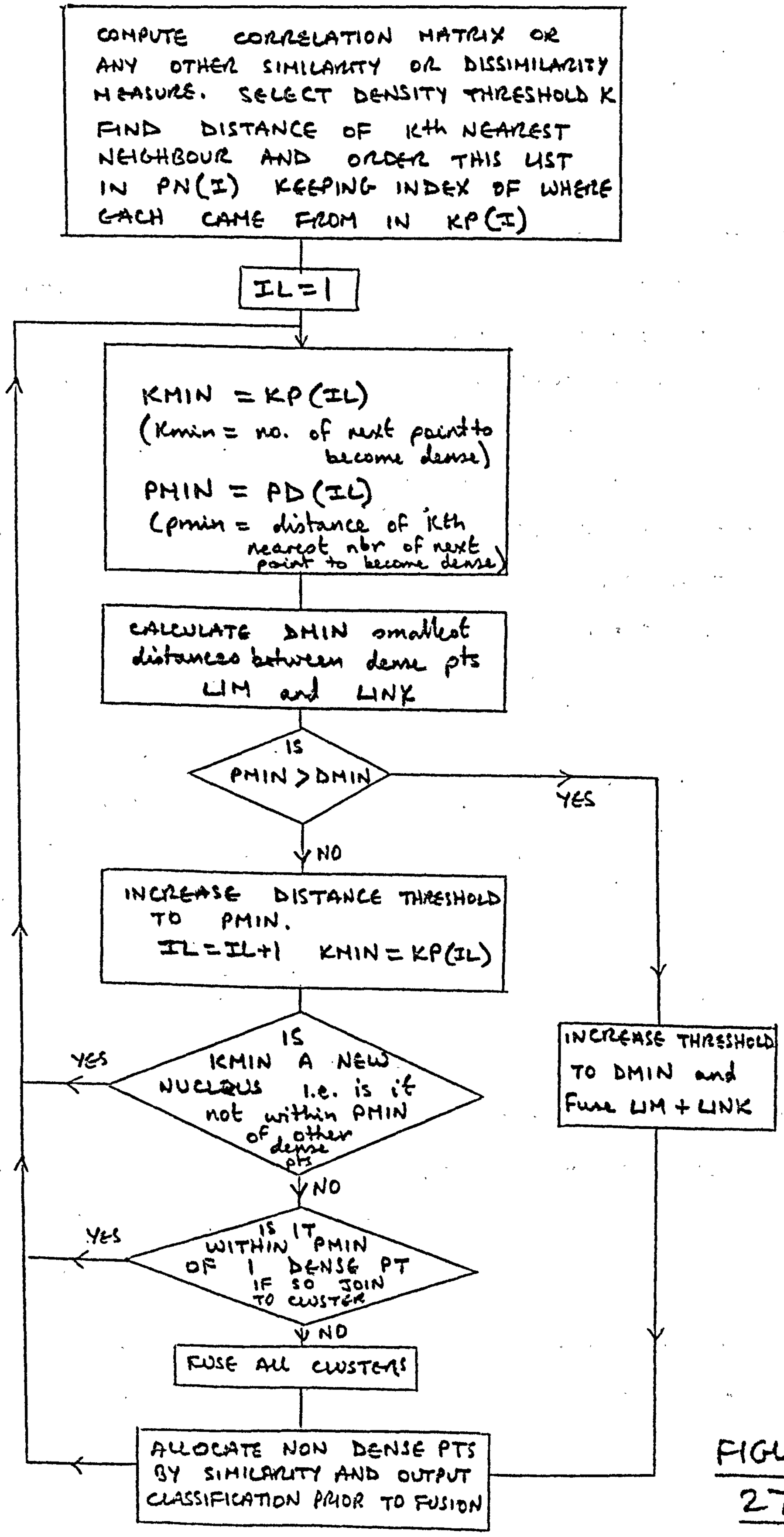

COMPUTE CORRELATION MATRIX OR ANY OTHER SIMILARITY OR DISSIMILARITY MEASURE. SELECT DENSITY THRESHOLD K
FIND DISTANCE OF KTH NEAREST NEIGHBOUR AND ORDER THIS LIST IN PN(I) KEEPING INDEX OF WHERE EACH CAME FROM IN KP(I)
IL=1
KMIN = KP(IL)
(Kmin = no. of next point to become dense)
PMIN = PD(IL)
(pmin = distance of Kth nearest nbr of next point to become dense)
CALCULATE DMIN smallest distances between dense pts LIM and LINK
IS PMIN > DMIN
YES
NO
INCREASE DISTANCE THRESHOLD TO PMIN.
IL=IL+1 KMIN = KP(IL)
INCREASE THRESHOLD TO DMIN and Fuse LIM + LINK
IS KMIN A NEW NUCLEUS i.e. is it not within PMIN of other dense pts
YES
NO
IS IT WITHIN PMIN OF 1 DENSE PT IF SO JOIN TO CLUSTER
Yes
NO
FUSE ALL CLUSTERS
ALLOCATE NON DENSE PTS BY SIMILARITY AND OUTPUT CLASSIFICATION PRIOR TO FUSION

FIGURE 27

The first step is to calculate the distance matrix D_{ij} (i=1,...n; j=1,...n) from the raw data matrix A_{ik} (i=1,...n; k=1,...m) where n is the number of observations in the study and m the number of variables they are measured on. The next step is to scan the distance matrix to see if any pairs of points are so close to one another that they can be considered as a single point (i.e. test to see if any elements of D_{ij} are less than d). If any such pairs exist each pair is considered as a single point (possibly with a change in weighting) at the centre of mass of the two original points, although as the points are very close, for computational simplicity one could use the co-ordinates for one of the original points.

Next the force acting upon each point i (i=1,...n) for which the mass $W_i \neq 0$, is calculated in each direction j (j=1,...m). This is given by the formula:

$$F_{ij} = \frac{1}{\sum_{k} W_k} \sum_{k=1}^{n} W_k \frac{(A_{ij}-A_{kj})}{(D_{ik}+C)^3}$$

The parameter C in the denominator is a positive constant to 'slow down' objects as they become very close to one another, by diminishing the force in this direction, and avoids the problem of very large F_{ij}, and points 'overshooting' their target, a kind of relaxation device.

Each point is moved in each direction in proportion to this force thus the movement M_{ij} of object i in the direction j is given by:

$$M_{ij} = S*F_{ij}$$

All movements take place at the same time, at the end of each iteration. Once a movement has taken place the distance matrix is recomputed and again if any two points are within a small distance d then they are amalgamated.

The procedure continues until all the objects have been amalgamated to one point. From the output, which lists at which iteration fusions occur, a dendrogram can be constructed.

Figure 28 gives the flowchart for the method, and the program is given in Appendix 1.

Three parameters are used in the method, the small distance d between objects to be amalgamated, the step length S which determines the extent of movements, and the slowing parameter C in the force expression. The parameters should be independent of the scaling and number of objects in the data used and so the data is scaled to have unit average interpoint distances.

If d is chosen too large points may be amalgamated which were not moving together, and if d is too small, a very large number of iterations will be necessary before all amalgamations have taken place. Our initial investigations with synthetic data have suggested that the method is fairly insensitive to the choice of d and that a suitable value is $d = .06 \pm .02$.

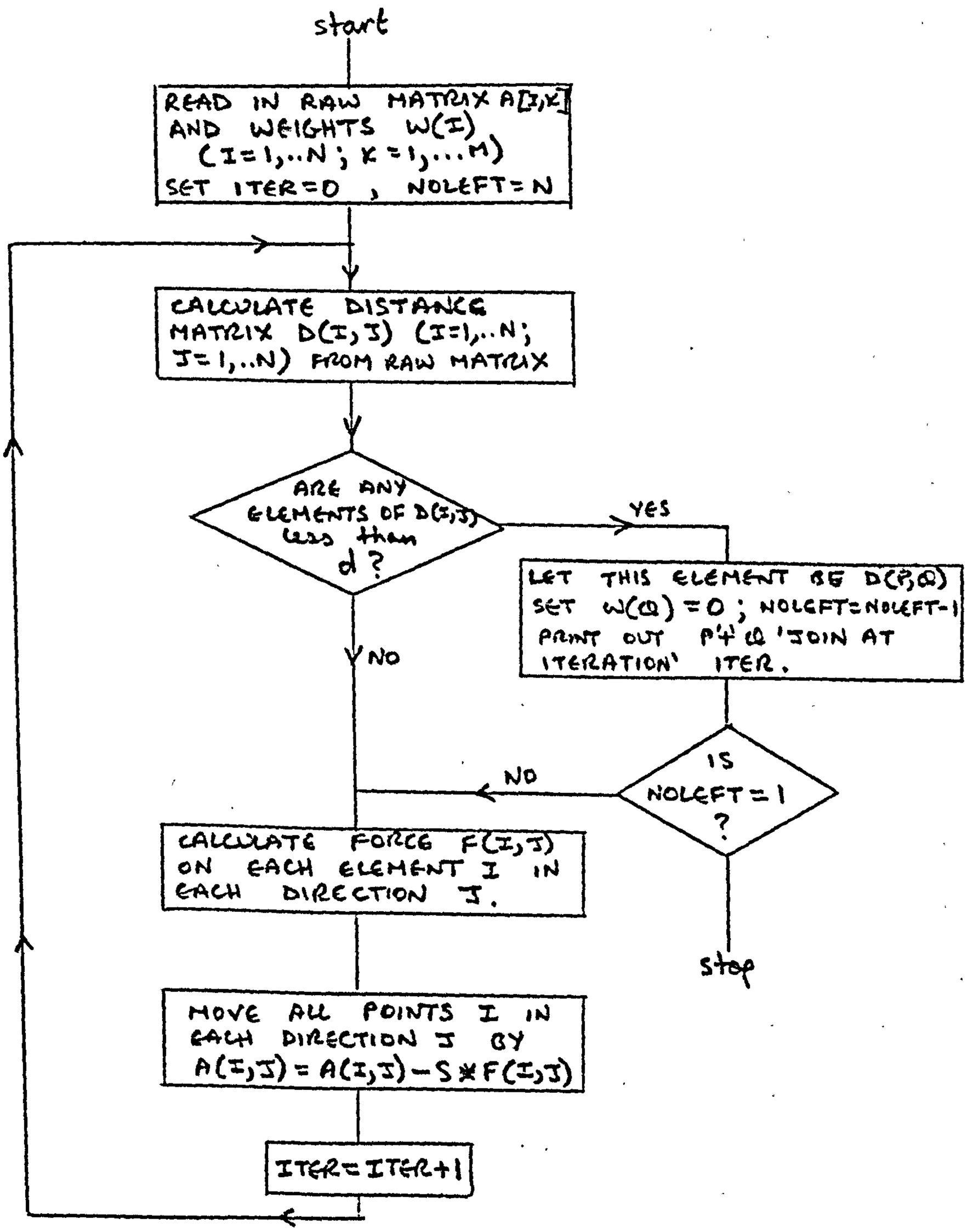

start
READ IN RAW MATRIX A[I,K] AND WEIGHTS W(I) (I=1,..N; K=1,...M)
SET ITER=0 , NOLEFT=N
CALCULATE DISTANCE MATRIX D(I,J) (I=1,..N; J=1,..N) FROM RAW MATRIX
ARE ANY ELEMENTS OF D(I,J) less than d ?
YES
LET THIS ELEMENT BE D(P,Q)
SET W(Q)=0 ; NOLEFT=NOLEFT-1
PRINT OUT P 'AND Q 'JOIN AT ITERATION' ITER.
IS NOLEFT=1 ?
NO
NO
CALCULATE FORCE F(I,J) ON EACH ELEMENT I IN EACH DIRECTION J.
MOVE ALL POINTS I IN EACH DIRECTION J BY A(I,J) = A(I,J) - S * F(I,J)
ITER=ITER+1
stop

FIGURE 28

The choice of the step length S is slightly more sensitive, if S is too large points may overshoot their target, and if S is too small points will only move slowly and hence a large number of iterations will be necessary. We suggest from our investigations that an appropriate value of S is around .005 depending on the level of accuracy required in the dendrogram.

The effect of C is to damp motion in cases where the force would have been very large. Thus if C is chosen too large motion will be very slow as points merge, and if it is chosen too small points may 'overshoot' the point to which they are attracted. Our investigations with C = .3 $\pm$ .1 have shown good results.

The choice of masses depends on the users particular needs. The question of reweighting when points have amalgamated is similar to the choice between the unweighted and weighted average linkage (see Sokal and Sneath 1963). If the masses are added when points amalgamate, then groups are attracted more to large groups than small groups. Another possibility is to give amalgamated points the mass of a single point, and it is this strategy that is recommended.

Other Miscellaneous Methods

Many forms of clustering have been suggested which are difficult to analyse by type. For completeness, in this section we will discuss briefly the known references giving methods not previously mentioned.

The method given by Rohlf and Sokal (Rohlf and Sokal 1967, Sokal and Rohlf 1966) for use in biological classification is to cover a scaled drawing of an animal with a computer punch card, and punch holes in the card to show the shape. These cards (one for each animal) are clustered by using the number of corresponding holes which are punched in each pair of cards as a similarity measure. This is only a crude approximation to the normal taxonomic procedure of taking various measurements on the animals, and poses difficult scaling problems. However one can visualize a more complex procedure that would entail a measurement of the distortion necessary to change the drawing of one animal into another.

Sawrey, Keller and Conger (1960) propose forming groups sequentially by the similarity with chosen points. The first point chosen is that which has the highest summad similarity with all other points. Any points with a similarity with this point above a certain level are joined to this group. Other groups are formed similarly after eliminating members of the first group from the investigation. The distances between group centres are calculated to see if any are close enough to merge, after which any unclassified members may be allocated to their nearest group centre. The method given by Hyvarinen (1962) is very similar. Also related is the method due to Pocock and Wishart (1967) which constructs spheres of set radius on each point, then the densest of these spheres are kept as

groups, and intersecting clusters are merged. The problem with such methods is the setting of the appropriate parameters.

Shepherd and Willmott (1968) propose a k-link procedure in which groups merge on the basis of the similarity between the first k neighbours. The method is an extension of Shepherd's (1966) where cluster centres are formed by single-link, but is not hierarchic - producing a clustering at a set level.

Fortier and Solomon (1966) try a partial enumeration method using random sampling to find a good solution to the maximization of the sum of squared within cluster correlations. Clusters are only formed if r^2 exceeds .5.

K.J. Jones' (1968) 'method of modality' rotates the data matrix to principal components and examines each one for evidence of clustering. This is measured by Kurtosis (the ratio of the 4th moment to the square of the 2nd moment - $\frac{N\Sigma(x-\bar{x})^4}{(\Sigma(x-\bar{x})^2)^2}$).

If this is less than 1.8 (the Kurtosis of a rectangular distribution), then Fisher's one-dimensional method is used to split into two groups. The method only works well with bimodal distributions along axes - this can be illustrated by a simple example, suppose we have 6 points at (0,0) and 3 at each of (-1,0) and (1,0), this gives Kurtosis of 2.0.

Rose (1964) gives a method for use with similarities of 0 or 1 (or by use of cut-off point in the similarity matrix) which form a graph linking points with similarity 1. The method eliminates links from this network by considering the shortest path between each pair of points through the network. Those which are used most times by these paths are eliminated as being probably bridges between clusters. The disadvantage of this method is the large amount of calculation involved in finding all the shortest paths.

Vinod (1969) gives an integer programming solution based on binary variables. This is formalized as:

$$\text{Minimize} \quad \sum\sum x_{iI} c_{iI}$$

(where $x_{iI} = 1$ if $i \in I$ and $x_{iI} = 0$ otherwise, and c_{ij} is the loss of information by putting i in the I^{th} group)

$$\text{s.t.} \quad \sum_j x_{ij} = 1 \quad \text{(for all i)} \quad \text{(each item is in one and only one group)}$$

$$\sum_j Y_j = m \quad \text{(there are m groups)}$$

$$Y_j \geqslant x_{ij} \quad (i=1,\ldots n, \text{ all } j)$$

An extension to Euclidean space is given by Vinod which attempts to minimize the sum of squares. This includes the constraint that if I and J are in the same group then all points nearer to I, than J is to I, should also be in that group. This can easily be shown to be false (see Rao 1971), but is a constraint which has been considered necessary and for which the criterion of error sum of squares has been criticized (see Ling 1971b). Rao (1971) gives integer

program formalizations for several criteria such as minimizing the total with group distance.

Jensen (1969) gives a dynamic programming approach which is basically enumeration after having eliminated unnecessary recalculation. As in Fisher's paper, by using the additive property of Euclidean distance squared, and keeping adequate records, one can build groups agglomeratively. The difficulty with integer and dynamic programming formulations is the high computation time.

McCormick et al (1972) give the 'bond energy algorithm' in which seriation is used to pick out groups as squares on the diagonal of the matrix. The procedure is also suggested for use on the raw matrix. Hartigan (1972) also uses the raw matrix to find groups of similar values to examine relations between sets of objects and sets of variables. The matrix is split into two halves either by row or column on the basis of the minimum sum of squares. This splitting continues on the two halves, producing a division of the raw matrix into varying sized rectangles. This can obviously only be applied where variables are in some sense comparable, and also the computation involved with large matrices becomes infeasible.

Other methods are given by Rohlf (1970) who gives a procedure for any shape groups, but the class of shapes must be specified beforehand, and Bromley (1966) uses inspection on the second order correlation matrix. Methods used in signal detection and pattern recognition are given by

Ball (1965). Hand sorting has been advocated by M. Kendall (1971) for less than 50 objects, and also suggested by Hollingsworth (1972), who uses a kind of manual iterative relocation.

Another useful approach to clustering is to examine interpoint distances to look for evidence of multiple modes. Inglis and D. Johnson (1970) give histograms of interpoint distances and show how clustered data gives rise to skewed or multi-modal distributions. The same approach is taken by R. Johnson and Wall (1969), but use the plot only as evidence of and not a method for clustering, since the clusters found by using the distance plot tend to overlap. Hills (1969) uses the z transform on correlations r:

$z = \frac{1}{2}\log(\frac{1+r}{1-r})$ these are then plotted in rank order of magnitude. This yields a curve which is compared with the line given by a random sample from a normal distribution. Kruskal (1972) also gives transformations to make clustering more apparent, but his results are disappointing.

An interesting method for cluster emphasis is to fit smooth surfaces to points, this has been attempted using the spline transforms of Boneva et al (1971) and also by trend surface analysis (see Merriam and Sneath 1966, Chorley and Haggett 1968).

2. OVERLAPPING METHODS

The importance of overlapping groups is twofold. Firstly, there are instances where these are specified from the type of analysis required, for example in sociology a certain person may belong to more than one social group, and in information retrieval a particular item of information may be of interest to several different people. Secondly, overlapping groups can be used to identify items whose cluster membership is uncertain.

We have already introduced our own nucleus method which can produce overlapping groups. Factor analysis also normally produces overlapping groups.

Most of the early work on overlapping groups was in information retrieval, at the Cambridge Language Research Unit (see Jones and Jackson 1967) where clusters have been classified into three types - chains, where n points are joined by n-1 links; cliques, where n points have $\frac{1}{2}n(n-1)$ links; and of more interest clumps which are well connected groups but with few connections with other groups. In fact, a lot of the work is called clump theory. The method which has gained prominence is due to Needham (1961, 1965, 1967) (see also Dale and Dale 1965 and Harrison 1968). The method isolates small groups (clumps) by iterative relocation attempting to maximize the bias of the group, defined by

$$\text{bias} = \sum_{j \in I} s_{ij} - \sum_{j \notin I} s_{ij}$$

which must be positive to form a clump. Various strategies are employed to ensure the groups are non-empty. Once a

clump has been found its members are not removed from the investigation, the procedure begins again from a different starting point, and hence groups may overlap.

Bonner (1964, 1966) gives a method which groups by iterative relocation. The method begins with a randomly chosen object and all objects with a certain similarity to this object form a group. From this group the procedure 'hill-climbs' to a better group and when no further improvement can be found a group has been defined. None of these objects may be chosen as centres of further clusters, but they are still able to be members of other clusters, and hence groups overlap. The approach of Harrison (1968) considers the density of the points around each object, to see if the density is greater than that due to chance (using a significance test) to define a cluster exists.

The method of Jardine and Sibson (1968a, b, 1971b) (see also Hodson 1970) called the β_k method allows clusters to overlap by k members. It is based on the single link method, but $\min(n_{i-1},k)$ links (where n_i = the total number of points in both groups) are necessary before groups may join. The main disadvantage of the original algorithm was time, but Cole and Wishart (1970) give a faster method of computation. Jardine and Sibson (1971) also mention the methods C_u which allows overlap of certain distance.

3. MONOTHETIC METHODS

A special type of clustering method is the class of monothetic methods, which have been mainly developed in the ecological field. These methods produce a hierarchical arrangement of objects by dividing groups into two parts at each stage of the process. The methods begin by selecting a variable, and dividing the objects into two groups depending on their score on this variable (these methods are normally based on binary data). The procedure continues by considering each of the sub-groups and dividing into two, on the basis of a further variable, which need not be the same variable for each sub-group. This process continues until groups can no longer be meaningfully sub-divided further. Provision is made in some methods for groups to join up if they are not significantly different. The variable chosen at any one stage is that which divides the sample into the most meaningful sub-groups, and it is in the exact meaning of this definition that methods vary.

The term monothetic was first used by Sneath (1962) to replace the previous use by Beckner (1959) of the word monotypic, since it had other meanings. The more usual, non-monothetic, methods are termed polythetic. Monothetic methods do not strictly come under the term of multivariate analysis since they do not conform to the requirement that all the variables must be considered together.

The pioneer work on monothetic methods was mainly carried out in ecology. An early work by Goodall (1953) was perhaps the first method proposed, and this has since been

updated and improved by Williams, Lance and Lambert (see Williams and Lance 1958, Williams and Lambert 1959), being now called association analysis. The computerized method for association analysis appeared in Williams and Lambert (1960). This method replaced Goodall's method and was one of the only monothetic methods in use in the early 1960's. The method was introduced into allied fields (see Lockhart and Hartman 1963, MacNaughton-Smith 1965) but the main area of application was still ecology, where the idea of 'indicator species' had long been prevalent. As monothetic methods became known in other fields, criticism mounted, especially from non ecologists. The main disadvantage of monothetic methods was that they did not obtain natural groups and that two objects which were identical in all ways except one could be forced into separate groups (see Sneath 1965, Bailey 1967). Williams and Dale (1965) have defended monothetic methods against criticism of misclassification on the grounds that such criticism automatically assumes that a polythetic system is desired. The methods seem, however, to work quite well in ecological examples, and the resultant clusters have the advantage that they have unique characteristics. In recent years Crawford and Wishart (1967, 1968) have introduced a fast and possibly more exact method. Also monothetic methods have become increasingly used in marketing, with a method known as the Automatic Interaction Detector (AID) developed by Sonquist and Morgan (1963) (see Assael 1970, Newman and Staelin 1971).

The method of Goodall (1953) was to pick at each stage the most abundant species (species being the variables, being

either present (1) or absent (0) in particular areas) as the variable to split the group, providing it had 'significant positive correlations' with other species. The measure of 'correlation' used was χ^2. The divisions continued until none of the variables defining each group were significantly negatively correlated and were thus said to be homogenous. Provision was made for the rejoining of classes, provided this would not create negative associations between group members.

The method of association analysis (Williams and Lambert 1959) is similar to Goodall's, but division is made on that species with the largest sum of associations (measured by χ^2) with other species. Inverse association analysis was suggested by the same group of ecologists (Williams and Lambert 1961) which was simply using association analysis on quadrats instead of on species, a further step came with nodal analysis (Lambert and Williams 1962) which applied both association analysis and its inverse to the same set of data, to examine connections between species and quadrats. Association analysis, however, gives rise to misclassifications according to Field (1969).

The method of Lockhart and Hartman (1963) was similar to the more usual methods of numerical taxonomy. The procedure was to select the two most similar objects, and then, ignoring variables on which these differed, selected the item most similar to these two, and so on until all objects had joined the groups, or until no property remained common to all. By beginning with different nucleus pairs, and consideration of graphs of number of members in group against

number of properties common to the group, a dendrogram may be built up. The method was tested against a polythetic grouping and was seen to have similar results.

The AID method of Sonquist and Morgan (1963) put monothetic methods on a more mathematical footing. Their method chooses the variable which reduces the unexplained sum of squares by the greatest amount as that which to divide on. Divisions end when no further sub-division reduces the error sum of squares by more than two per cent. This article suggests the main use of the method as in investigation of the importance of the variables. A problem related to this and other monothetic methods, pointed out by Assael (1970) is that if two variables are very close in their discriminating power, the choice of one over the other can give rise to totally different results. The method has been used and discussed by Staelin (1971).

The researches of Crawford and Wishart (1967) were to develop a method which could handle large data blocks efficiently. The method is based on the assumption that species which are frequent and occur in densely populated areas are the best for determining ecological groups. The method calculates a set element potential (SEP),

$$S_i = \frac{1}{x}\sum_i x_{ij} \left(\sum_i x_{ij} \left(\sum_k x_{ik} \right) \right)$$

for each quadrat i which is a measure of the significance of the quadrat. From this, after suitable sealing of S_i so it lies between 0 and 1, we label the new SEP as S_i and compute

$$u_j^2 = \left(\sum_i x_{ij} S_i - \sum_i (x_{ij} \sum_i S_i) \right)^2$$

for each species j. This expression is a measure of the interaction between species and group. The division is made on that species with maximum interaction u_j^2. Crawford and Wishart compared the method with association analysis and found it much quicker on large data sets and no less accurate. In their later paper (1968) they describe an agglomerative method to check for possible misclassifications in their method.

C.3 COMPARISONS OF OTHER RESEARCHERS

Considering the large number of methods which have been suggested there have been surprisingly few comparative studies. Such studies as there have been, fall into three types:

1. The application of methods on real data where success is measured either by the expected or most meaningful result.

2. Applying methods to synthetic data with known properties.

3. Theoretical properties which methods should possess.

We will proceed to consider the work that has been done in each type, a critique of the methods of comparison will be mainly left to the next section.

1. Real data comparisons

Watson et al (1966) in a botany example used Nearest Neighbour and Centroid and could find no real difference in results and in fact suggested that "taxonomic groups.....are not particularly sensitive to variations in the analytical approach".

Moore and Russell (1967) in a pedology study used Nearest Neighbour, Furthest Neighbour, Centroid and Flexible ($\beta = -0.25$) and could not make definite conclusions because of the lack of a satisfactory means of fit. Their dendrograms differed to a large extent at the higher hierarchical levels, and they suggested the use of one or more sorting methods at a low level, and extreme caution at

higher levels. 'T Mannetje (1967) analysed a set of botanical data by the same four methods and was only able to conclude that nearest neighbour was not as good as the others, and also suggested the use of several methods.

Lange, Stenhouse and Offler (1965) compared Association Analysis and Nearest Neighbour on the same set of fossil data and found remarkable agreement. They suggest that both methods can lead to valid but different classifications.

Gower (1967) compared Centroid, Median, Association Analysis and the Edwards and Cavalli-Sforza method verbally, pointing out the impracticality of the latter.

Solomon (1971) compares Weighted Average with factor analysis, the β-coefficient method, and that of Fortier and Solomon. He concludes that "there is a remarkable overlap in the product of all four procedures", something which is not at all clear from the results, no cluster of the 19 objects being found by all four methods. He concludes by suggesting use of the simplest or most economical method.

Crawford et al (1970) use several methods in an ecology example but only as aids to interpret the data and not to compare methods.

Wishart (1971) uses Mode analysis and Beale's method to demonstrate their different approaches, the latter being designed only for round groups.

Lambert and Williams (1966) compare Association Analysis with Centroid (used with an information statistic) on five

sets of ecology data and conclude the second method is "currently both more rigorous and more flexible". Boyce (1968) compared Centroid, Group Average and Weighted Average on zoological data and found Group Average the best, although there was "relatively little difference". Colman (1968) compared the Tanimoto method, Harrison's method and Nearest Neighbour and found Tanimoto's method the least satisfactory.

Bartko et al (1971) compare Rubin's hill-climbing method, Lorr's method, Friedman and Rubin's and Furthest Neighbour on a medical example and found furthest neighbour to be the only method to work satisfactorily. Day and Heeler (1971) also compared Rubin's method with furthest neighbour and were surprised to find the latter method worked better.

Campbell (1970) analysed soil data with principal co-ordinate analysis, Flexible ($\beta = -0.25$), Centroid and Median methods and suggested the use of an ordination method with a cluster method for analyses. Flexible was concluded the best method. Flexible ($\beta = -0.25$) was also found to give the best results in a study by El-Gazzar et al (1968) when it was compared with Centroid (with information measure), although one conclusion was that "the similarity between the alternative hierarchies is remarkable".

Hall (1967) compared Centroid (with information measure) and Group Average against his Heterogeneity Analysis, which fared best.

Sneath (1969) ranks several methods according to his experience and judgment in the following order:

1. Group Average
2. Weighted Average
3. Furthest Neighbour
4. Centroid (with information measure)
5. Nearest Neighbour, Association Analysis

Wishart (1969) gives the following order of success in a geology study:

1. Flexible ($\beta = -0.25$)
2. Ward's Method
3. Mode ($k = 3$)
4. Furthest Neighbour
5. Nearest Neighbour, Median, Group Average, Centroid

Pritchard and Anderson (1971) used six methods to analyse three ecology data sets, with the order of efficiency:

1. Ward's Method
2. Furthest Neighbour
3. Group Average
4. Centroid
5. Nearest Neighbour

Association Analysis is discussed as a special type of method to be used to help interpret the results from other methods and is "not recommended to be used on its own". They also state that "perhaps the most striking conclusion concerns the magnitude and variety of the artefacts produced by small changes in the method".

2. Synthetic data comparison

Lankford (1969) uses Centroid, Ward's and Neely's methods on a constructed two-dimensional data set with very straggly groups. Neely's method (a version of nearest neighbour) not surprisingly did best.

Burr (1970) tried Nearest Neighbour, Furthest Neighbour, Centroid, Group Average, Ward's Method and his own method which minimizes variance at each fusion, on a simplex, which all methods managed to give a symmetrical dendrogram except Centroid which gave chained reversals. This is regarded as failure by Burr but from the argument on page 159, can be considered as no less a success.

Sneath (1966) illustrates the use of Nearest Neighbour, Furthest Neighbour and Group Average on a random set of 20 points in two dimensions. No real conclusions are given except "the clusters produced by the three methods were remarkably constant". The similarity of the results of group average and furthest neighbour were shown to be very similar by use of the cophenetic correlation between their results. The cophenetic correlation (due to Sokal and Rohlf 1962) is the correlation between the elements in the original similarity matrix and the elements of the similarity matrix reconstructed from a dendrogram. Sokal and Rohlf use the measure on real data and find Group Average performs better than Weighted Average. Farris (1969) however shows that the cophenetic correlation (CPCC) measure will always give the highest results with the Group Average Method and concludes

"if it is desired to term 'optimal' those classifications in which most 'similar' OTU's are clustered together, the CPCC should not be employed as an optimality criterion". (Lessig 1972 also uses the CPCC, and Farris 1973 suggests a better measure for dendrogram comparison.)

Rand (1971) has analysed two methods - a method which produces a hierarchy minimizing the sum of all within group distances at each step and one which minimizes the average of the average within group distances. The methods are compared on their results from a synthetic set of data, then with sets including random perturbation and missing data.

Cunningham and Ogilvie (1972) analysed seven methods on 6 sets of synthetic data, and randomly permitted 2 sets which had a large number of ties, three times, and also perturbed 3 sets with an error term. To compare the methods they used two measures of fit between the original and derived distance matrices - Kendall's tau (for the rank correlation of the original dissimilarities d_{ij} and recovered dissimilarities d_{ij}^*) and a stress measure $\frac{\Sigma(d_{ij} - d_{ij}^*)^2}{\Sigma(d_{ij}^2)}$.

The results were:

1. Group Average
2. Weighted Average
3. Furthest Neighbour
4. Median, Nearest Neighbour, Centroid, Ward's

Another extensive study is due to Strauss (1971) who analysed several methods on 7 sets of binary and 7 sets of continuous synthetic data, and later 4 of the methods were

used on another data set. The method of comparison was a measure of fit related to the number of elements that had been correctly assigned. The results were:

1. Beale's method
2. Flexible ($\beta = -0.25$)
3. A divisive information statistic method of Strauss
4. Ward's method
5. Group average
6. Centroid
7. Median
8. Nearest Neighbour
9. Association Analysis

The results discussed so far can be ranked into a very rough order. This is shown below

	STRAUSS	CUNNINGHAM	SNEATH	WISHART	PRITCHARD	CAMPBELL	BOYCE
Flexible ($\beta = -0.25$)	1	.	.	1	.	1	.
Ward's Method	2	4	.	2	1	.	.
Group Average	3	1	1	4	3	.	1
Weighted Average	.	2	2	.	.	.	2
Furthest Neighbour	.	3	3	3	2	.	.
Centroid	4	4	.	4	4	2	2
Median	5	4	.	4	..	2	.
Nearest Neighbour	6	4	4	4	5	.	.

3. Theoretical comparisons

Some of the simplest examples are that of a particular method failing in a situation where the grouping is obvious. Switzer (1968) shows Friedman and Rubin's method failing with two-dimensional groups. Ling (1971) gives examples where the Edwards and Cavalli-Sforza method fails and also Ward's.

Proctor (1966) gives theoretic grounds why the group average method should be replaced by centroid.

There have been two major works on theoretical criteria, by Fisher and Van Ness (1971) (also Van Ness 1973) and Jardine and Sibson (1968b, 1971b) (also Sibson 1970).

Fisher and Van Ness (1971) give the following properties which 'admissible' clustering methods should possess:

(a) Order independent - if the order in which objects are numbered is changed then the results should be unaltered. This criterion can be violated by methods which include random starting points, and also in some methods when ties exist.

(b) Convex admissibility - the convex hulls of clusters should not intersect. This seems reasonable for the case of round clusters, but there can be instances where clusters curve round others which could not be found by methods obeying this criterion - in fact the opposite of this criterion would be a necessary one.

(c) Connected admissibility - the minimum spanning trees of separate clusters should not intersect. This only has meaning in two dimensions, in which clusters can be visualized better than found mechanically.

(d) Well structured (k-group) - if there exists a clustering so that all within cluster distances are smaller than all between cluster distances then it should be found. This is violated even by some iterative relocation methods which can find local optima.

(e) Well structured (exact-tree) - ultrametric data should be preserved. This is only applicable to hierarchical methods.

(f) Well structured (perfect) - if all within group similarities are s_1 and between group similarities are s_2 $(s_2 > s_1)$ then this grouping should be found.

(g) Point proportion - if we duplicate a point any number of times the clusters should not change. This may be reasonable under certain circumstances, but there may often be reasons why this point should be weighted more - in the centre of a cluster the same set of values can easily be repeated by chance, especially with binary data.

(h) Cluster proportion - if we duplicate a cluster any number of times then the clustering is unaltered. Again this is a matter of user decision and not a reasonable property in all instances.

(i) Monotone - if a monotone transformation is applied to the similarity matrix then the clustering should not be changed. Such criteria are of importance in view of the large number of different similarity measures which exist, but if the measure that one uses has strict properties itself then this criterion need not be upheld.

(j) Cluster omission - if we remove a cluster then the remaining clusters are unchanged. This is the same sort of criterion as (h).

Van Ness (1973) adds two more conditions:

(k) Equal admissibility - a method should always do 'better' than that expected from a random division into clusters.

(l) Repeatable admissibility - some methods have corresponding discriminant analysis algorithms (see Fisher and Van Ness 1973). This condition states that if a method is clustered in a certain way and the corresponding discriminant analysis used then no misclassifications should occur. This requirement, however, introduces the properties of another technique, and any method without this property could fail because of the corresponding discriminant analysis and not the cluster analysis.

Jardine and Sibson (1968b) follow a similar development. They suggest criteria (a), (d) and (e) of Fisher and Van Ness and introduce three more; related particularly to hierarchic methods:

(m) Continuity - 'small' changes in the data should produce 'small' changes in the dendrogram. This would be a reasonable property for a cluster method, but if the data is considered exact, then this is an unnecessary criterion. Williams et al (1971) argue that this property is artificial and "do not see how the situation could arise" - but surely robustness under error would be an advantage. (See page 148 where this property is seen to lead to chaining.)

(n) Scale freedom - the dendrogram should be invariant under scalar multiplication of the similarity measure. (This is a weaker form of criterion (i).)

(o) Minimum distortion - the transformation to a hierarchy should be in some sense optimum - this is a weak property which nearly all methods conform to.

Jardine and Sibson examine the hierarchic methods that can be performed with the algorithm of Lance and Williams and conclude that single link is the only one which satisfies all the criteria - we have shown (page 147) that the continuity property leads to chaining, and so this is not surprising.

Williams et al (1971) in a reply to Jardine and Sibson give three new criteria:

(p) Intense grouping - grouping should be more intense than that implied by the original similarity matrix. This appears to be an attempt to force structure on the data - surely the whole point of this type of analysis is to represent the data in the best way.

(q) Insensitivity to outliers - the grouping should be relatively insensitive to outlying values. They state that single-link fails to uphold this property, but since it is the method which identifies outliers most easily, we fail to see this is true.

(r) Ultrametric data should not always be preserved - this is the opposite of criterion (e) and Williams et al give an example where this would appear reasonable.

The difficulty that arises in producing criteria is that one tends to think of the type of clustering which one is used to, and consider the properties necessary in particular types of study to be necessary for the whole field of clustering. The formulation of criteria should be a

preliminary stage of a particular study in order to determine the type of method to be used, but the generalization of these to other studies is often incorrect. One of the advantages of cluster analysis is its broad and varied approaches, and it should be used more as a technique to fit a problem (like dynamic programming) than a technique to fit problems to (like linear programming).

C.4 CHOICE OF METHODS FOR STUDY

In our investigations we restrict ourselves to the polythetic methods which do not produce overlapping groups. This is for several reasons - these form a distinct class of methods for finding a certain type of cluster, this class is the most used, and it is here where the majority of methods exist.

Of all the hierarchical methods which exist, those which have had the most use are the ones which can be used with the algorithm of Lance and Williams. Some of these are also ones which have been used by other researchers in comparisons. We include all eight methods in our study. Nearest neighbour is included, despite its poor showing in other studies, because of its special properties (on the criteria of Jardine and Sibson, and Fisher and Van Ness nearest neighbour obeys all criteria except convex admissibility). We also include our own extension of flexible method. These methods are normally performed by agglomerative algorithms, and so we wish to include some which are normally divisive in nature - the dissimilarity analysis of MacNaughton-Smith is the only one to have been used to any extent, to this we add the profile clustering method of McQuitty, as one of the few methods of this type. Of all these, only single link has properties which make it suitable for finding clusters of any shape. We know of no other present accepted method with this property - thus we include our Klink and Nucleus methods. Other methods which we have discussed are very similar to some of the ones we have chosen for inclusion.

Of the remaining types, we have discussed them on pages and have pointed out some of their disadvantages - mainly in computation time. The method of reduction to an ultra-metric space as proposed by Roux (1972) and Hartigan (1967) is an interesting approach but neither of their methods to perform this operation are satisfactory (Roux states his own method to be inferior to weighted average and the time of Hartigan's analysis is excessive). The only other approach of note is that of Burr (1970) who minimizes the within-cluster variance agglomeratively. This method was discovered too late for inclusion in our tests, and also we know of no published application.

Iterative relocation methods are less abundant - Beale's method, especially when used from a starting position given by each object in its own group, or that given by Ward's method, seems to have advantages over most other proposed methods. It is also the method which, in this country at least, has been most widely applied. The disadvantage of the K-means method of MacQueen or the similar ISODATA of Ball and Hall is that they need parameters which vary with the application. To Beale's method we add the group average relocation method attributed to Wishart. Neither of these methods are suitable for finding straggly clusters, and neither are any of the methods outlined on pages 188-193 Hence we introduce our own Neighbourhood Method explained on page 196

Of the miscellaneous methods we can eliminate several of the heuristic hand-type methods as having more exact

counterparts in the hierarchical methods which are to be included, also the mathematical programming and partial enumeration types which have high computation time. Mode analysis is one method which has been used successfully in a number of studies, and is one of the few methods in existence for finding straggly groups - this method is included. The disadvantages of other methods have been discussed on pages 201-206. The use of data examination to look for evidence of clusters is interesting, but seems to be of particular use in conjunction with other methods, and not as methods themselves. We include our condensation model as an example of a completely different approach.

Some of the eighteen methods we have chosen for inclusion need parameters to be set for their use. All of these are methods of our own except Mode analysis, for this we use Wishart's recommended range of K = 3,...6 and add to it the values K = 1 and 2, so that the results can be compared more easily with single-link. We have a similar decision with the Klink and Nucleus methods where we need a parameter that measures overlap. In both these cases the parameter will normally be a function of the number of objects under study. We use K = 2, 3, 4 for the Klink method, and K = 1,...5 for the nucleus method. In both these cases if we reduced the smallest parameter value we would obtain the single link method. The neighbourhood method also reduces to nearest neighbour and our parameter is again an integer measure which is related to the 'straggliness' of clusters obtained. We use K = 1,...4.

This leaves us with the condensation and extension of flexible methods, each of which has more than one parameter. The range of appropriate values has been covered in the description of the method. We chose d = 0.04, 0.06, 0.08; s = 0.004, 0.006; c = 0.2, 0.3, 0.4. Less values are chosen for s since it has less impact on the results than other parameters, being more of a parameter affecting the time taken by the method. The extension of flexible method has two parameters α and β - we wish to include weighted average ($\alpha = 0.5$, $\beta = 0$), median ($\alpha = 0.5$, $\beta = -0.25$), and those values given by the flexible method ($2\alpha + \beta = 1$, with $\alpha = 0.5$ to 0.9). We began our investigations using $\alpha = 0.4, 0.5, \ldots 0.9$ and $\beta = -1.2, -1.1, \ldots 0.5$ with a few extra points - the median method values, and more points along the line of Lance and Williams' flexible method. After several trial runs the value of $\alpha = 0.4$ was seen to be far inferior to $\alpha = 0.5$, and so $\alpha = 0.45$ was used instead. At this stage $\alpha = 0.55$ was also included and also more isolated points around the flexible line. This gave 142 pairs of values.

All methods were evaluated on eight data sets which exhibited very marked grouping - all of the methods had reasonable results. The profile clustering method however had high running time and had a tendency to force the data into two sets artificially, especially when groups of differing size were used. This method was therefore dropped from investigation.

Thus seventeen methods were selected for detailed investigation. These are listed in Table 3. The parameters used in the extended flexible method are shown in Table 4. If we include the use of a method with different parameters as different investigations - then we have 186 methods or variations of methods to be evaluated.

1. NEAREST NEIGHBOUR

2. FURTHEST NEIGHBOUR

3. WEIGHTED AVERAGE

4. GROUP AVERAGE

5. CENTROID

6. MEDIAN

7. FLEXIBLE

8. EXTENDED FLEXIBLE (for parameters see Table 2)

9. WARD'S METHOD

10. KLINK METHOD (K = 2, 3, 4)

11. NUCLEUS METHOD (K = 1, 2, 3, 4)

12. DISSIMILARITY METHOD

13. BEALE'S METHOD

14. GROUP AVERAGE RELOCATION

15. NEIGHBOURHOOD METHOD (K = 1, 2, 3, 4, 5)

16. MODE ANALYSIS (K = 1, 2, 3, 4, 5, 6)

17. CONDENSATION MODEL (d = 0.04, 0.06, 0.08;
s = 0.004, 0.006; c = 0.2, 0.3, 0.4)

TABLE 3

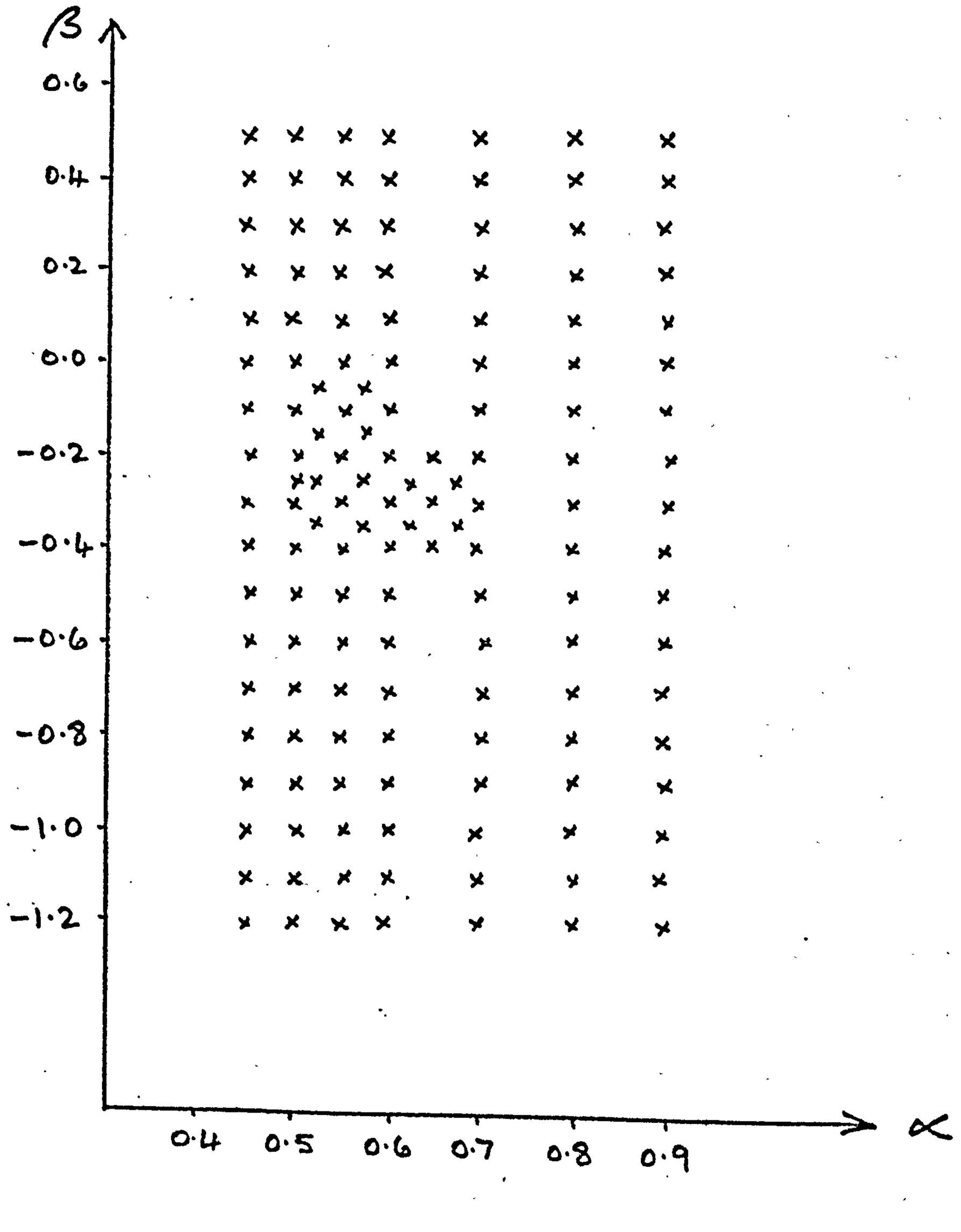


TABLE 4

C.5 COMPARISON OF METHODS

As discussed in Section C.3 several papers have examined cluster methods on their ability to produce meaningful groups in real data sets. However with real data there is no certainty that 'meaningful' groups exist. A typical example is the most analysed set of data in multivariate analysis - Fisher's (1936) iris data. This data set consists of four measurements on fifty plants of each of three sub-species of iris, and has been used as a cluster analysis example by several authors (Friedman and Rubin 1967, Scicon 1971, and Solomon 1971). One set of plants (iris setosa) is clearly separable on two of the variables, but the other two sets seem to have no natural division (see Kendall 1966). Note that in this example any method which has a tendency to find equal sized groups will give a good solution - both Friedman and Rubins, and Beale's (Scicon) methods are of this type and show good divisions into two groups whereas Solomon uses weighted average and finds no such clear grouping.

A similar problem can occur with random data from known distributions - for example a fairly common approach (Wishart 1969, Strauss 1971) is to use two-dimensional sets of random points from normal distributions with different means. Here the groups may overlap and it may be impossible for methods to separate them exactly.

For our tests we hence discard these approaches and use synthetic data which can be easily interpreted visually as falling into a particular number of clusters. Thus we are

immediately restricted to a maximum of three dimensions, and for ease of analysis we use two-dimensional data. We are thus assuming that if a method works well with 2D data then it will perform well with data of other dimensionality. Another possible criticism of this approach is that synthetic data is different from real data, this is in some respects a valid criticism, since in some instances the error in a particular set of real data can give rise to limitations on the choice of method. However the groups we have chosen for study are such that it would be reasonable to expect any method, that was of practical use, to find them.

The clustering of a set of data may vary in six general ways:

1. Number of clusters
2. Number of objects in each cluster
3. The separation of clusters
4. The 'shape' of each cluster
5. Relative cluster areas
6. Relative cluster densities

In each of these variables we could have an infinite range, but we can impose certain practical limits on some of them for the purposes of our investigations.

1. Number of clusters - in real examples the number of clusters found (and looked for) is normally fairly low, so it seems reasonable to concentrate on cases where few clusters exist.

2. In our comparisons we restrict ourselves to 30 points throughout - as a number large enough to form clusters, but low enough for fast computation. The use of a single number of points throughout facilitates easy comparison. Given a particular number of clusters, we vary cluster sizes to include equal sized, large with small, and outliers, as a cross-section of typical cases.

3. We normally attempted to make the separation between some clusters small, in order to make successful clustering more difficult. Clusters were included which 'touched', and thus certain points could be members of different groups in different analyses, but the grouping in each case was evaluated as correct.

4. We made an important distinction between 'round' clusters and other shapes. This was due to the different aims of cluster methods. With 'straggly' groups we concentrated on roughly elliptical shapes.

5. The area covered by each cluster was varied by using mixtures of clusters covering several different sized areas.

6. The overall density of each cluster is defined by the size, shape and number of elements. The density within the cluster was normally either uniform or with fairly central modes.

The number of tests was set at 64 'round' group tests and 32 'straggly' tests. The concentration on round groups was due to the large number of methods which only find groups of this type. With our 186 methods or variations of method outlined in Sections C.2 and C.4 this amounts to 17,856 cluster analyses.

Having outlined our battery of tests, the next important decision is the measurement of similarity, to use in our methods. Since we are concerned with attempting to find groups which we can see by inspection, we must consider the way in which the eye groups points. It seems reasonable to assume that one uses some form of distance strictly monotonic with Euclidean distance. We thus use either Euclidean distance or its square, in all our methods. The use of these measures is further supported by their wide use and growing acceptance. Note that some methods are invariant under monotonic transformations of Euclidean distance, such as nearest and furthest neighbours, and simply represent distortions of scale in the resultant dendrograms. The validity of using groups which have been assessed visually can be judged by the reader, from the tests (which are in general the most difficult ones) shown later in this section.

Scaling problems do not arise, as our axes are weighted correctly visually. Also correlated variables do not occur to any extent and so in these cases Mahalanobis' distance is the same as Euclidean distance.

The most difficult area of our experimental design is deciding whether a method has succeeded in finding a particular grouping or not.

If the data falls into t groups, then we inspect the groupings produced by each method at the t-group stage. If this grouping is the same as the visual one then the method has to some extent at least been successful. The

t+2, t+1 and t-1, t-2,...2 group stages are also investigated to see if their interpretation of the data seems as good as or better than the initial visual groupings, if so then this grouping is also accepted. In cases where the t-group solution was not the same as that anticipated, the range of solutions giving t+2, t+1,...2 groups are considered to see if any are as good as the original visual grouping. If none of the solutions fit the data then the method has failed. This is the first method of evaluation - a simple yes or no to whether a good grouping has been found.

Although the method may have at one stage of its grouping, found a good interpretation of the data, it may be difficult to select this number of groups as the best number of groups. In order to test this, we need to develop a null hypothesis for each method. Thus we must test each method on random data in order to determine significant differences from random results. This gives rise to the problem of what type of random data to use - i.e. from what distribution. For our investigations we used six types of random data, in two dimensions:

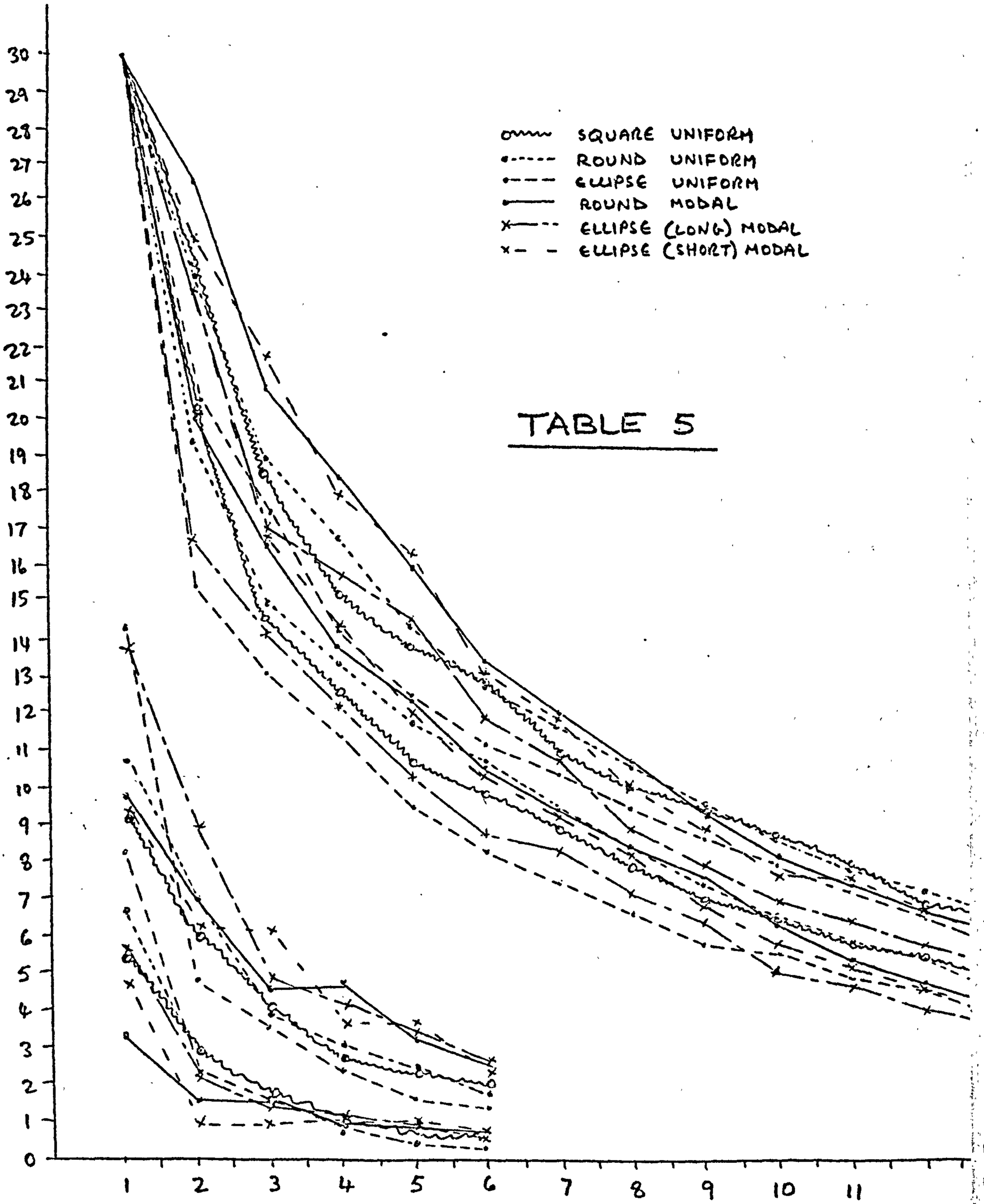
1. Uniform distribution over a unit square.
2. Uniform distribution over a unit circle.
3. Uniform distribution over an ellipse with major axis twice the length of the minor.
4. Normal distribution with unit standard deviation along both axes.

5. Normal distribution with standard deviation 1 along x axis and 2.5 along y axis.

6. Normal distribution with standard deviation 1 along x axis and 1.5 along y axis.

We took ten sets of each, thus giving us 60 more data sets to be analysed by our methods, which combined with our 96 tests gives 29,000 clusterings to be performed.

The following construction of a null hypothesis is that which was followed for most of the cluster methods. From the random data results for each method we form the graph of the objective function for each of the 60 data sets. This is done for each set of 10 in turn and the boundary of the points for each set is drawn. This gives a graph similar to that shown in Table 5, which is part of that produced by the group average relocation method. (Note: that all the lines go through a common point is a property of this particular method and in general does not occur.) The lines form fairly smooth curves indicating that ten random sets for each type of data is probably sufficient. As shown in Table 5 we also experimented with first differences in the objective functions but these gave very wide boundaries and were little use as significance tests. From the complete set of lines we produce, by hand, smoothed boundaries. The boundaries cannot be produced statistically from the sixty points since they are from different distributions. The two boundaries are drawn so as to include all the sixty points for each level, and also to be as 'smooth' as possible. Points may come outside the boundaries in cases where they appear to be separate from the other points - this is subject to the



SQUARE UNIFORM
ROUND UNIFORM
ELLIPSE UNIFORM
ROUND MODAL
ELLIPSE (LONG) MODAL
ELLIPSE (SHORT) MODAL

TABLE 5

condition that a maximum of one of the random sets of data has points outside the boundaries.

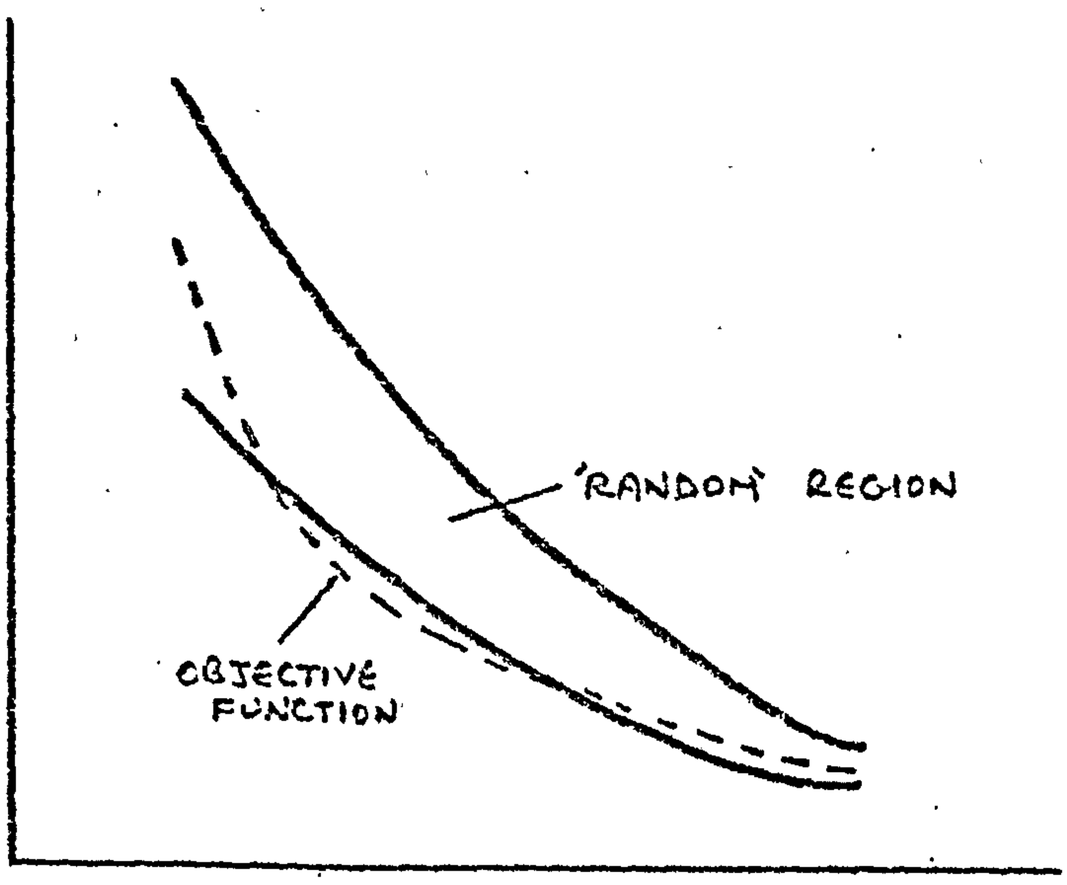
These boundaries form a null hypothesis - if the graph of the objective function for one particular test lies entirely between these lines then the method has failed to distinguish the test data from random data. (Some null hypothesis diagrams are given in Section C.6.)

This forms a fairly strict null hypothesis equivalent to about a 1% level of significance or less. It may be that we can presume that our data varies equally over all dimensions (for example if we normalize) and thus we could eliminate our random data sets 3, 5 and 6. This gives a third evaluation of a method, that is not as exacting as our full null hypothesis. These alternatives are termed the "full null hypothesis" and the "round null hypothesis" in the following text.

Methods produce graphs of objective functions which have some points below the null hypothesis boundary, this indicates clustering. If some points fall above the boundaries then this indicates that the points are more equally spaced than random points. It is possible to have some points above and some points below the boundary - this indicates either clusters of equally spaced points, or equally spaced clusters.

For the comparisons all results (including the random data results) were scaled to unit variance. The random data configurations were all inspected visually to see if any 'obvious clusters' were present. Although there was a surprising amount of apparent structure, none of the configurations were as 'clustered' as the 96 tests.

We now have a technique which tells us if a method has found clusters in the data. We need now to determine how many clusters have been found. If only one point is beneath the null hypothesis boundary then we can obviously pick this out as the number of clusters, so our problem is when several points lie beneath the boundary. If we had two or three markedly separate groups then we expect to obtain objective functions as shown below.



This leads us to the conclusion that the level at which the objective function curve begins to converge with the boundary is the level of clusters given (i.e. where the objective function is 'parallel' to the edge of the 'random' region).

We are thus ready to proceed. We now have three typas of 'success' for a method:

1. Were meaningful groups found at any level in the grouping procedure? This method of measuring success has the advantage that it can be determined exactly. Also if one had opportunity for data checking such as splitting data, repeating experiments or had prior knowledge or expectation, then this level of comparison might be sufficient. Also a change in objective function could improve apparent failures on the other tests of success.

2. Assuming that variables have the same dispersion, are the results significantly different from those on random data? Are the correct number of groups found? In our study most of the configurations had roughly equal variance in each dimension, so this may be considered a valid test.

Note that methods which find straggly groups should be more invariant under axes scaling and hence these null hypotheses should hardly be different from those of the full set of random data.

3. Are the results significantly different from random data? This is the test which involves fewest assumptions, and is hence a very strict test.

With the first method of comparison, methods either fail or succeed. With the null hypothesis methods we can have success, but failure may be of several types:

1. The wrong clusters may be found, but no results be found significant.

2. The wrong clusters may be found, but significant results found on the null hypothesis test.

3. The right clusters may be found, but no results be found significant.

4. The right clusters may be found, but the wrong number of clusters be found significant.

Of these types of failure type 2 is the most dangerous since it gives a misleading result, type 4 also gives wrong results but these may not be quite so dangerous. Types 1 and 3 and the least misleading types of failure because they give

no result rather than wrong results - of these type 3 may be least disastrous since it has been nearest to the correct results.

The procedure outlined above for forming null hypotheses was not applicable to all of the methods under discussion, the procedure in cases such as Mode analysis will be outlined later.

Results

We will consider first the seven methods given by the mathematical expression of Lance and Williams. The table below shows the successes of each method.

	64 ROUND TESTS Level of success			32 STRAGGLY TESTS Level of success		
	1	2	3	1	2	3
NEAREST NEIGHBOUR	53	23	15	27	16	13
FURTHEST NEIGHBOUR	50	36	19	5	0	0
MEDIAN	58	38	30	10	1	1
GROUP AVERAGE	57	47	25	12	1	0
CENTROID	61	55	32	14	2	1
WEIGHTED AVERAGE	58	45	29	11	1	0
WARD'S	61	42	33	12	1	0

The most striking feature of the results is the failure of all the methods, except nearest neighbour, on the straggly tests (with furthest neighbour being particularly unsuccessful). However nearest neighbour performs less well on the round tests, and together with furthest neighbour gives the worst results. The results of the other five

methods are fairly similar, although centroid, on the above figures, dominates all but nearest neighbour and Ward's.

With the exception of nearest neighbour, which performs differently from the other methods we can suggest the following partial order of success:

1. Centroid, Ward's
3. Median, Weighted Average
5. Group Average
6. Furthest Neighbour

It must be pointed out however that the first five methods all give good results and our experiments have not differentiated them to a great extent.

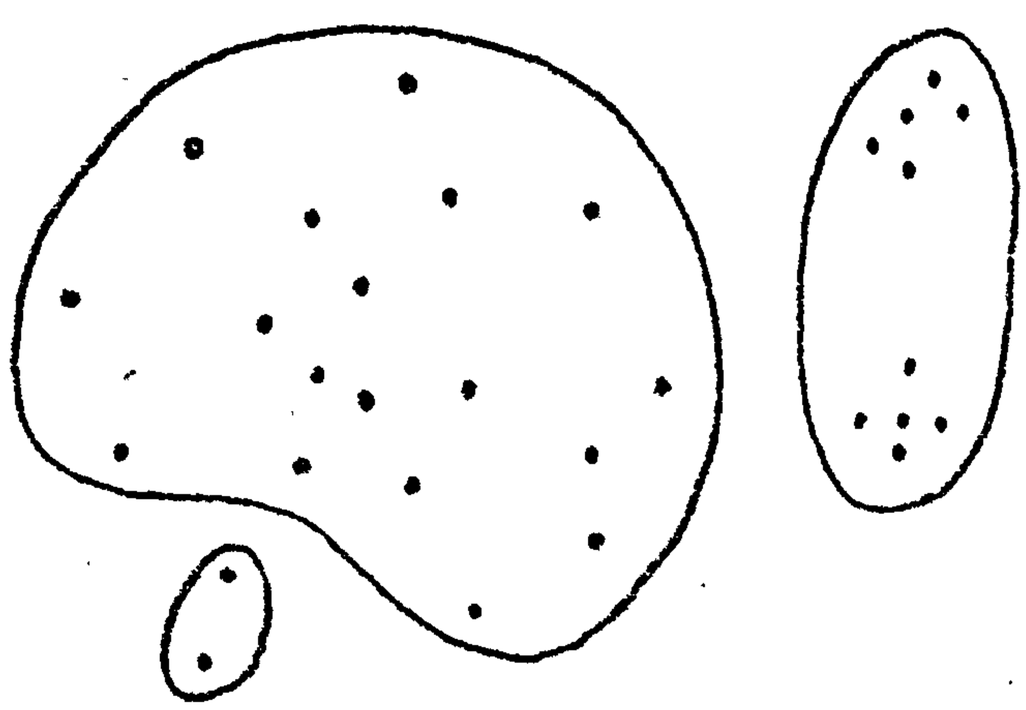
It may be, however, that the different methods do not fail on the same tests and a method which does badly overall may do better than the others with a particular type of group. It will also be useful to examine the failures made by each method in order to determine their properties, with a view to possible method improvements. Here we concentrate on the results of the first success level, and will leave consideration of nearest neighbour until later.

In the round tests all of the six methods succeeded on fifty of the tests, thus all the methods dominated furthest neighbour. All of the methods failed on two of the tests. In the straggly groups all the methods succeeded on four particular tests and failed on sixteen. Here, furthest neighbour was dominated in each test by all but the median

method. The only other dominances are centroid which dominates group average in both round and straggly groups, and Ward's method which dominates weighted average in every test.

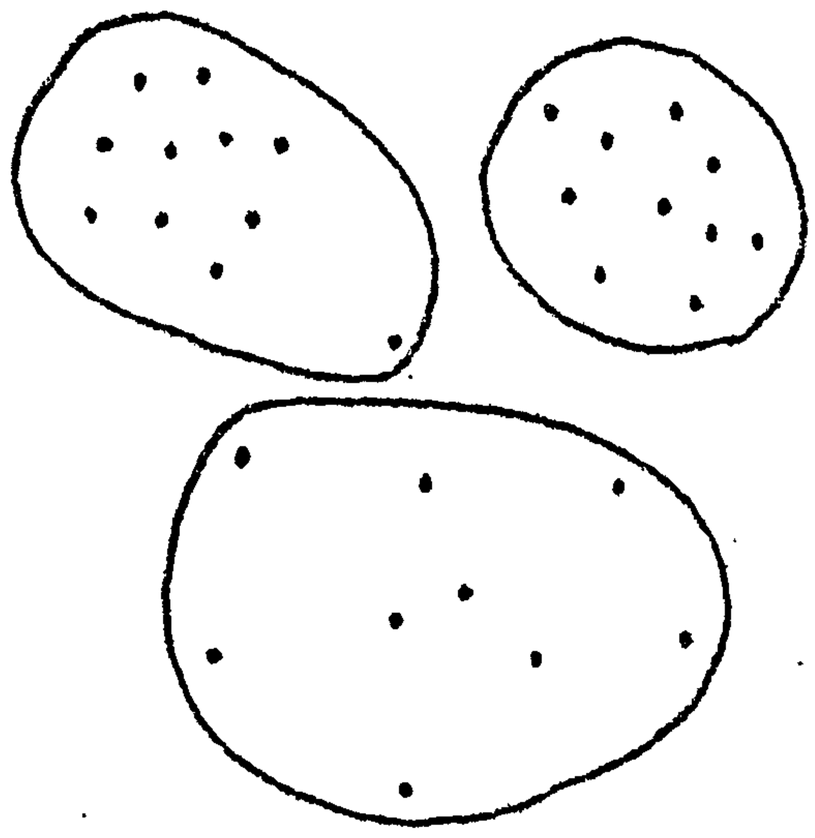
We consider now some of the failures of these six methods in order to examine their properties, firstly on the round group tests.

Centroid - this failed on three of the round tests, two of which all the methods failed on. All three tests involved two dense clusters near to each other and a less dense cluster, covering greater area. For example the three group solution given by the centroid method in test 42 is as follows:



TEST 42
CENTROID

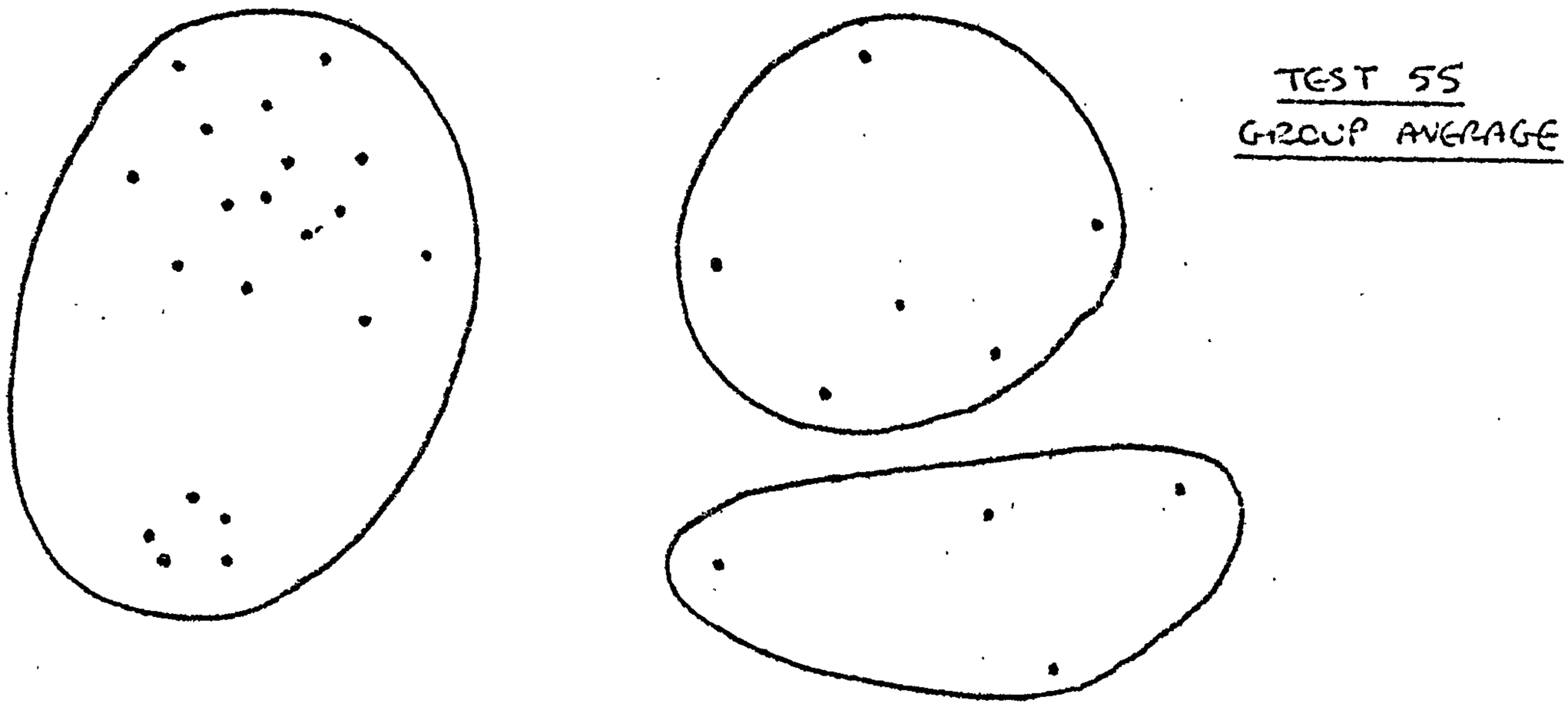
In test 28 the three group solution given by centroid was:



TEST 28
CENTROID

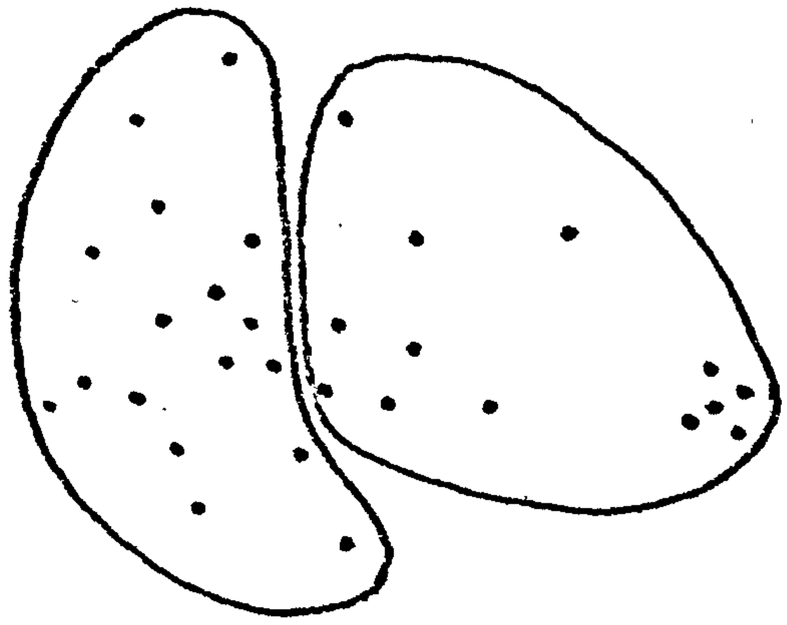
It can be seen that the difficulties of the method are caused by the calculated distance of peripheral members of the dense cluster to its centre being too large. Thus it is not surprising that centroid dominates group average since group average replaces a group by a single point as the grouping proceeds and centroid replaces by an area.

Group Average - this failed the three tests that centroid failed and four others. The failure in test 28 was the same as that shown for centroid above, and that for test 42 very similar to that shown above. All the failures were in cases where a less dense group is present. In test 55 we had the following three cluster result:

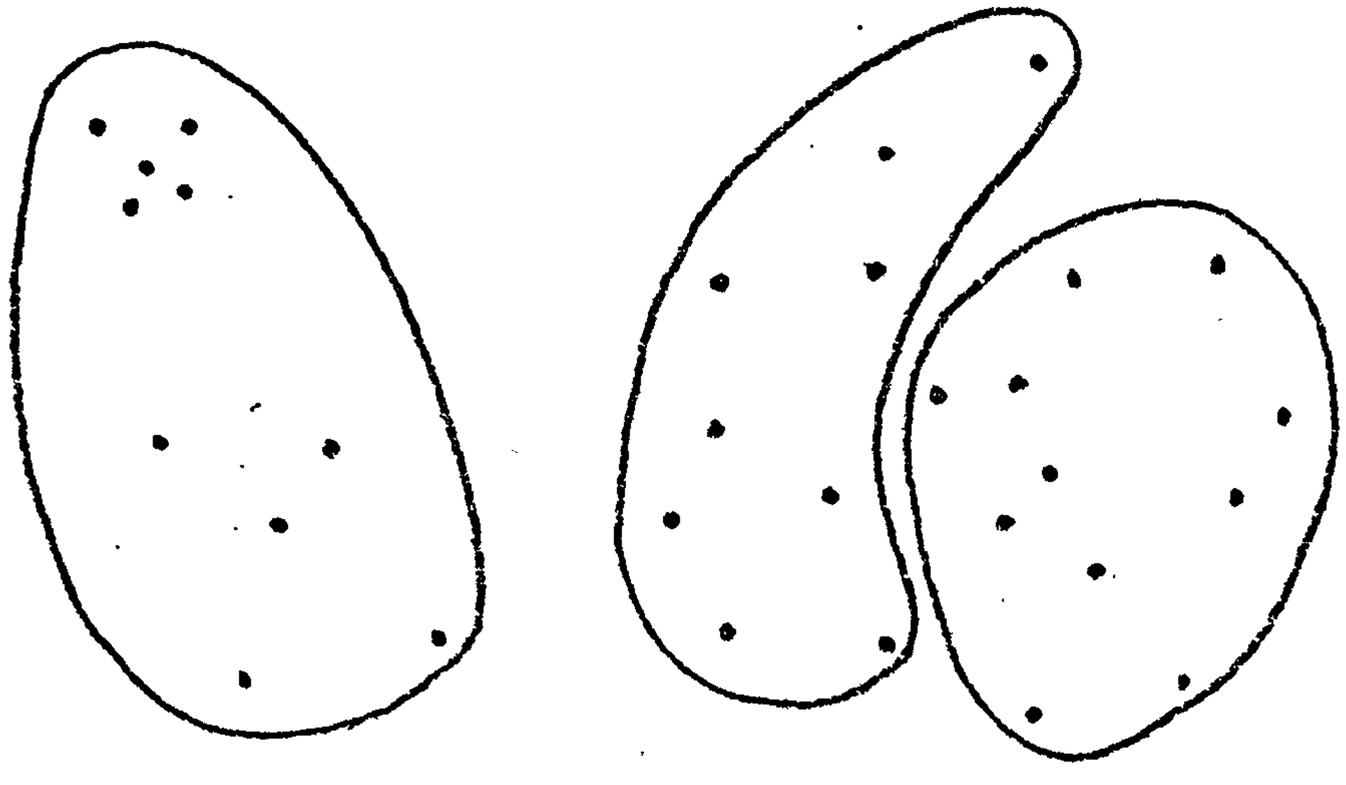


As we have mentioned, group average exhibits more difficulty than centroid with cases where dense and not so dense groups are present.

Weighted Average - this failed 6 tests, all of which involved small and large groups together. We can illustrate this by the results on tests 16 and 40 where the following groupings were found.



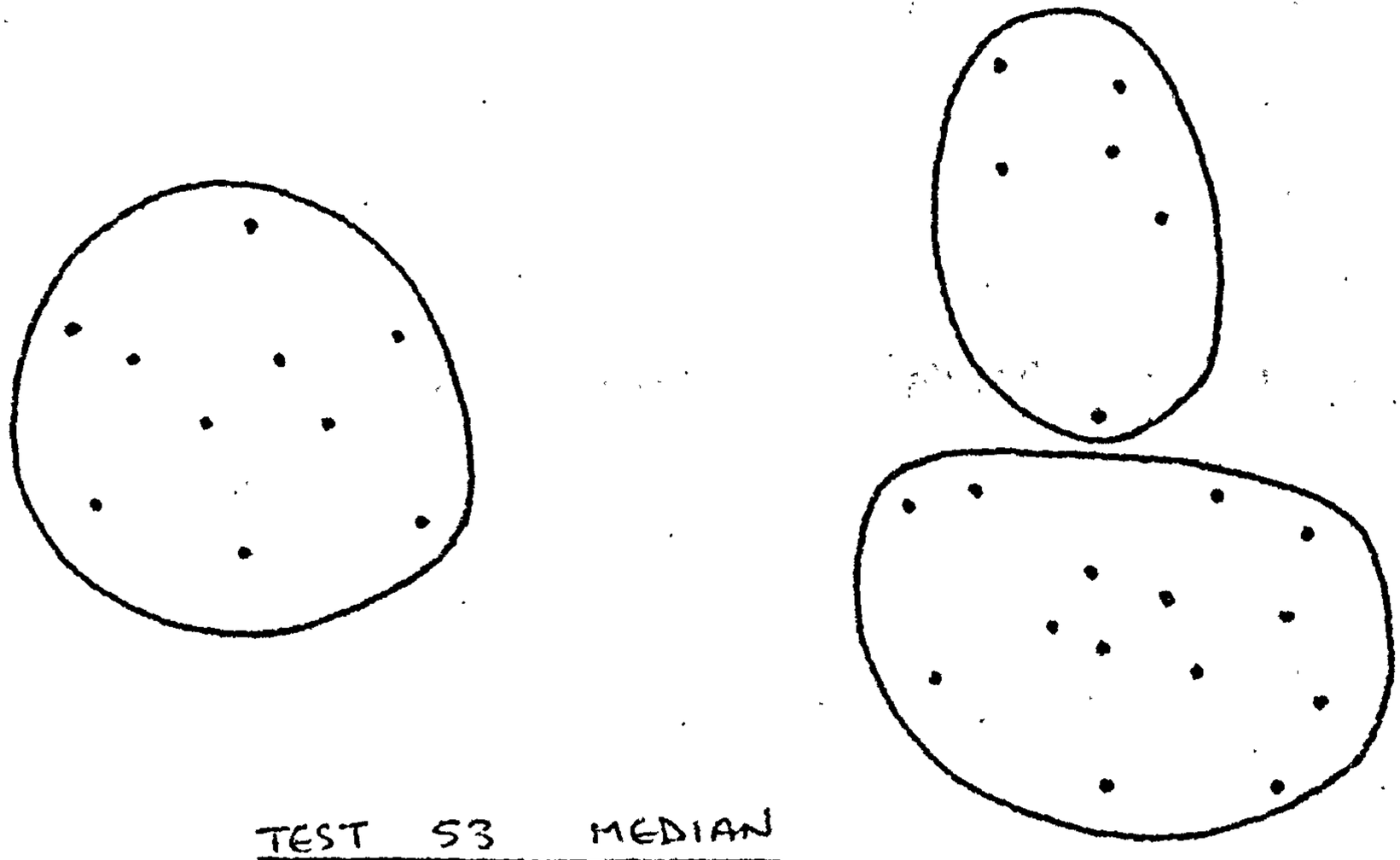
TEST 16
WEIGHTED AVERAGE



TEST 40
WEIGHTED
AVERAGE

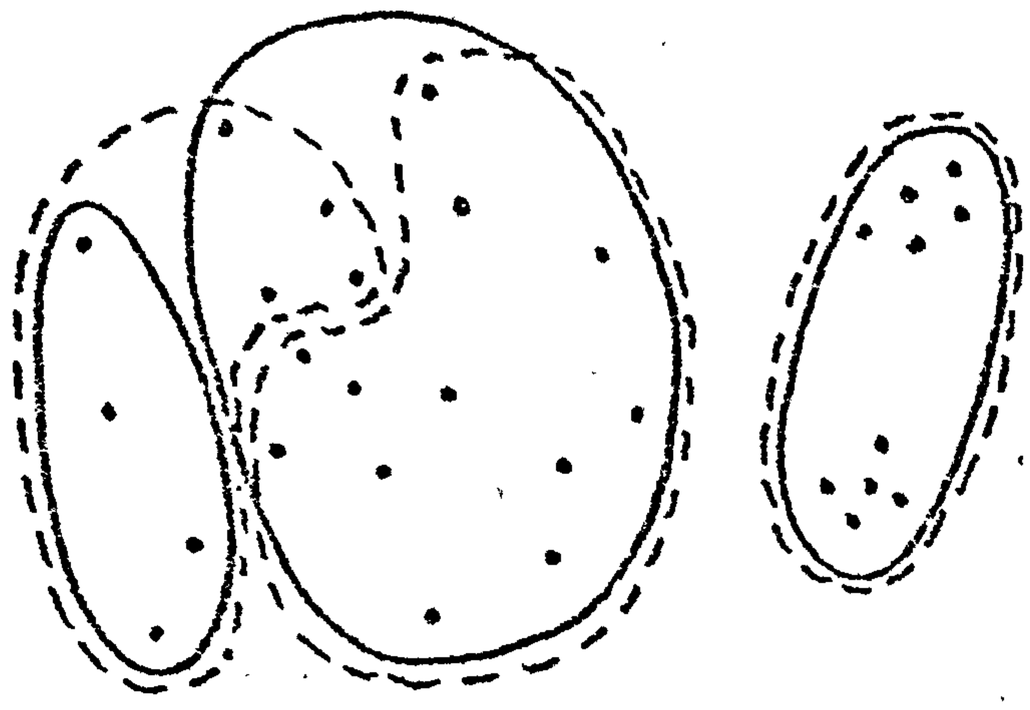
This defect is caused by the lack of weight attached to a cluster when joining a single point - this causes the centre of a small cluster to 'drift' away from its centre of mass. This causes clusters to have a tendency to be similar in their number of points.

Median - the median method also failed six tests, five of them in common with group average, and in three of these giving the same groups as group average. Median however succeeded on test 55 and failed test 53 (see median solution below).



TEST 53 MEDIAN

Another interesting comparison is with test 42 where the group average and median solutions were as follows:

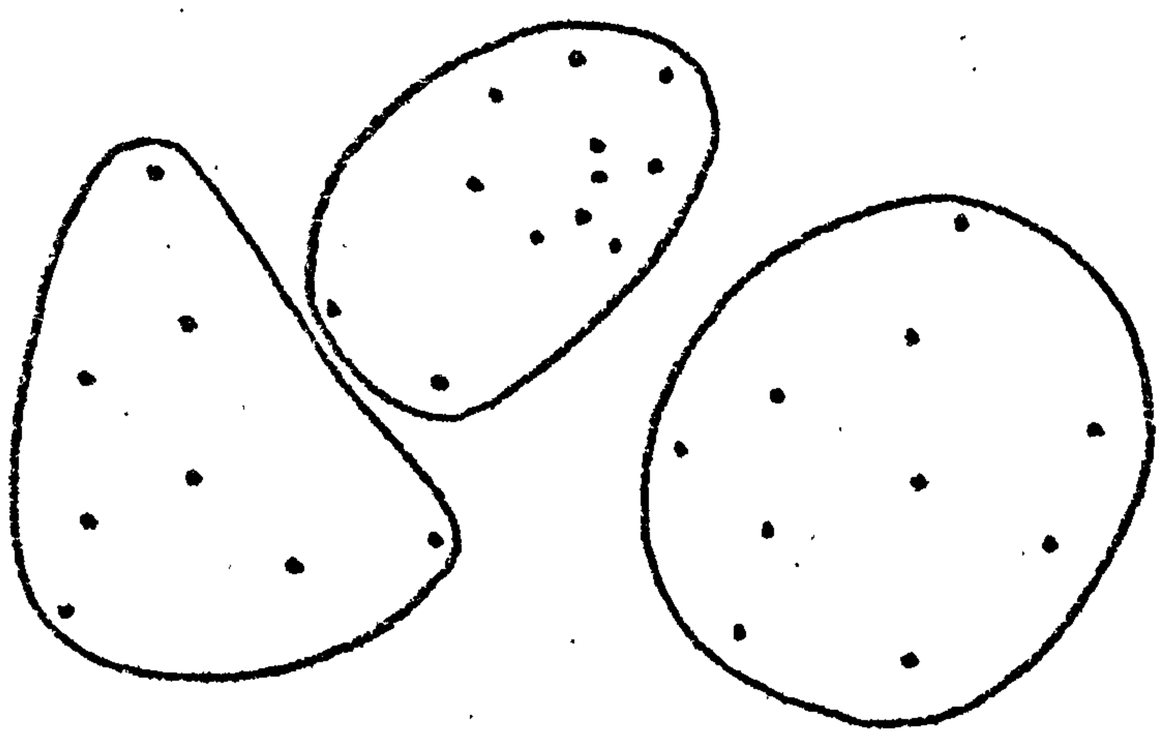


The exact nature of the type of failure of the median method is difficult to specify. It appears to have similar failings to group average. The defect of median is to join

small groups earlier than larger ones, often causing a small group to 'take' a single point which is late to link to a group, from a nearer larger cluster.

Ward's Method - this method also showed a slight tendency to fragment large groups. Three failure were noted - test 40 was split up in the same manner as weighted average, and the two methods that all these six methods failed on, tests 42 and 44. The deficiency of Ward's method then is the same as that of weighted average but to a much less extent.

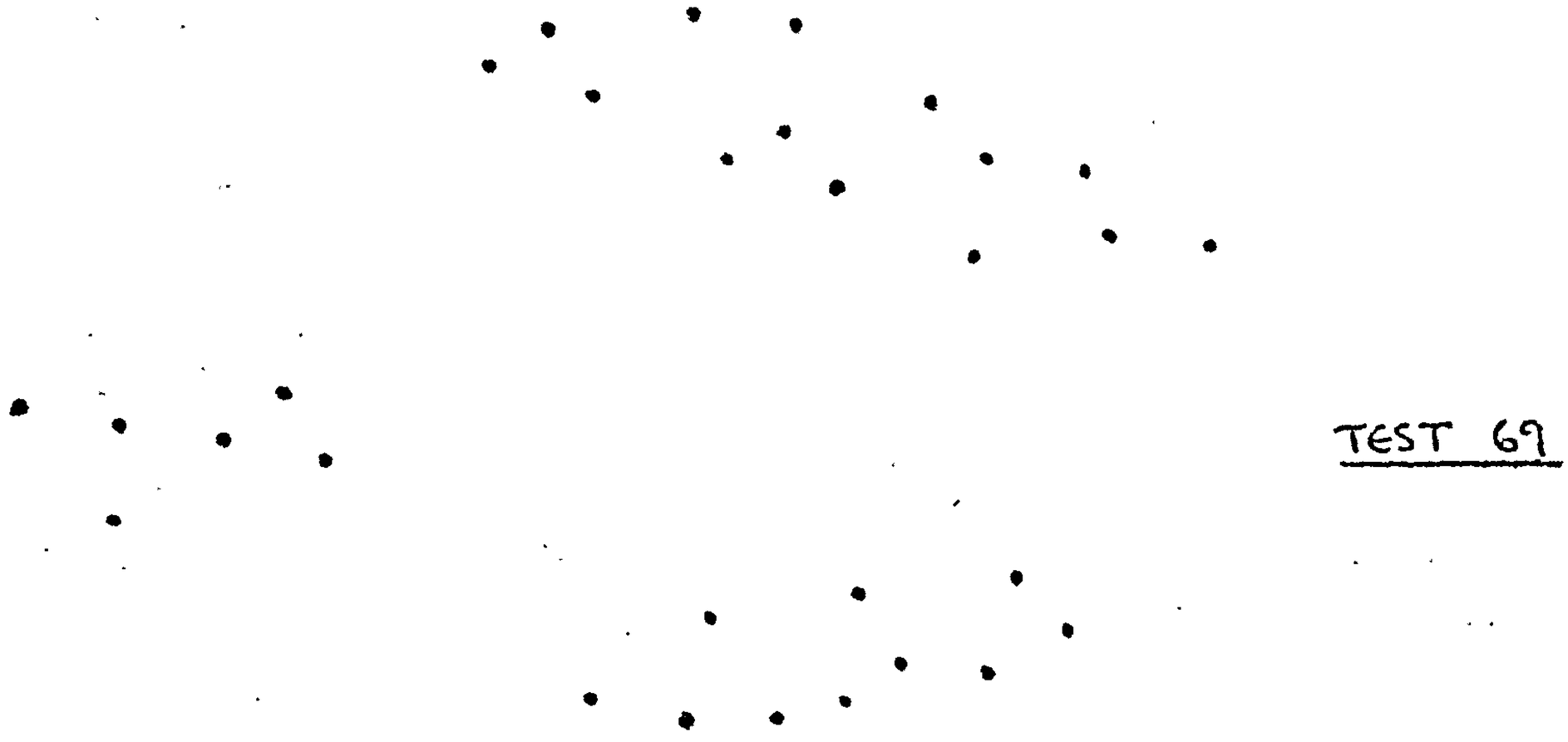
Furthest Neighbour - this was dominated by all the other five methods, note that in all these cases we have groups of differing areas and peripheral points of large area clusters are often clustered with nearby clusters of smaller area. This can easily be illustrated - consider test 23, in which all the other methods succeeded in finding the correct three groups:



TEST 23
FURTHEST
NEIGHBOUR

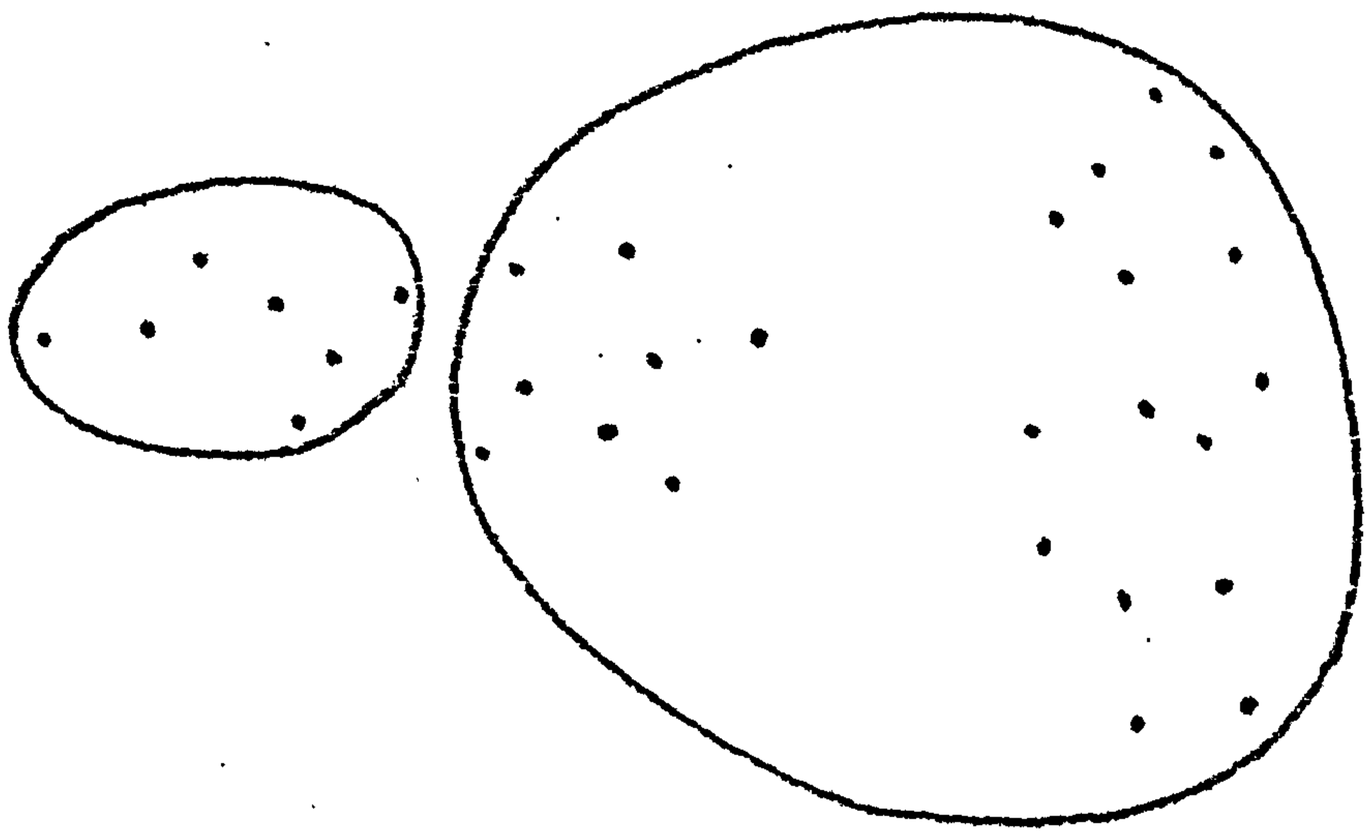
We now consider the performance of the six methods on the straggly groups, but since the methods are not designed for, and found few of the, straggly groups, we will consider

them for why they succeeded, rather than why they failed. Of the four tests which all succeeded on, all exhibited clear gaps between groups, e.g:



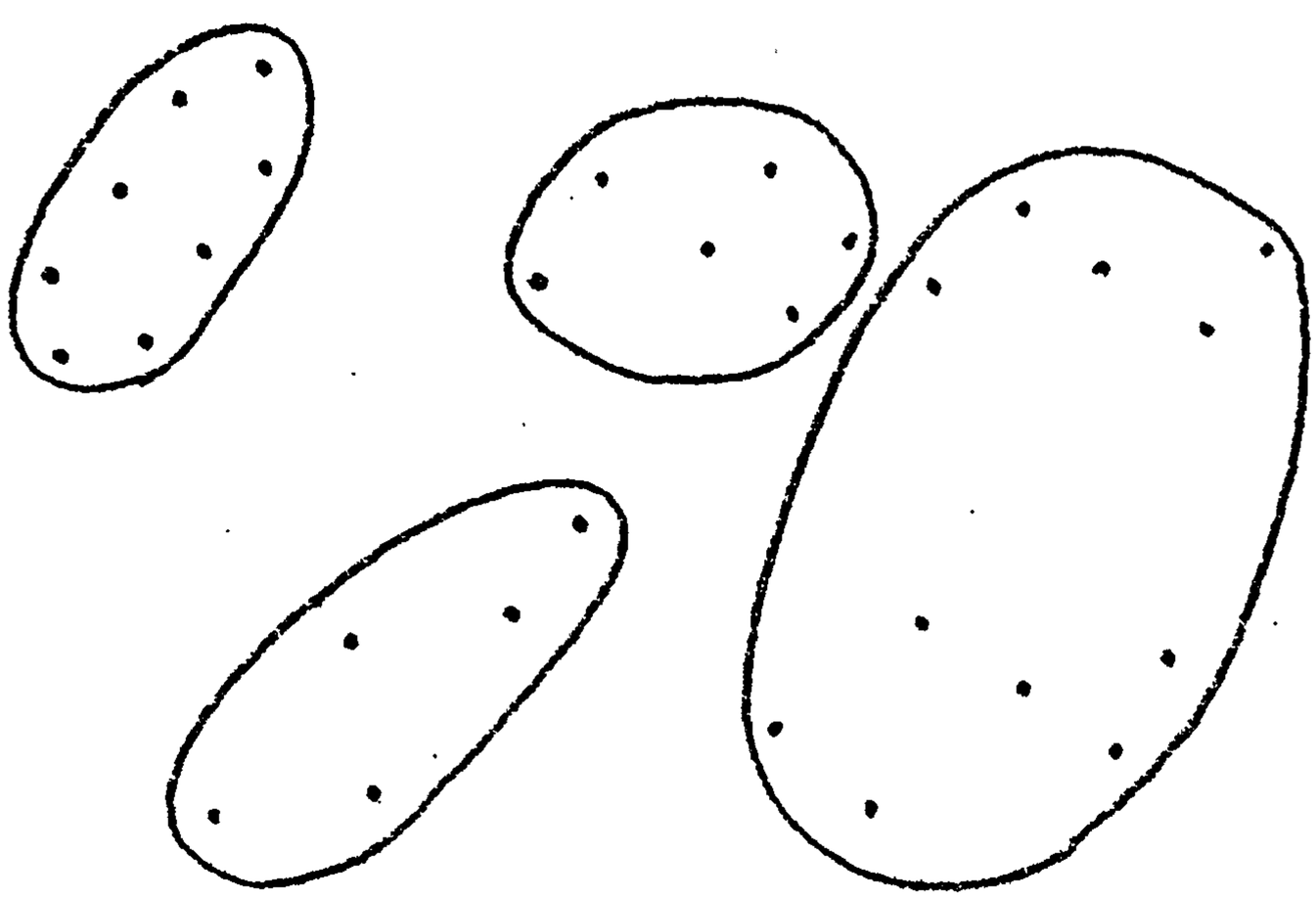
TEST 69

Furthest neighbour only succeeded on these four tests and one other, which only median failed - this is test 67, we show the result:



TEST 67
MEDIAN

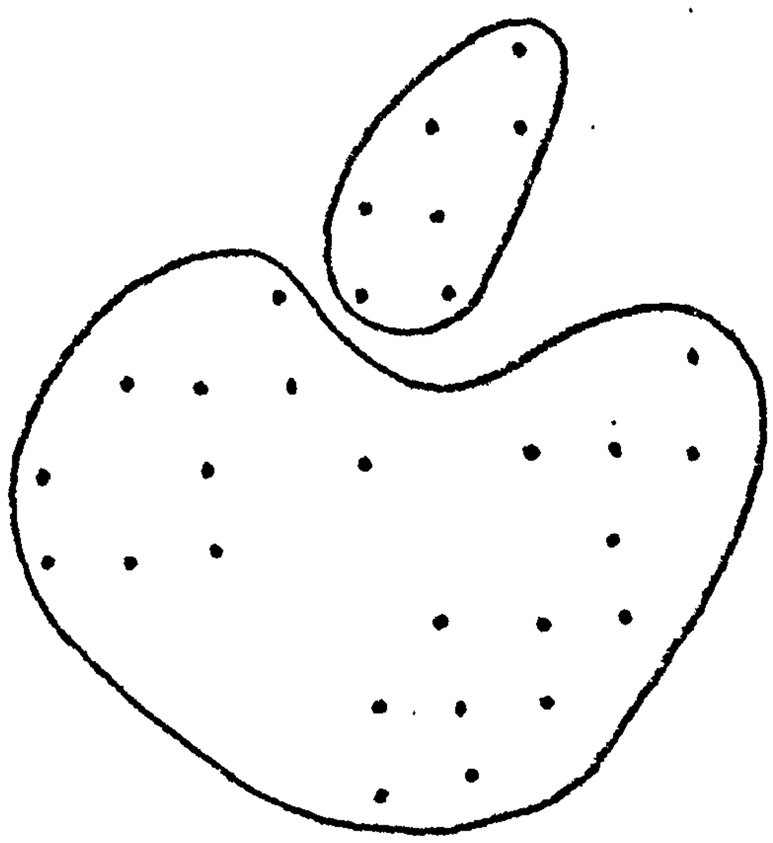
This can be seen as a similar fault to that which occurred with round groups. Three more tests were analysed correctly by the methods, excluding furthest neighbour. One of these is given below - test 70, with the furthest neighbour solution.



TEST 70
FURTHEST
NEIGHBOUR

The success of some of the methods with such groups shows how surprisingly good they are at finding groups. The furthest neighbour solution can be seen to be due to points at the ends of long clusters having low similarity with their own clusters. Furthest neighbour was intended to form 'compact' clusters, but it fails to do this at higher levels in the hierarchy.

Of the other five methods, only median fails test 83 - shown below:



TEST 83
MEDIAN

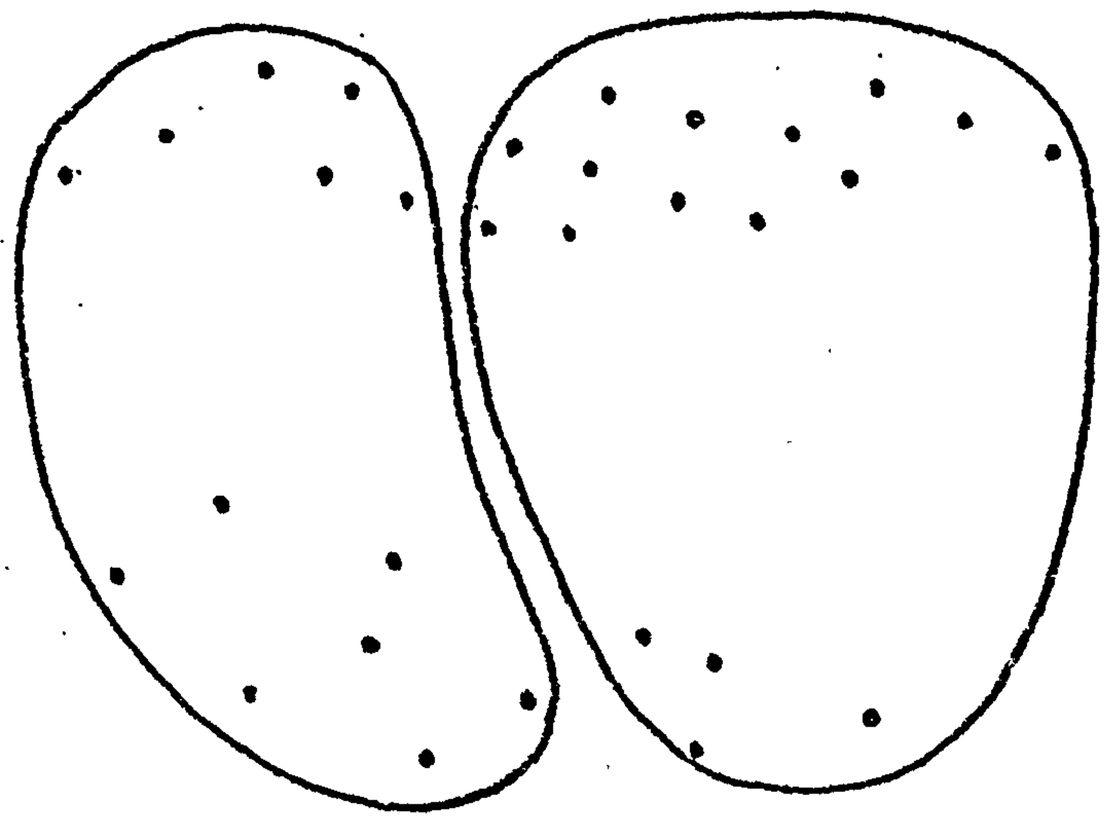
This defect appears to be that median forms 'rounder' groups than the other methods.

Only weighted average of these five methods failed test 85:

TEST 85.
WEIGHTED
AVERAGE

This was caused by three roughly equal groups being formed and the wrong two of them joining. Thus this fault is due to weighted average's tendency to form equal-sized groups.

Both median and weighted average failed in the following case (test 66) and gave the same groups:



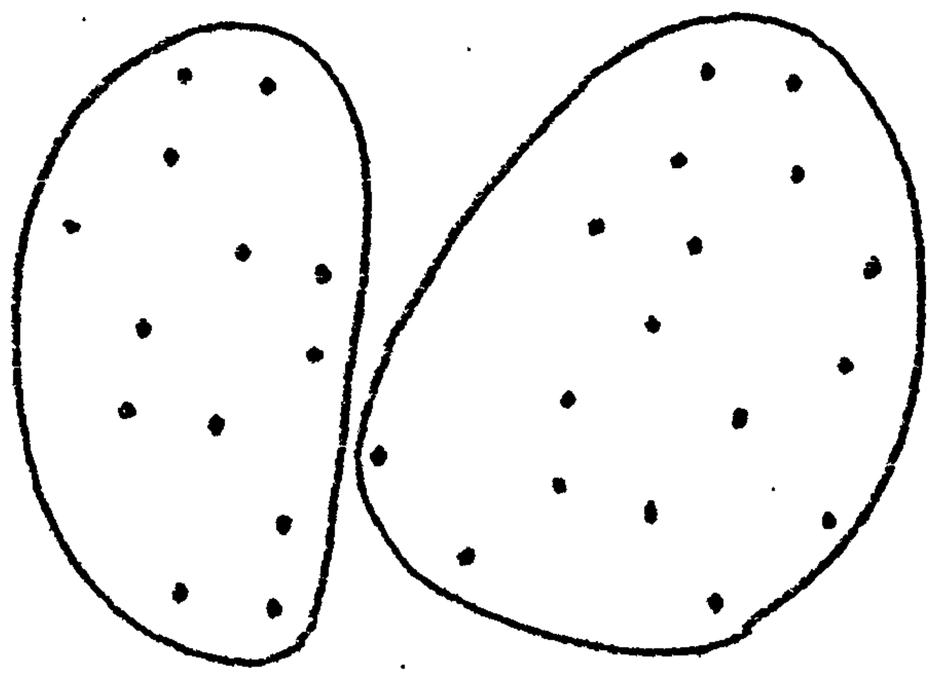
TEST 66

WEIGHTED AV.

MEDIAN

This was caused in both cases from the four cluster solution by the joining of the two left hand clusters.

This was in the weighted average case caused again by the preference of forming equal-sized groups, and the preference of median to join smaller groups. However with test 75 median and weighted average were the only methods which found the correct groupings. The other methods gave the grouping:



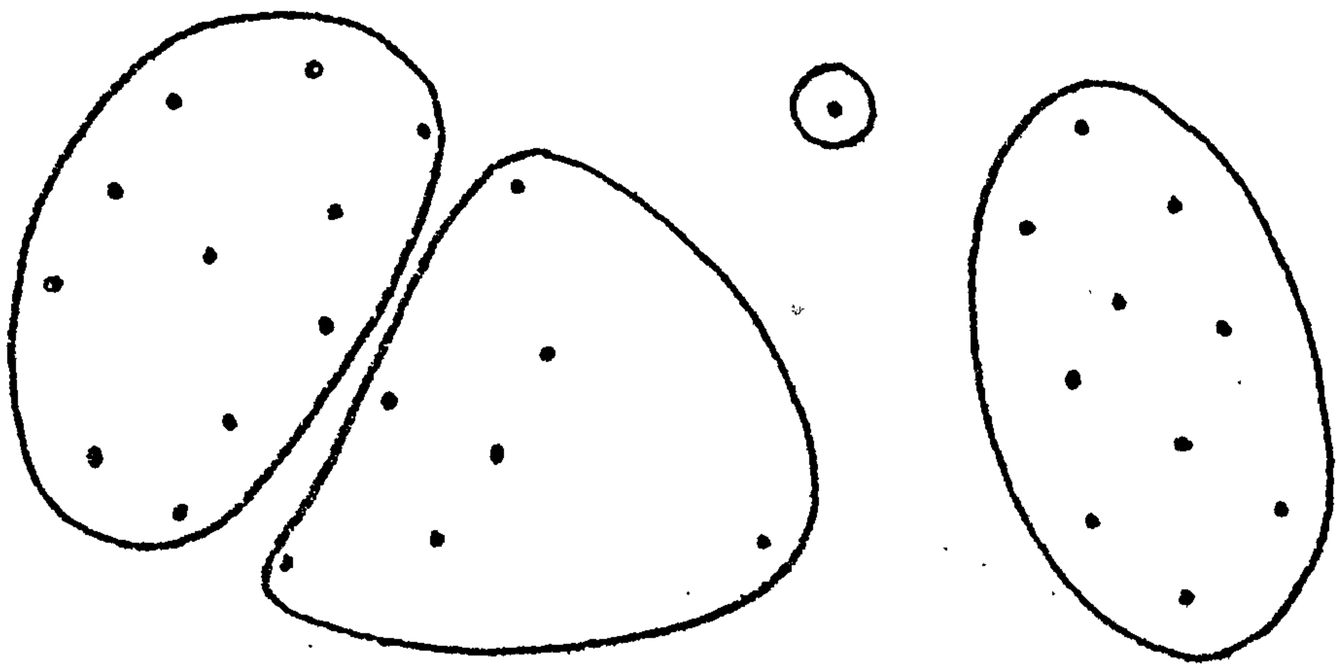
TEST 75

WARD

CENTROID

GROUP AV.

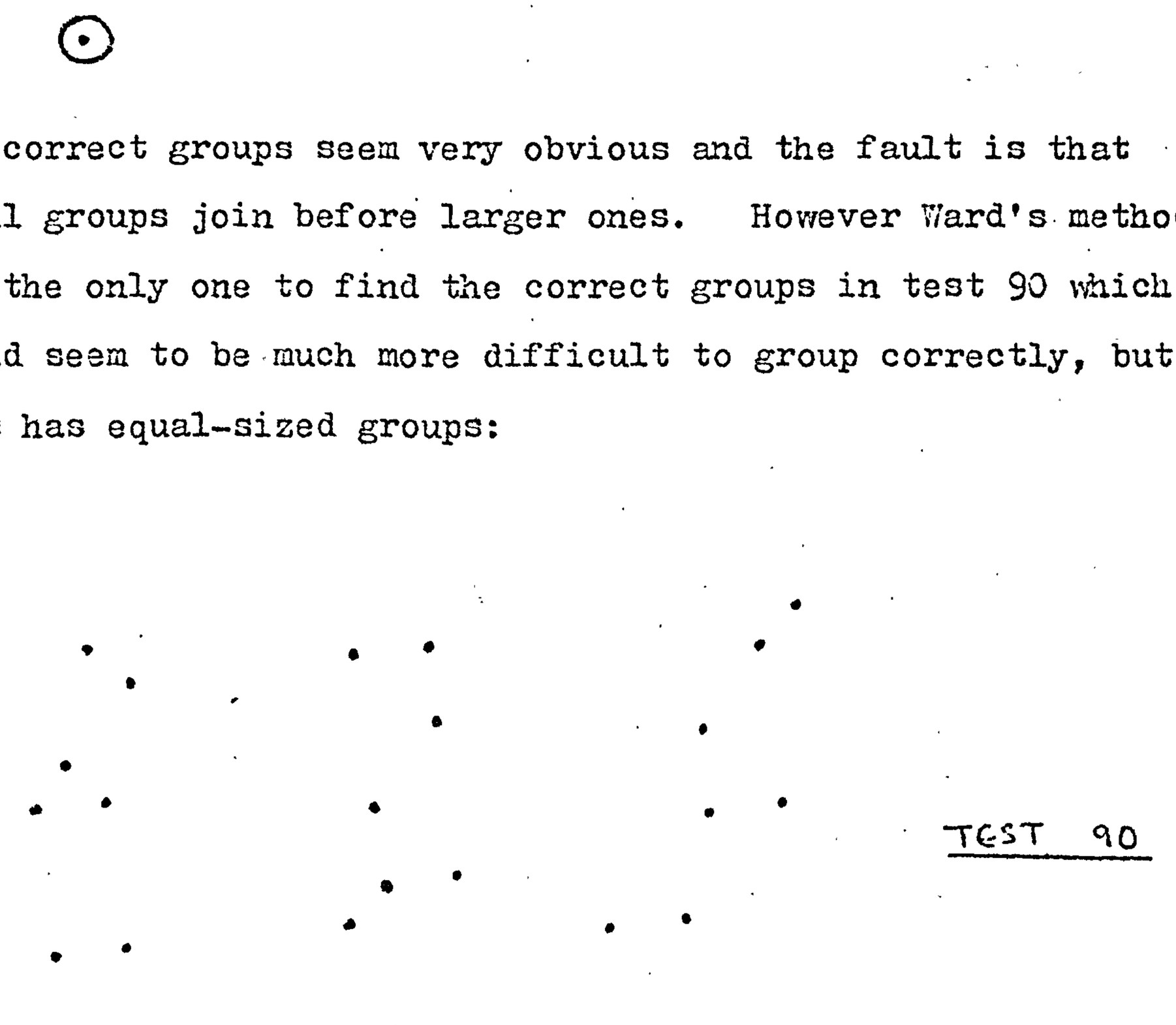
Of the five methods only Ward's failed test 94 - one which included outliers:



TEST 95

WARD

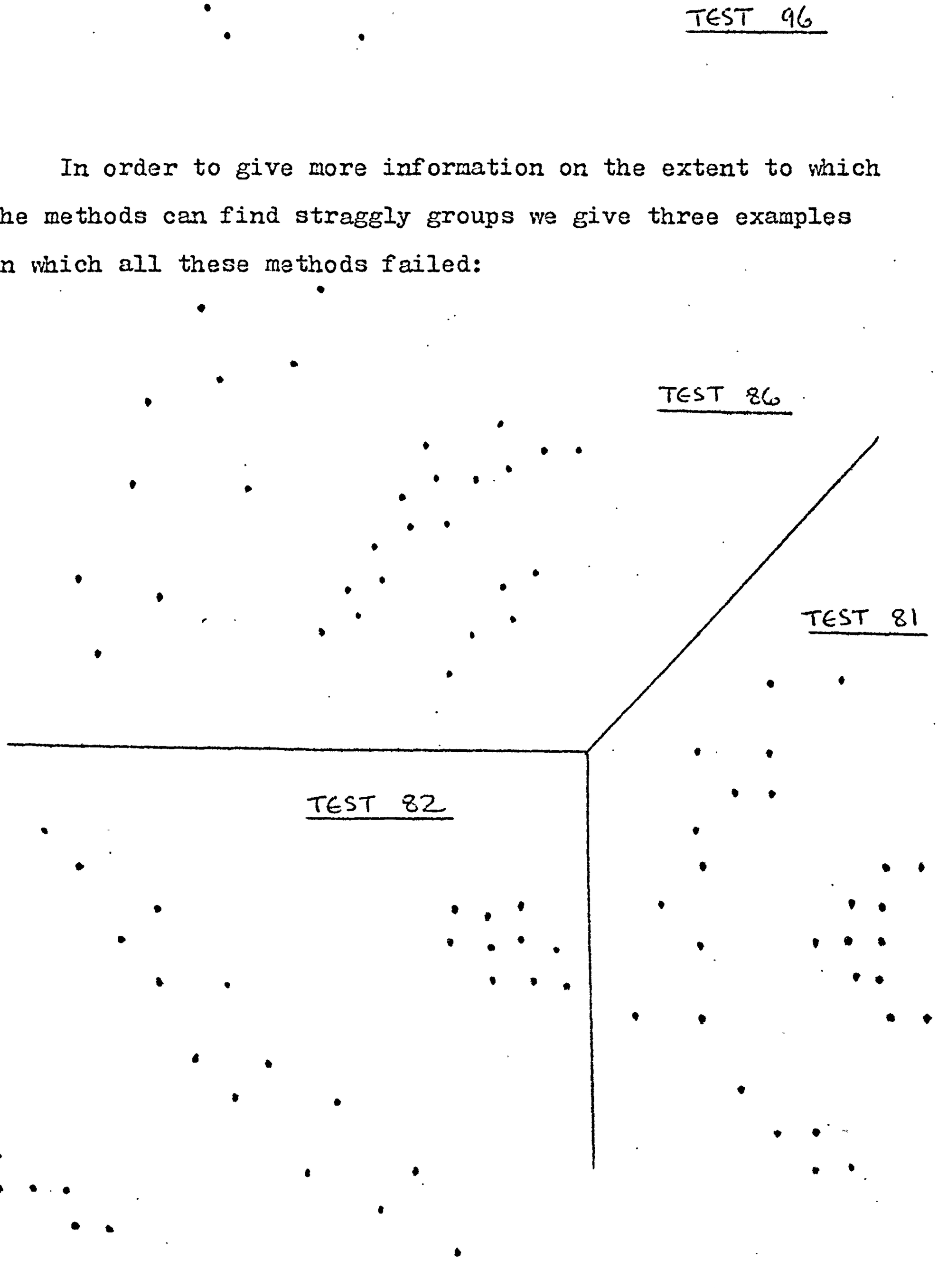
The correct groups seem very obvious and the fault is that small groups join before larger ones. However Ward's method was the only one to find the correct groups in test 90 which would seem to be much more difficult to group correctly, but this has equal-sized groups:



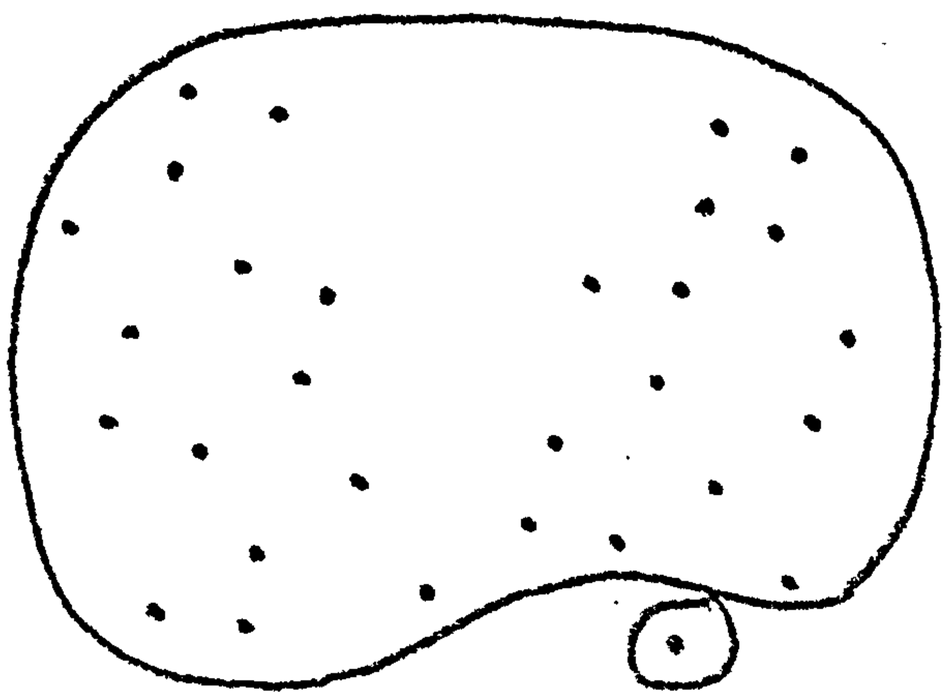
TEST 90

In another test with outliers (test 96) centroid was the only one to succeed:

In order to give more information on the extent to which the methods can find straggly groups we give three examples on which all these methods failed:



Nearest Neighbour - we can now examine the behaviour of nearest neighbour. In the above three examples the top two were successfully completed, but the lowest one was not found, since the two denser groups merge before the third forms. The method is also notorious for its inability to separate groups which 'touch', or have isolated points between them. For example the two group solution in test 75 is as follows:



TEST 75
NEAREST
NEIGHBOUR

The method succeeded on the outlier tests perfectly, and the defect in the above grouping can be seen as over-enthusiastic search for outliers.

We can summarize our findings so far, as follows:

1. Nearest neighbour performs better with straggly groups than the other methods and worse on round groups. The other methods had very similar results (except perhaps furthest neighbour).

2. Median, Group Average, Weighted Average, Centroid and Ward's methods are in general superior to furthest neighbour which is dependent on the maximum width across

the cluster, and hence fails with clusters covering different sized areas or straggly clusters.

3. Centroid performed better than group average which tended to lose points from groups which cover larger areas to nearby groups of smaller area. This is due to the replacing of a group by a single point at the centre.

4. Weighted average has the opposite defect in that the group was often represented by points outside the densest part of the cluster. This has the effect of tending to form clusters of equal numbers of points.

5. The median method had similar results to group average in general.

6. Ward's method dominated weighted average with round groups, having similar defects, but to a lesser extent. Ward's method had difficulty identifying outliers.

7. Centroid had difficulty when dense groups and less dense groups were present, but together with Ward's method had the highest success rate.

8. The extent of 'straggliness', which the methods (except furthest neighbour) could cope with, was larger than one would have expected from the literature.

9. Nearest neighbour was seen to fail where denser groups are close and also where aberrant points existed between clusters. The method was seen to identify outliers easily.

The results of each method on each test are shown in Table 6.

	ROUND GROUPS																																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
NEAREST NEIGHBOUR			X											X									X				X	X	X	X	X							X					
FURTHEST NEIGHBOUR			X													X							X				X	X	X		X								X	X		X	X
MEDIAN																											X		X		X											X	
GROUP AVERAGE																											X	X	X		X											X	
CENTROID																												X														X	
WEIGHTED AVERAGE																X																								X		X	X
WARDS																																								X		X	
MODE 1			X											X									X					X															
2			X						X			X		X									X				X	X			X												
3			X			X		X	X			X	X	X		X	X	X				X	X				X	X	X		X		X		X		X	X		X			X
4		X	X			X	X	X	X			X	X	X		X	X	X			X	X	X				X	X	X	X	X	X	X	X	X		X	X		X	X		X
5		X	X			X	X	X	X			X	X	X	X	X	X	X		X	X	X	X				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6		X	X			X	X	X	X		X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
DISSIMILARITY ANALYSIS						X									X	X											X	X					X					X	X			X	X
K-LINK 2			X																				X					X	X	X	X							X					
3			X										X	X							X	X	X				X	X		X							X	X					
4			X										X	X									X				X	X		X		X	X		X		X	X		X			
NUCLEUS 1			X									X		X				X			X	X	X					X	X	X													
2			X						X				X	X				X			X	X	X				X	X	X	X		X	X				X	X				X	
3			X			X			X				X	X	X						X	X	X				X	X	X	X	X	X		X	X		X	X	X	X	X	X	X
4			X			X		X	X				X	X	X	X	X	X			X	X	X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
BEALES																																								X		X	X
GROUP AVERAGE RELOC						X						X			X	X							X				X	X	X		X	X	X	X	X							X	X
NEIGHBOURHOOD 1																												X															
2																							X																				
3																																											
4																							X																				
5																							X																				

TABLE 6

☒ FAILURE ☐ SUCCESS

ROUND GROUPS

STRAGGLY GROUPS

44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 | 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96

N. NBR
F. NBR
MED.
G. A.
CENT.
W. A.
WARD
MODE 1
2
3
4
5
6
DISSIM.
KLINK 2
3
4
NUC. 1
2
3
4
BEALE
G.A.R.
NBR 1
2
3
4
5

☒ FAILURE ☐ SUCCESS

Constructed Null Hypotheses Results

We now consider the methods in the light of the findings from the random tests, as described previously on pages 240-1. We will not analyse in detail the results from the null hypotheses assuming equal variance along axes, since this may be too sweeping an assumption. We will, however, give a table which shows the number of successes and failures of each type.

ROUND TESTS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
N. Neighbour	23	15	11	14	1
F. Neighbour	36	7	6	8	7
Median	38	8	3	11	4
Group Average	47	6	2	4	5
Centroid	55	2	1	4	2
Weighted Average	41	5	2	8	4
Ward's	42	9	1	9	3

Centroid dominates all the other methods on the above figures, also weighted average and Ward's method dominate median, and Ward's and group average dominate furthest neighbour. Centroid had the highest success, and nearest neighbour fared badly.

STRAGGLY GROUPS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
N. Neighbour	16	8	4	3	1
F. Neighbour	0	3	20	2	7
Median	1	4	18	5	4
Group Average	1	6	17	5	3
Centroid	2	7	8	5	10
Weighted Average	1	4	8	6	13
Ward's	1	8	16	3	4

Nearest neighbour is easily the best, and weighted average and centroid are the ones which can give the most misleading results. The results of each method on the round cluster hypothesis are shown in Table 7.

We can examine the results on the full null hypothesis in more detail, since it involves less assumptions. We first present the two tables.

ROUND GROUPS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
N. Neighbour	15	32	11	5	1
F. Neighbour	19	27	11	4	3
Median	30	21	3	6	4
Group Average	25	29	6	3	1
Centroid	32	25	2	4	1
Weighted Average	29	28	4	2	1
Ward's	33	23	3	4	1

Text cut off in original

TABLE 7

ROUND CLUSTER HYPOTHESIS

ROUND GROUPS (1–23)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
NEAREST NEIGHBOUR	W	W	X			R			W				R	X	R	R	R	R	W	W	S	R	X
FURTHEST NEIGHBOUR			X						W					R	R	X							S
MEDIAN			X				W	W	W	W	W	W			R	R							
GROUP AVERAGE																R							
CENTROID																R							
WEIGHTED AVERAGE															R	X							
WARDS														R	R	R							
K-LINK 2	W	W	X			R	W	W	W	W	W	W	R	R	R	R	R	R	W		W	R	S
3		W	S				W	W		W		R	X	X	R	R	R	R			S	X	X
4	W	R	S				W	W	W			W	X	X	R	R	R	R			W		S
NUCLEUS 1		W	X			W			R	W	W	S	R	X	R	R	X	X	W	W	S	X	X
2		W	X			W	W	W	S			R	X	X	R	R		X	W	W	S	X	X
3			X			X		W	S				S	X	S		W	W			S	X	X
4		W	S			X		X	X				X	X	X	X	X	S	W	W	S	X	X
BEALES																R							
GROUP AVERAGE RELOC.						S						S		R	X	X	R	R				R	X

ROUND GROUPS (24–45)

	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
NEAREST NEIGHBOUR	W	W	W	X	X	X	X	X		R	W			R	X	R	W		R		W	R
FURTHEST NEIGHBOUR			W	X	X	S	W	S	R	R	R		W	R	R	S	S	W	X	X	S	
MEDIAN				X	R	S	R	S	R							R	W	W	X	R	S	
GROUP AVERAGE				X	X	S	W	S	R	R						R	W		S	R	S	
CENTROID					X												W		S		S	
WEIGHTED AVERAGE				R	R		W		R	R	W			W	W	W	S	W	S	X	S	
WARDS				W	W		W		R	R	R	W	W	W	R	X	S	W	S	[illegible]	S	
K-LINK 2			W	R	X	S	X	S	R	R	W		W	R	X	R	W		R			R
3				X	X		X		R	R		W	W	X	X	R	W					R
4				X	X		X		X	X	R	X	W	X	X	R	S	W	R	X		X
NUCLEUS 1			W	R	X	S	X	W		W	R	R		W	R	R	R				R	R
2		W		X	X	S	X	W	S	S		R		S	X	R	R	W	S	[illegible]	R	W
3				X	X	S	X	S	X	W	S	X	W	S	X	X	X	S	S	S	X	W
4		W	S	X	X	S	X	S	X	X	X	X	X	X	X	X	X	X	X	X	S	X
BEALES									R	R	W	W	W	R	R	R	S	W	S	S	S	
GROUP AVERAGE RELOC.				S	S	S		S	X	X	X	S	W	R	R	R	W	W	S	S		S

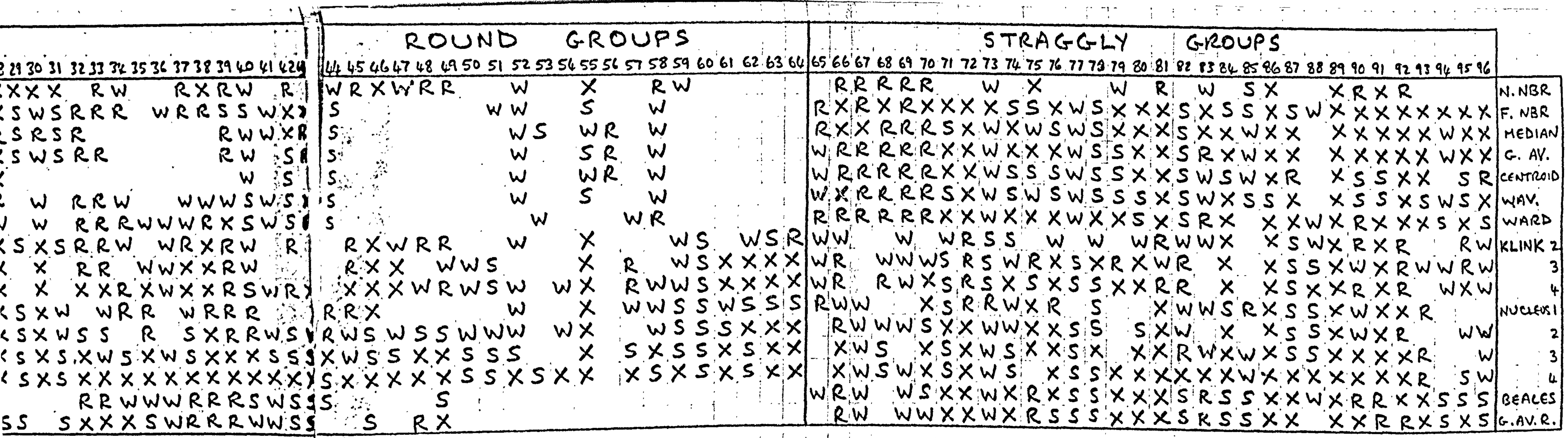
☐ SUCCESS

[R] RIGHT CLUSTERS + NOT SIGNIFICANT

[X] WRONG CLUSTERS + NOT SIGNIFICANT

[W] RIGHT CLUSTERS + WRONG NO. SIGNIF.

[S] WRONG CLUSTER + SIGNIFICANT



LIGHT CLUSTERS
WRONG NO. SIGNIF.

S WRONG CLUSTER + SIGNIFICANT

☐ SUCCESS

R RIGHT CLUSTERS + NOT SIGNIFICANT

X WRONG CLUSTERS + NOT SIGNIFICANT

W RIGHT CLUSTERS + WRONG NO. SIGNIFIC.

S WRONG CLUSTERS + SIGNIFICANT

STRAGGLY GROUPS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO SIGNIF.	WRONG CLUSTERS AND SIGNIF.
N. Neighbour	13	11	5	3	0
F. Neighbour	0	4	26	1	1
Median	1	7	21	2	1
Group Average	0	9	19	3	1
Centroid	1	11	17	2	1
Weighted Average	0	10	20	1	1
Ward's	0	11	19	1	1

Again the most striking feature is the failure of all but the nearest neighbour on the straggly groups. Another important feature is the low figures in the two right hand columns, indicating how 'safe' the methods are - giving misleading results in a maximum of 13.5% of cases (median), and as few as 5.2% (weighted average), and most methods give completely wrong results in about 1% of the tests.

The best methods achieve a 50% success with round groups, and negligible success with straggly - whereas nearest neighbour has 23% success on round and 41% on straggly. The higher percentage on straggly groups is caused by these tests being somewhat easier to group by eye, and hence easier for a straggly-type method.

The methods, with the exclusion of nearest neighbour, are difficult to differentiate on the basis of these results, but general conclusions can be drawn. Furthest neighbour is

probably least successful, being only slightly better than nearest neighbour on round groups. Centroid and Ward's method are the most successful with almost identical results. Centroid dominates median, and weighted average dominates group average. The results of each method on the full null hypothesis are shown in Table 8.

We now consider the extension of Lance and Williams' flexible method. This accounts for a large number of our investigations. Table 9 shows the number of failures on the 64 round tests. Rough contours have been added to highlight the behaviour of α and β. There was a wide range of success from 60-100%. There was a marked difference from $\alpha = 0.45$ to $\alpha = 0.5$. In general, the lower the value of β, the better the results. The results on Lance and Williams' line are of interest - the level of success is not sensitive in the region $\alpha = 0.5$ to 0.7, but the suggested value of $\alpha = 0.625$ by Lance and Williams appears to be a little high. If a dendrogram is required then the best position on this diagram is at $\alpha = 0.575$, $\beta = -0.150$. With these parameters there were three failures on the round tests, the same three as those with Ward's method.

The results with straggly groups are shown in Table 10, as with most of the methods discussed previously, there was little success. An interesting result is the slight improvement as β increases, as opposed to the improvement with decreasing β noted on the round tests. Also for each

ROUND GROUPS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
NEAREST NEIGHBOUR	R	R	X		R	R		R	R				R	X	R	R	R	R	W	W	S	R
FURTHEST NEIGHBOUR		R	X	W	R	R		R	R	W	R	R	R	R	R	X		R				R
MEDIAN		R	X				W	R	R	W	W	W		R	R	R						R
GROUP AVERAGE	W	R	R			R			R	R			R	R	R	R	R	R				R
CENTROID		R	R										R	R	R	R	R	R				R
WEIGHTED AVERAGE		R	R			R							R	R	R	X	R	R				R
WARDS													R	R	R	R						R
K-LINK 2	W	R	R			R	W	W	R	W	W	W	R	R	R	R	R	R	W		W	R
3	R	R	X			R	R	R	R	W		R	X	X	R	R	R	R			X	X
4		R	X				R	R			W	W	X	X	R	R	R	R			W	R
NUCLEUS 1		R	X		R	R		R	R	W	W	S	R	X	R	R	X	X	W	W	S	X
2	R	R	X		R	R	R	R	X		R	R	X	X	R	R	R	X	W	W	X	X
3		R	X		R	X	R	R	X	R	R	R	X	X	X	R	R	R			X	X
4	R	R	X			X	R	X	X			R	X	X	X	X	X	X	R	W	X	X
BEALES			R						R			R	R	R	R	R	R	R				R
GROUP AVERAGE RELOC.			R			S		R	R			S	R	R	X	X	R	R				R
MODE 3			X			S		S	X			X	X	X		X	S	X				S
4		X	X			S	S	X	X			X	X	X	R	S	S	S			S	S
5		X	X			S	X	X	X			X	X	X	X	X	X	X		S	S	X
6		X	X			S	X	X	X		S	X	X	X	X	X	X	X		S	X	X

	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
NEAREST NEIGHBOUR	X	W	W	R	X	X	X	X	X	R	R	R			R	X	R	R	R	R	R
FURTHEST NEIGHBOUR	X			W	X	X	S	R	X	R	R	R	R	R	R	R	X	X	R	X	X
MEDIAN	R				X	R	S	R	S	R	R	R				R	R	R		X	R
GROUP AVERAGE	R				X	X	X	R	X	R	R	R	R	R	R	R	R	R	R	X	R
CENTROID	R				R	X		R		R	R	R				R	R	R		X	R
WEIGHTED AVERAGE	R				R	R	R	R	R	R	R	R	R	R	R	R	R	X	R	X	X
WARDS	R				R	R		W		R	R	R	R	R	R	R	X	X	R	X	R
K-LINK 2	X			W	R	X	X	X	X	R	R	R		W	R	X	R	R		R	R
3	X			R	X	X	R	X	R	R	R	R	W	R	X	X	R	R	R	R	R
4	X			R	X	X	W	X	R	X	X	R	X	R	X	X	R	X	R	R	X
NUCLEUS 1	X	R	R	R	R	X	X	X	R	R	R	R	R	R	R	R	R	R	R		R
2	X		R	R	X	X	X	X	R	X	S	R	R	R	X	X	R	R	R	X	R
3	X		R	R	X	X	X	X	X	X	W	X	X	R	X	X	X	X	X	X	X
4	X	R	R	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
BEALES	R				R	R		R		R	R	R	R	R	R	R	R	X	R	X	X
GROUP AVERAGE RELOC.	X				X	X	S		S	X	X	X	X	R	R	R	R	R	R	X	X
MODE 3	X				S	S	S		S		S		S		X	X		S			S
4	S				X	X	S	S	S	S	X	S	S		X	X		X	X		S
5	X				X	X	S	S	S	X	X	X	S	S	X	X	S	S	S	X	S
6	X	S	S	S	X	X	S	X	S	X	X	X	S	S	X	X	X	S	X	X	S

TABLE 8

FULL NULL HYPOTHESIS

☐ SUCCESS [R] RIGHT CLUSTERS + NOT SIGNIFICANT [X] WRONG CLUSTERS + NOT SIGNIFICANT [W] RIGHT CLUSTERS + WRONG NO. SIGNIF. [S] WRONG CLUSTERS + SIGNIFICANT

ROUND GROUPS

	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
N.NBR	R	R	X	R	R	R		R	R		R	X			R	W					R
F.NBR	S				R	R	R	W	R	S	R	X			R						
MEDIAN	S				R	R			W	S		R	R		W						
GRP. AV	S		W		R	R			R			X	R		W						
CENTROID	S	W	W	W	R	R			W		R	R	R		R						
WGT. AV	S				R	R			W		R	R			W						
WARD	S		R		R	R		W	R	W		R		W	R						
K-LINK 2		R	X	R	R	R			R			X			R	R	X		R		R
3		R	X	X	R	R	R	X			R	X		R	R	R	X	X	X	X	X
4	W	X	X	X	R	R	R	X	R		R	X		R	R	R	X	X	X	X	X
NUCLEUS 1	R	R	X	R	R	R		R	R		R	X		W	W	S	S	R	X	S	X
2	R	W	S	W	X	X	W	R	R		R	X			R	X	S	X	X	X	X
3	X	W	S	S	X	X	S	S	S		R	X		S	X	X	X	X	X	X	X
4	S	X	X	X	X	X	X	X	X	S	X	X		X	S	X	X	X	X	X	X
BEALES	S				R	X			R												
GRP. AV. R	S		S		R	X															
MODE 3			S			R				S						S	S		S	S	S
4			X	S		S	S	S	S	S						S	S	S	X	S	S
5	S	S	S	S	X	X	S	S	S	S		X	S	S	S	S	S	S	X	S	S
6	S	X	X	S	X	S	S	X	S	S	X	X	S	S	X	S	S	S	X	S	X

STRAGGLY GROUPS

	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
N.NBR		R	R	R	R	R		R	W		X	R			W		R		W		X	X			X	R	X	R	R			
F.NBR	R	X	R	X	R	X	X	X	X	X	X	X	R	X	X	X	X	X	X	X	S	X	X	W	X	X	X	X	X	X	X	X
MEDIAN	R	X	X	R	R	R	X	X	W	X	R	X	R	S	X	X	X	X	X	X	W	X	X		X	X	X	X	X	R	X	X
GRP. AV	R	R	R	R	R	R	X	X	W	X	X	X	W	X	X	X	X	S	R	X	R	X	X	W	X	X	X	X	X	R	X	X
CENTROID	R	R	R	R	R	R	X	X	W	X	X	X	W	X	X	X	X	S	R	X	R	X	R		X	X	X	X	X	R	X	R
WGT. AV	R	X	R	R	R	R	X	X	R	X	R	X	R	S	X	X	X	X	R	X	X	X	X	W	X	X	X	X	X	R	X	X
WARD	R	R	R	R	R	R	X	X	R	X	X	X	R	X	X	X	X	S	R	X	R	X	X	W	X	R	X	X	X	X	X	X
K-LINK 2	W				R	R	W	R	S	S		R		W		W	R	W	W	X		X	X	R	X	R	X	R	R	R	R	R
3	W	R		R	R	R	X	R	X	R	R	X	S	X	R	X	R	R	R	X		X	X	S	X	R	X	R	R	W	R	R
4	R	R		R	R	X	X	R	X	X	X	X	S	X	X	X	R	R	R	X		X	X	X	X	R	X	R	R	W	R	R
NUCLEUS 1	R	R				X	S	R	R	R	X	R		X		W	X	W	W	X	R	X	X	X	X	R	X	X	R		R	R
2		R	W	R	R	X	X	X	W	R	X	X	S	X	R	X	X	R	R	X		X	X	X	X	R	X	R	R	R	R	R
3	R	X	W	X	R	X	X	X	R	X	X	X	X	X	R	X	X	R	R	X	R	X	X	X	X	X	X	X	R	R	R	R
4	R	X	W	X	R	X	X	X	R	X	R	X	X	X	X	X	X	X	X	X	R	X	X	X	X	X	X	X	R	R	X	R
BEALES	R	R	W	R	R	X	X	X	R	X	R	X	X	X	X	X	X	S	R	X	X	X	X		X	R	R	X	X	X	X	X
GRP. AV. R	R	R	W	R	W	R	X	X	W	X	R	X	S	X	X	X	X	S	R	X	X	X	X		X	X	R	R	X	X	X	X
MODE 3						S		S		X	X			X	S					S	X	S	R	S	X	R	S	X		S	S	S
4						S		S		X		X	S	X		S	S		X	S	X	X	X	S	X	X	S	X	S	X	X	
5						X	X	X		X	R	X	S	X	X	X	X		X	S	X	X	X	S	X	X	X	X	X	X	X	S
6			X			X	X	X		X	X	X	S	X	X	X	X		X	X		X	X	S	X	X	X	X	X	X	S	S

☐ SUCCESS

[R] RIGHT CLUSTERS + NOT SIGNIFICANT

[X] WRONG CLUSTERS + NOT SIGNIFICANT

[W] RIGHT CLUSTERS + WRONG NO. SIGNIF.

[S] WRONG CLUSTERS + SIGNIFICANT

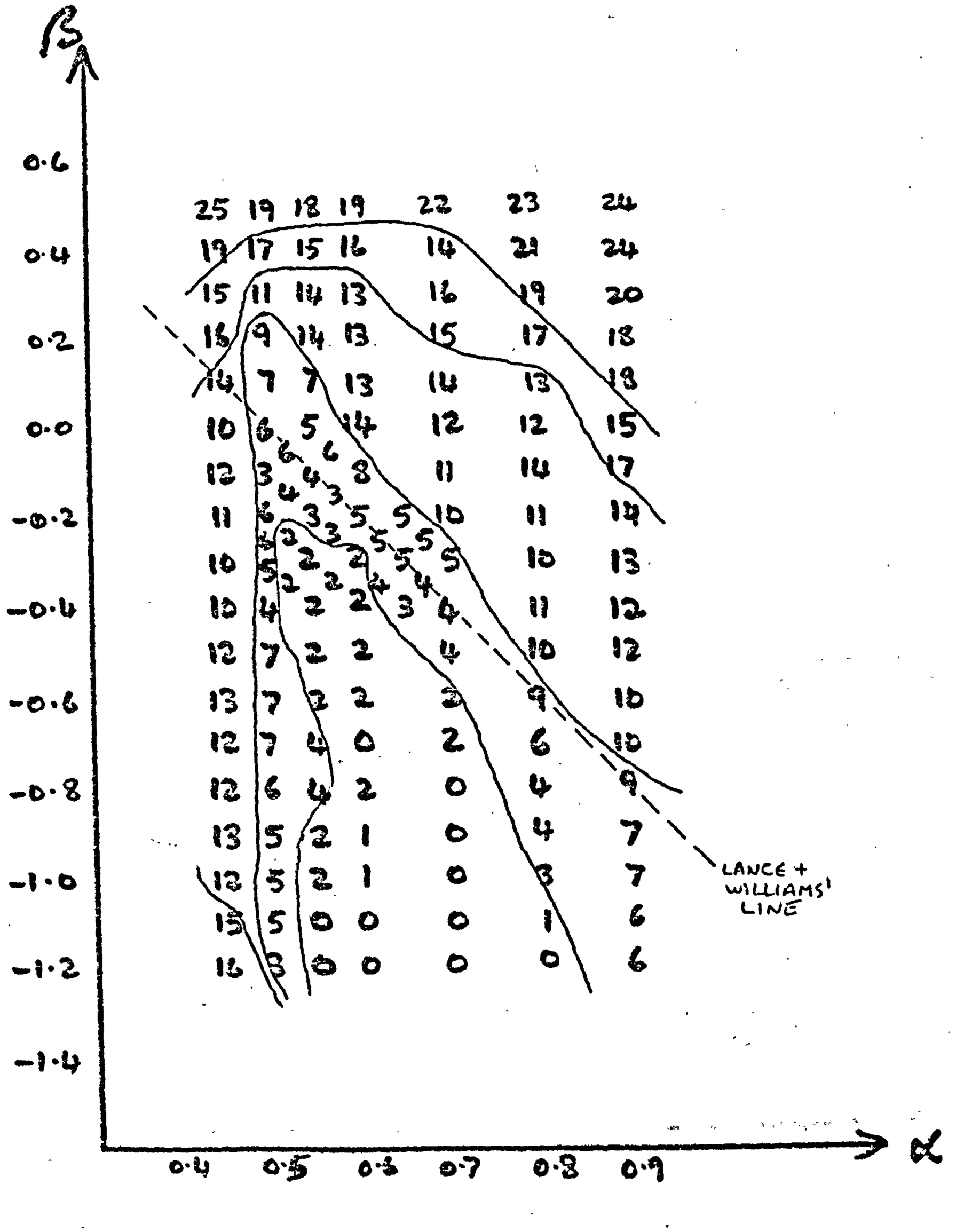


TABLE 9 - ROUND TESTS
FAILURES

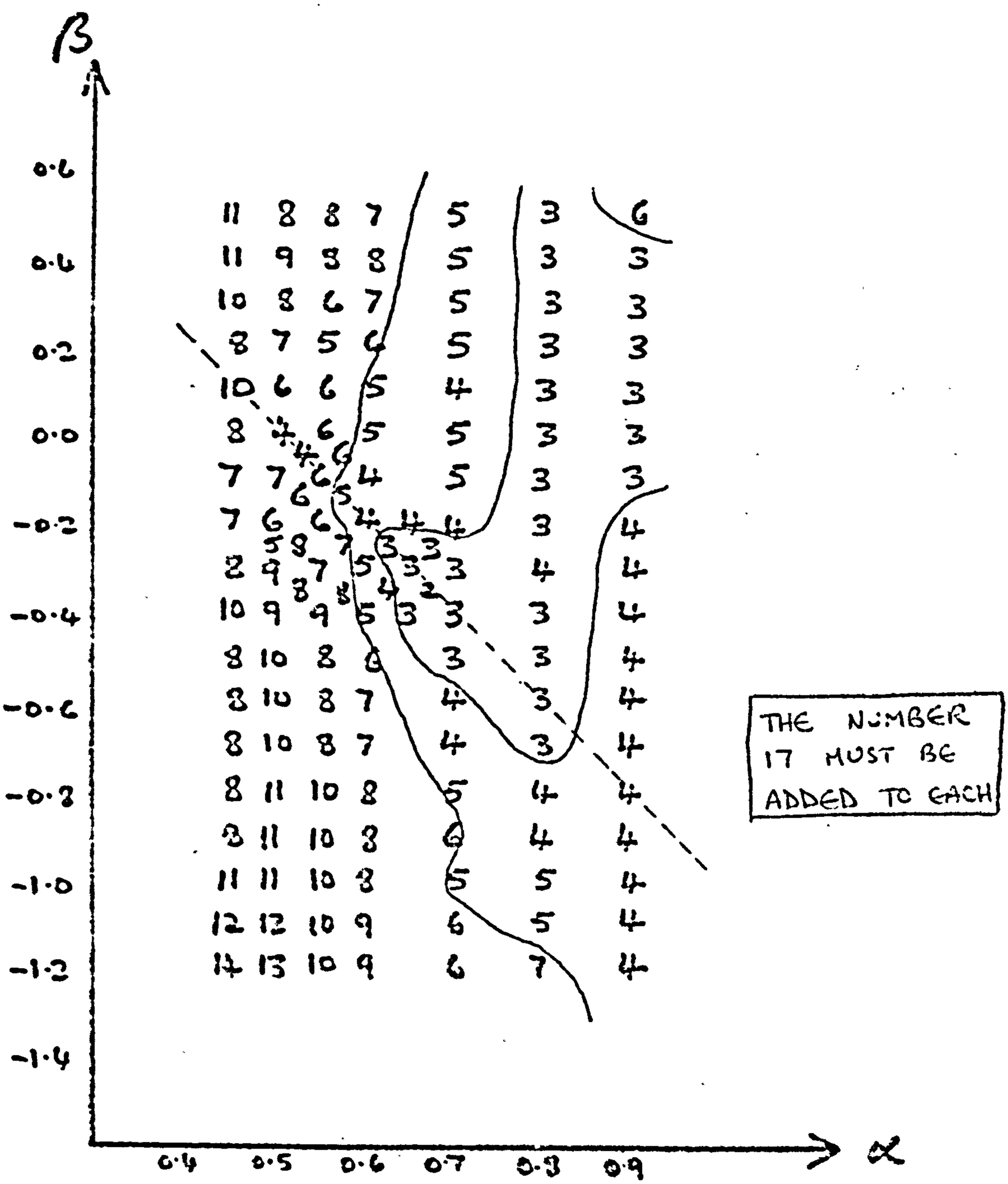


TABLE 10 - STRAGGLY TESTS FAILURES

value of α there is a tendency to reach a minimum on Lance and Williams' line, where a value of α between 0.6 and 0.8 seems best. Values between $\alpha = 0.575$ and 0.675 perform best on our full 96 tests and do almost as well as the centroid and Ward's method, and better than the other methods investigated.

If dendrograms are not required then optimum values are values of $\alpha = 0.7$ with β in the range -0.8 to -1.2 and possibly beyond. All of these successfully completed all 64 of the round tests, but were not quite as successful as some of the other hierarchical methods with straggly groups.

Not all the failures with particular parameters are of the same type. If we divide the tests into two classes - those with equal sized groups and those with unequal groups we obtain two charts as shown in Tables 11 and 12. Contours have been drawn at levels which show the pattern most clearly. Parameter values in the upper right of the graph do least well on unequal groups and $\alpha = 0.45$ values do worst on equal groups.

To examine the reasons for this effect we look at particular tests where this pattern of failure is most clearly illustrated. Firstly we look at the behaviour of the values around $\alpha = 0.9$, $\beta = 0.5$. The upper right failure pattern was exhibited most by tests 34 and 40. The results are shown below:

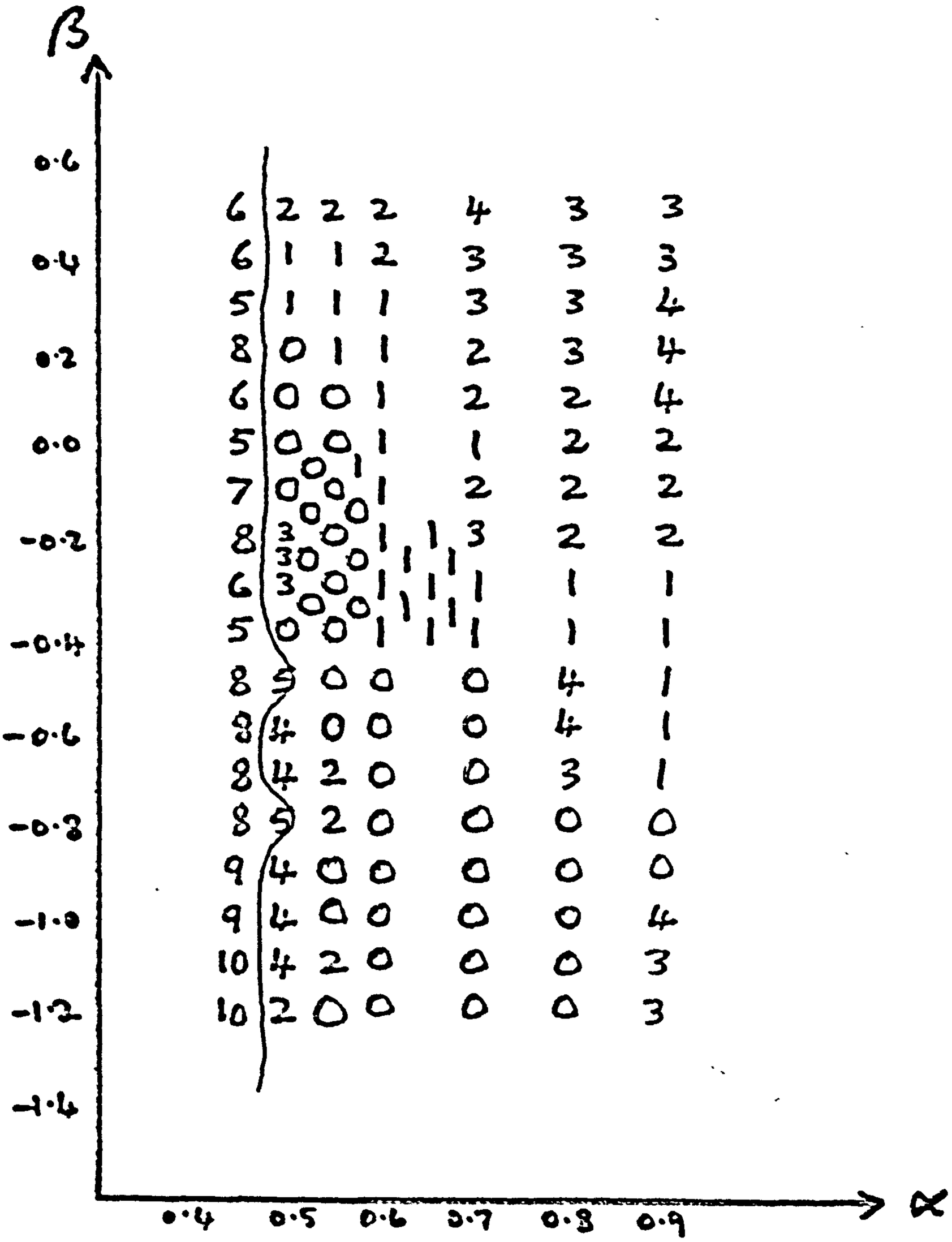


TABLE II - EQUAL CLUSTERS

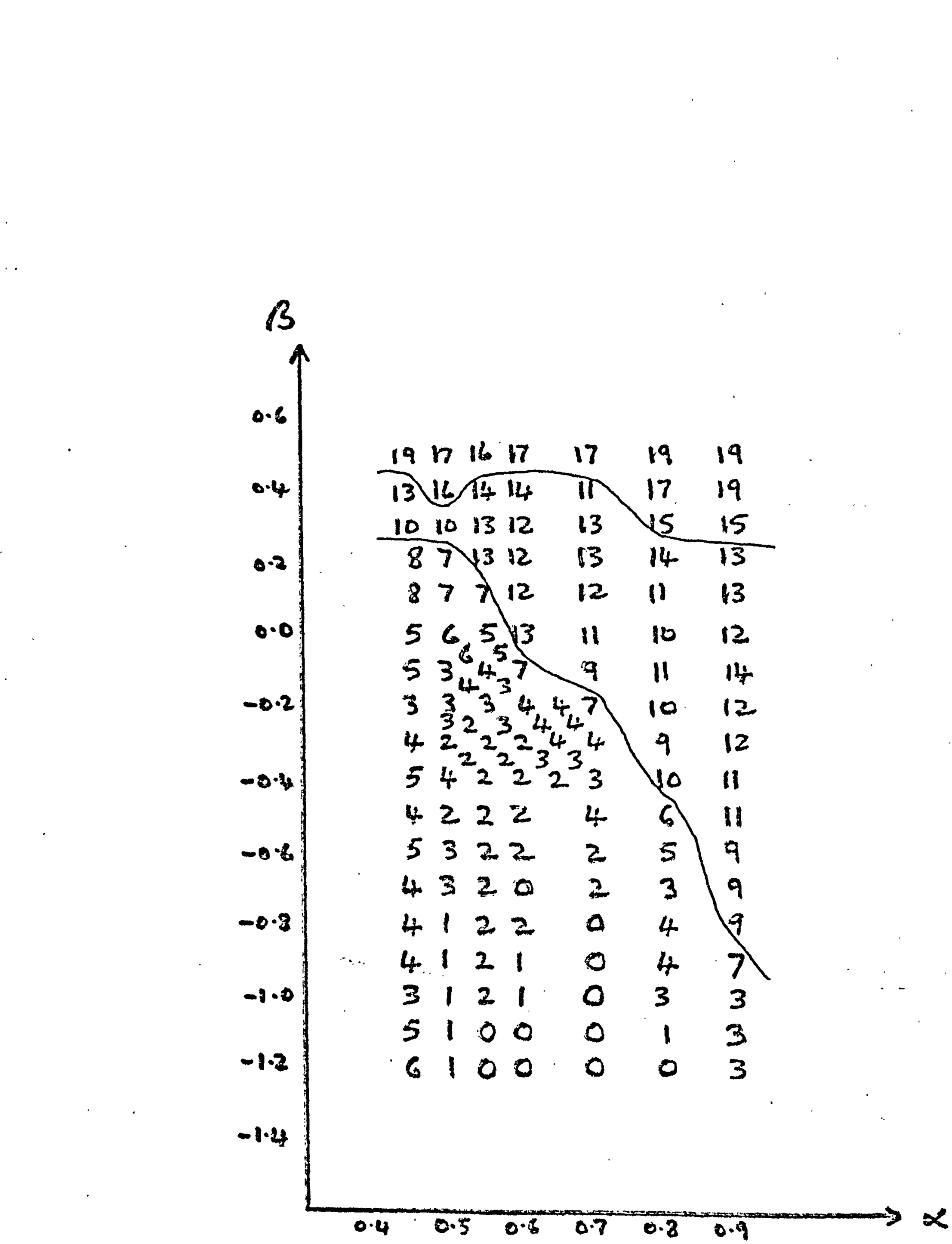
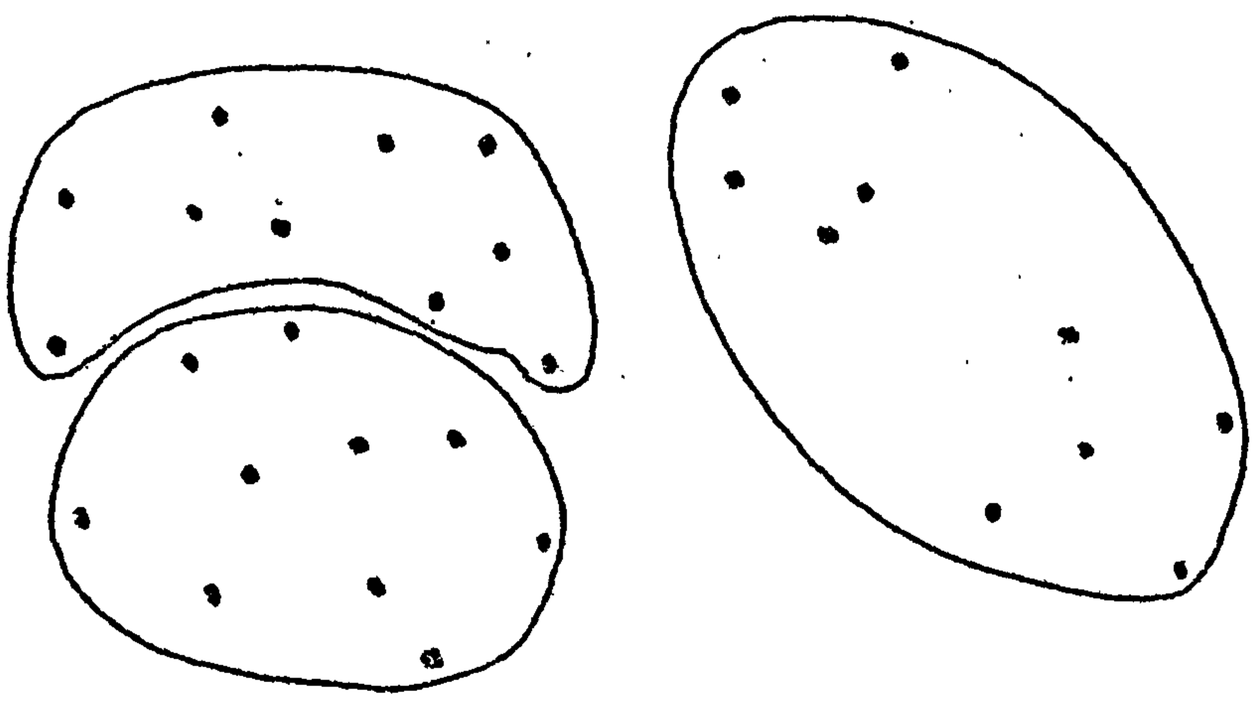
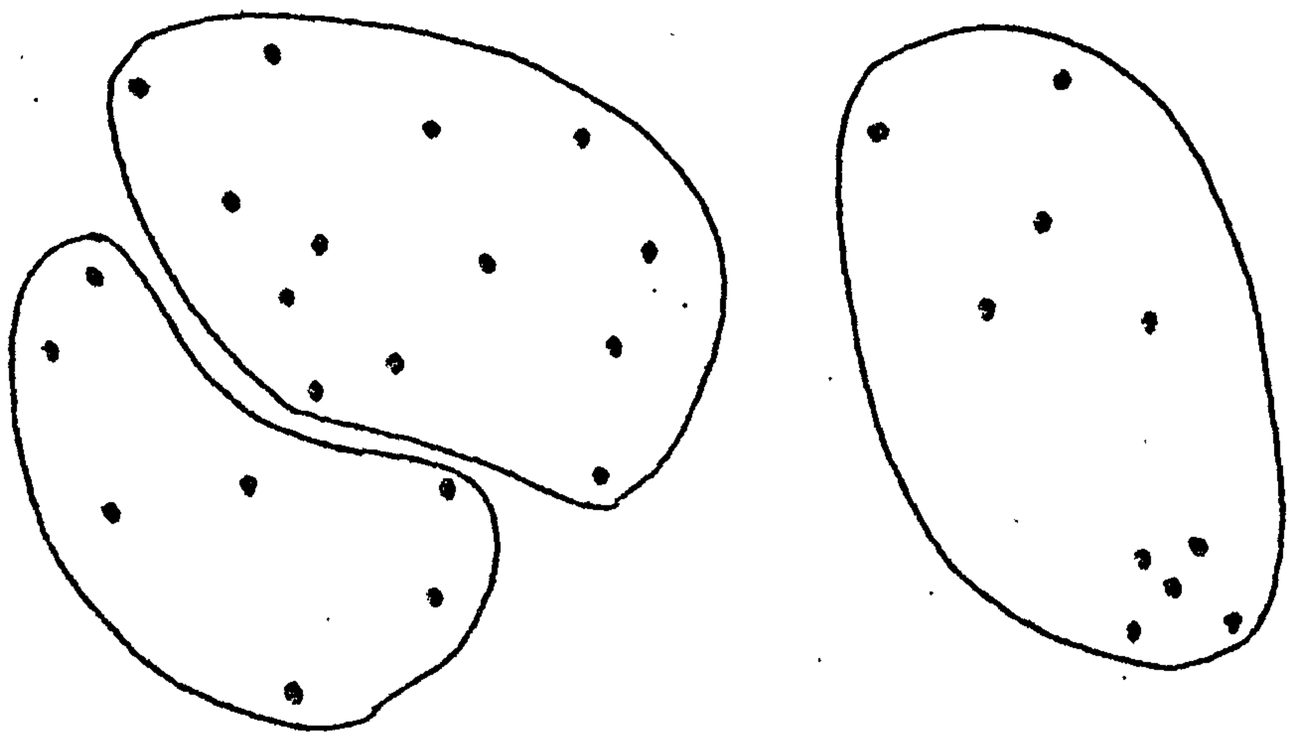


TABLE 12 — UNEQUAL CLUSTERS



TEST 34

$\alpha = 0.9 \quad \beta = 0.5$



TEST 40

$\alpha = 0.9 \quad \beta = 0.5$

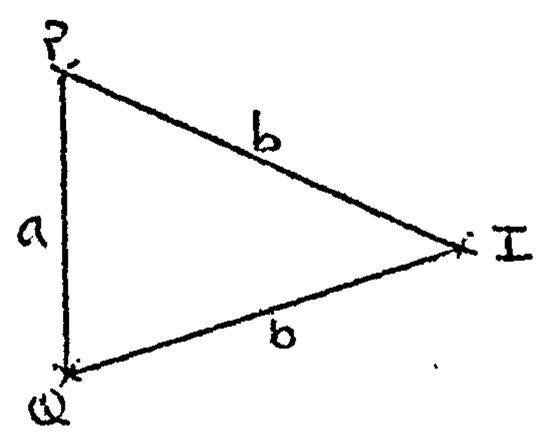
These failures were exhibited by a large number of parameters exclusively in the upper left corner. The reason for failure is fairly evident - that smaller groups have greater tendency to join than larger, and thus roughly equal sized groups are formed. This also explains the greater success of these parameter values on equal cluster tests.

This can be seen from the way the method works. The method is based on the way new distances are calculated when points merge. We have the new distance from point J to the new cluster

$$D(P',J) = \alpha(D(P,J) + D(Q,J)) + \beta D(P,Q)$$

If α and β are large, in particular if $\alpha > 0.5$ and $\beta > 0$, then the new calculated distances will tend to be much larger than D(P,J) or D(Q,J) and hence as clusters form they will be 'moved away' from the rest of the points. This causes objects to cluster into pairs, then into pairs of pairs, etc. This is the opposite effect to chaining where new clusters are 'moved' nearer to other objects. We call this effect pairing.

If we consider the case where we have three points:



then we would wish the new distance to the cluster to be no less than b.

I.e. $\quad b \geq 2b\alpha = a\beta$

as $b \to \infty$ the inequality becomes $\alpha \leq \frac{1}{2}$

as $b \to a$ we have $1 \geq 2\alpha + \beta$

and typically if $b = 3a/2$ we have $3 \geq 6\alpha + 2\beta$.

These lines are shown in Table 13.

The failure with $\alpha = 0.45$ can be investigated by examination of tests 22 and 27:

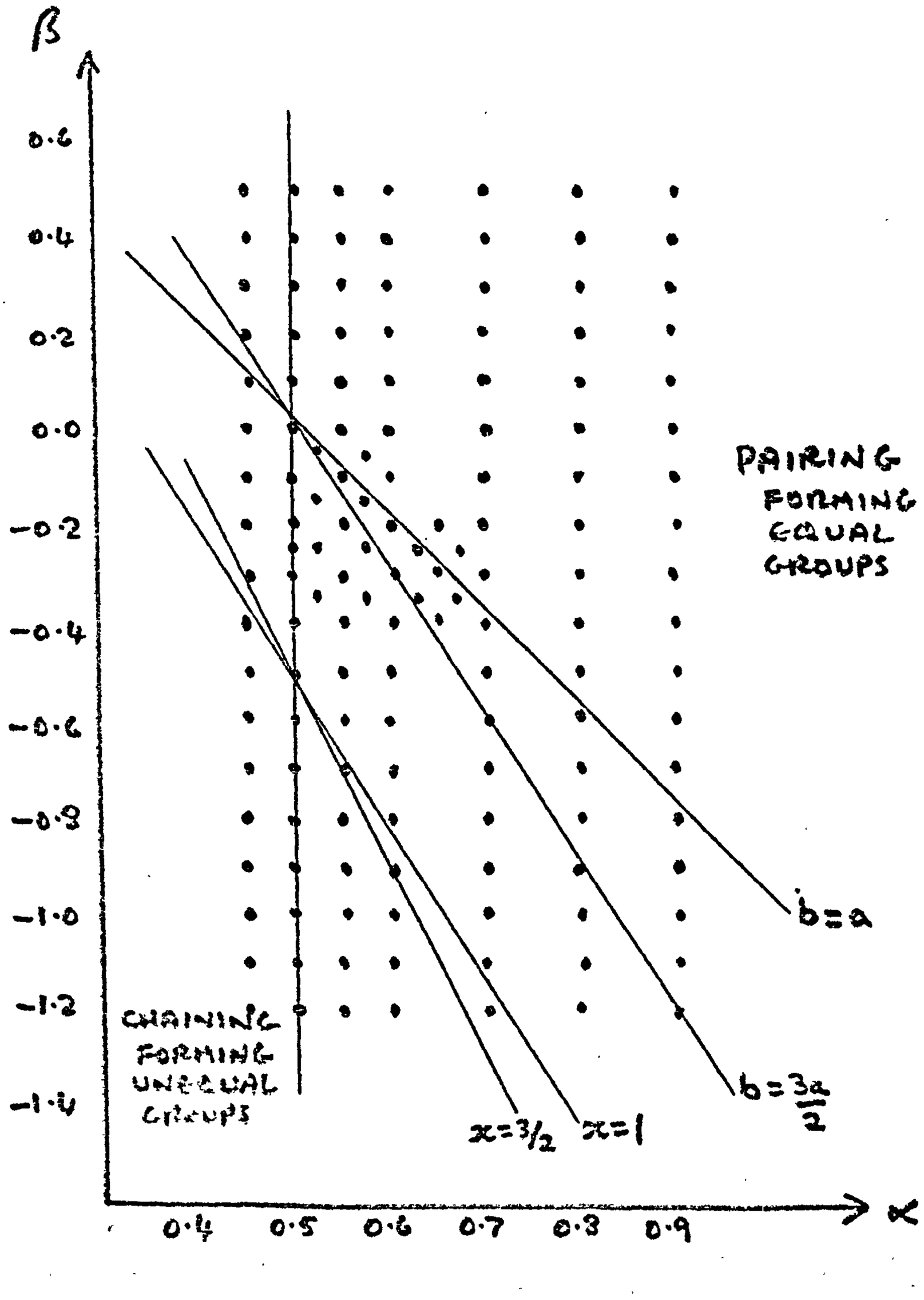
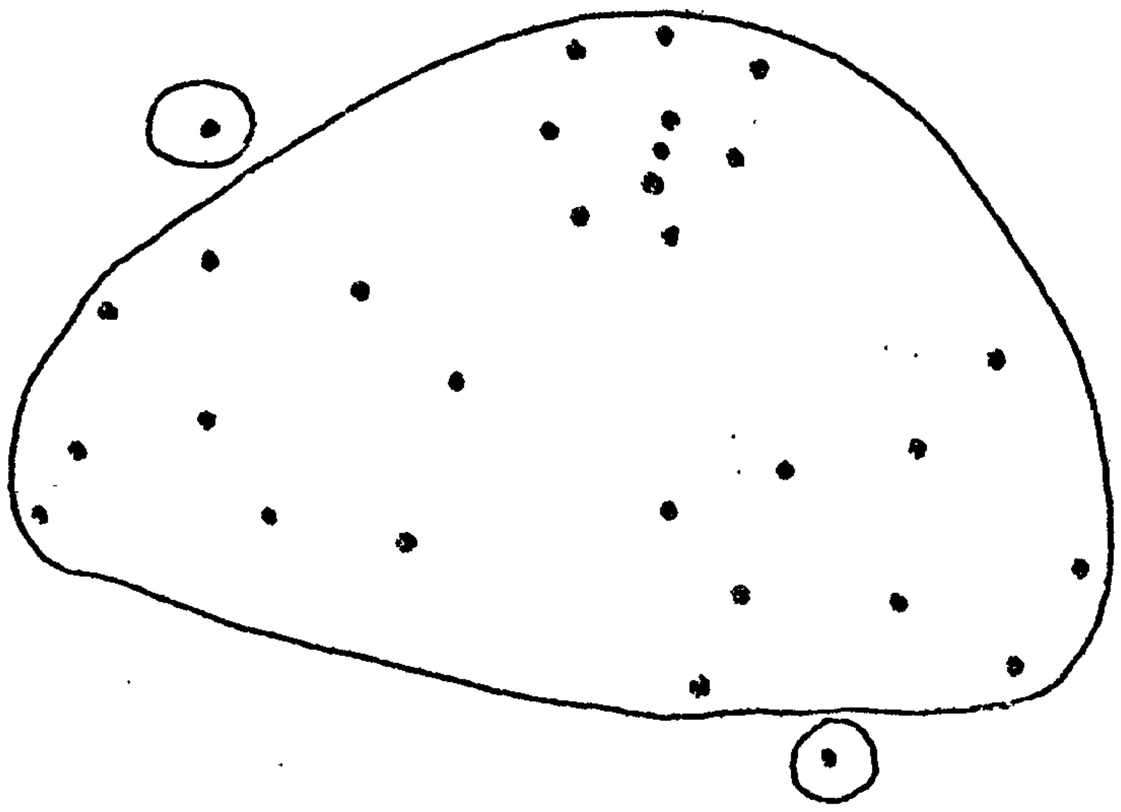
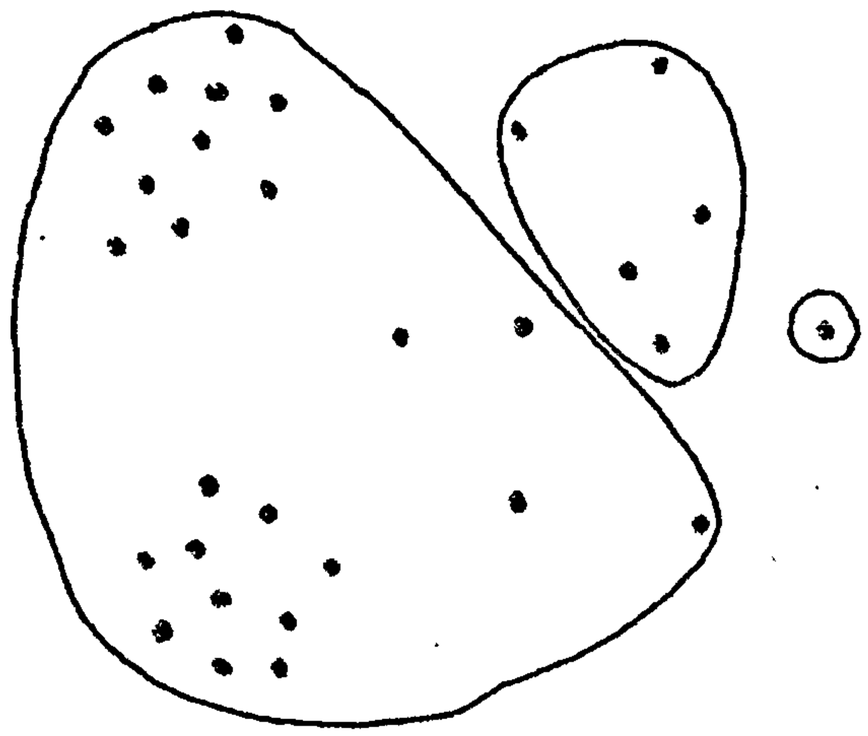

β
0·6
0·4
0·2
0·0
−0·2
−0·4
−0·6
−0·8
−1·0
−1·2
−1·4
PAIRING FORMING EQUAL GROUPS
CHAINING FORMING UNEQUAL GROUPS
b=a
b=3a/2
x=3/2
x=1
0·4
0·5
0·6
0·7
0·8
0·9
x

TABLE 13



TEST 22

$\alpha = 0.45$

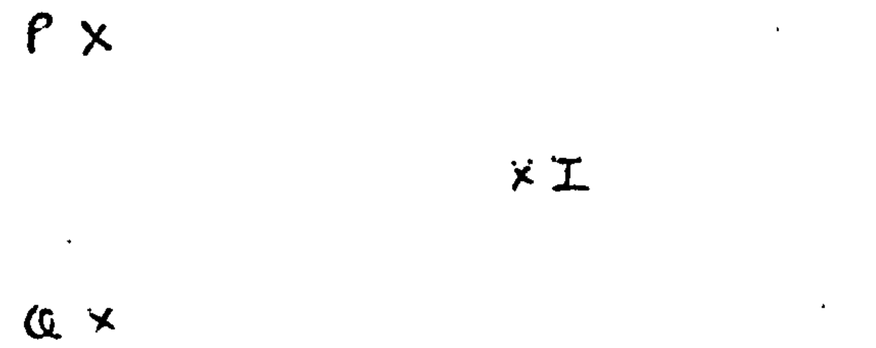


TEST 27

$\alpha = 0.45$

Test 22 failed with $\beta = -0.9$ to -1.2 and test 27 with $\beta = -0.5$ to -1.2, both with $\alpha = 0.45$.

This appears to be the opposite effect with larger clusters joining in preference to smaller ones. From the equation showing the distance calculation it can be seen that if α is less than 0.5 and also β is negative then the new distance will tend to be smaller than either D(P,J) or D(Q,J). We may wish this to be true in cases where we have



but certainly not where we have the case

Here our new distance

$$D = (2x + 1)\alpha a + \beta a$$

and we wish our new distance to be at least ax. So we require

$$ax \geq (2x + 1)\alpha a + \beta a$$

as $x \to \infty$ the equation becomes $\alpha = 0.5$

as $x \to 1$ the equation becomes $1 = 3\alpha + \beta$

Thus we have a family of lines each of which includes $\alpha < 0.5$ and $\beta \leq -0.5$ as a region in which chaining can occur.

This gives the situation in Table 13 which gives a good correspondence to our results with round groups.

In order to reduce the volume of computation required for the null hypothesis stages, we eliminated some parameters from our investigations. This elimination was carried out from Table 13 and the tests results (Tables 9 and 10), we reduced to a level which would still include the weighted average and median methods. The parameters used in the second half of the study are shown in Table 14.

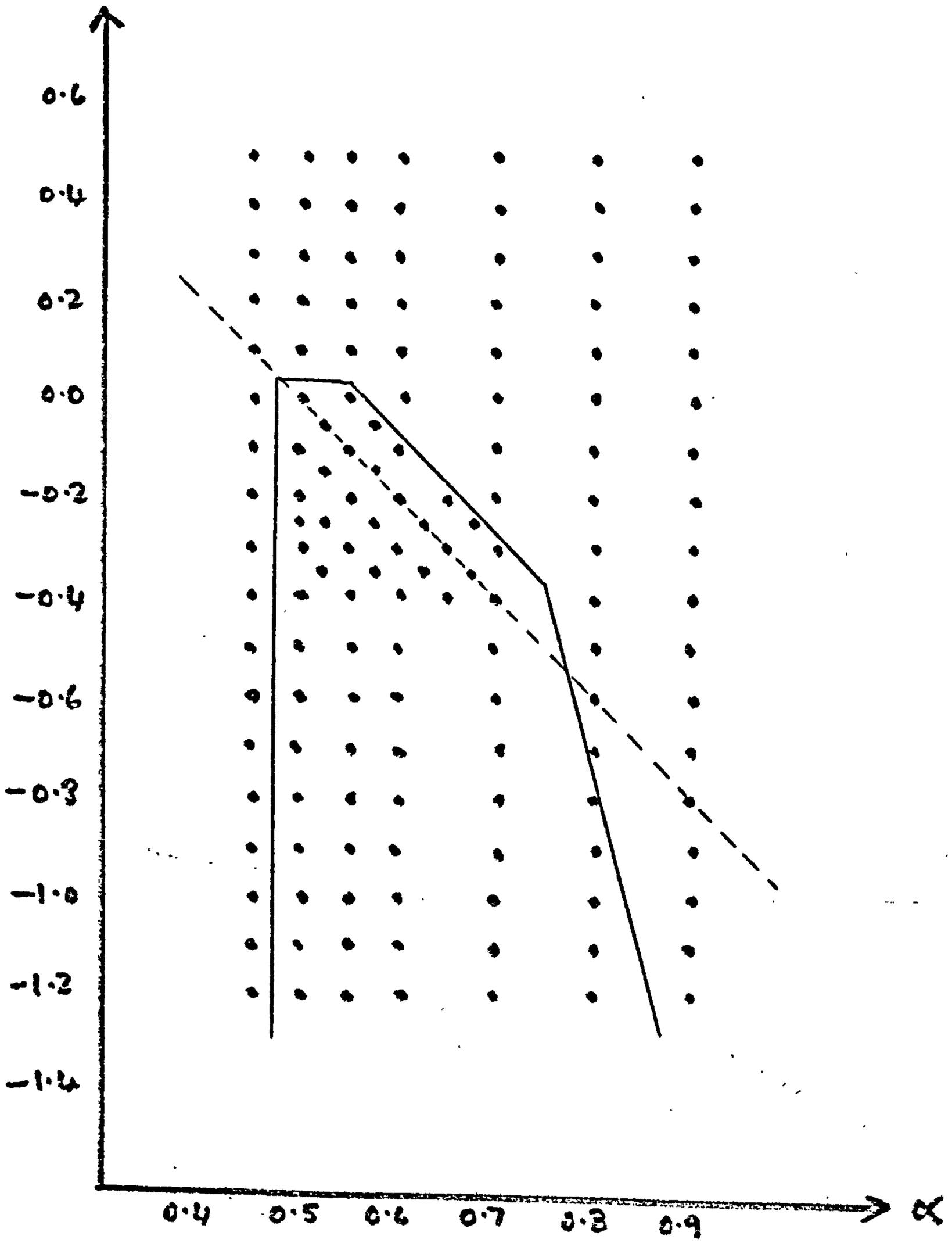


TABLE 14

As before, the results for the null hypothesis assuming equal variance along each dimension, are presented with little examination. These are shown in the Tables 15 and 16. The contours are similar to those in Tables 9 and 10, but are more marked. The best levels of success are as good as centroid, and better than the other methods discussed so far.

We now move on to the full null hypothesis. These results are shown in Tables 17 and 18. Little success was recorded with the straggly groups, but there is a tendency for values with high β to do slightly better.

The successes with round tests are much better than that with the other methods, and reaches 70%. The contours are again similar to those already noted. We can tabulate our best values of α and β.

Parameters		64 ROUND TESTS Level of success			32 STRAGGLY TESTS Level of success		
α	β	1	2	3	1	2	3
0.6	-0.5	62	56	40	9	2	1
0.6	-0.6	62	56	40	8	2	1
0.6	-0.7	62	55	45	8	2	1
0.7	-0.8	64	52	42	10	3	1
0.7	-0.9	64	55	43	9	1	1
0.7	-1.0	64	55	39	10	1	0
0.7	-1.1	64	57	45	9	2	0
0.7	-1.2	64	59	43	9	2	0
0.8	-1.1	63	53	38	10	1	0
0.8	-1.2	64	54	38	8	1	0

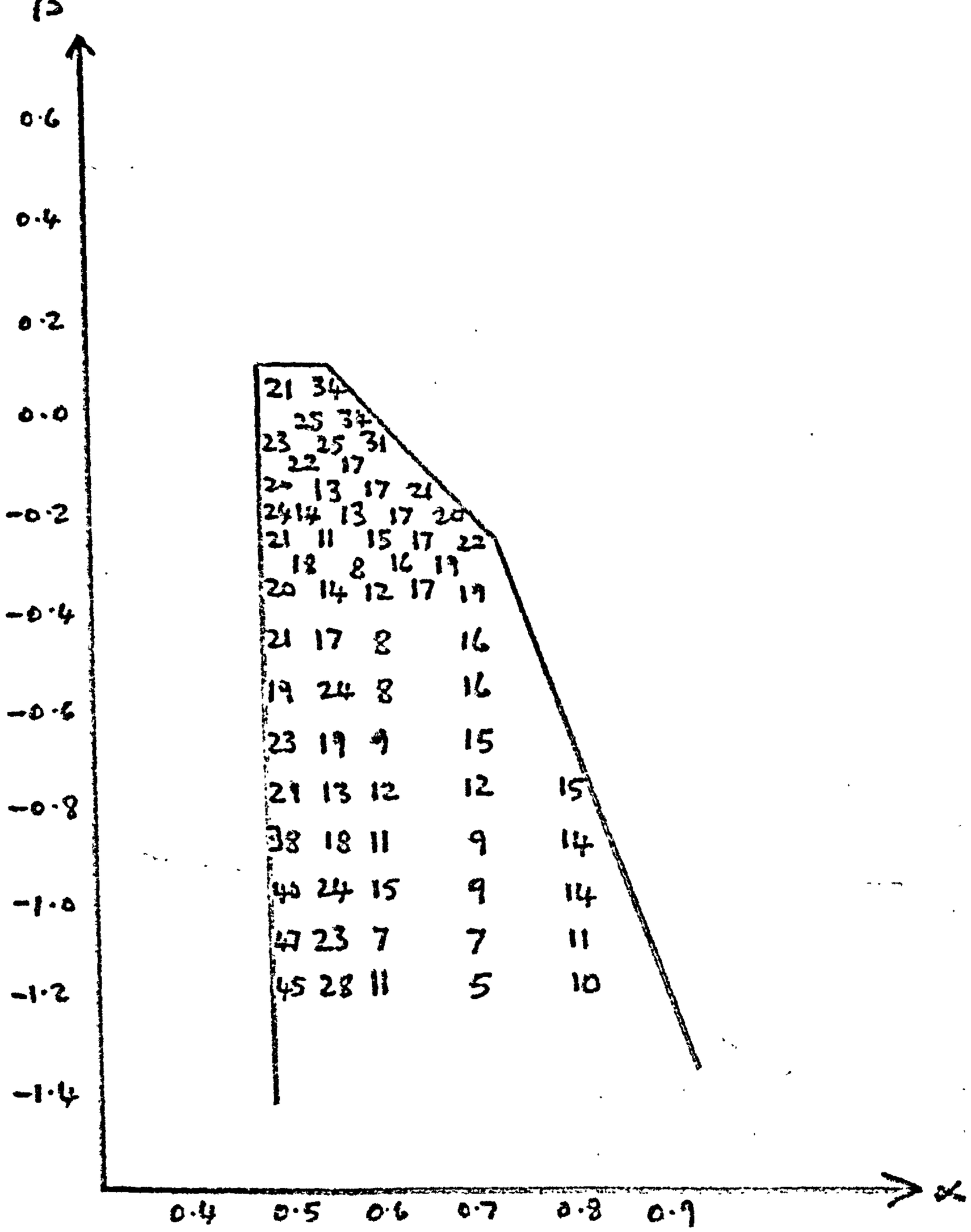


TABLE 15 ROUND TEST FAILURES

ROUND HYPOTHESIS

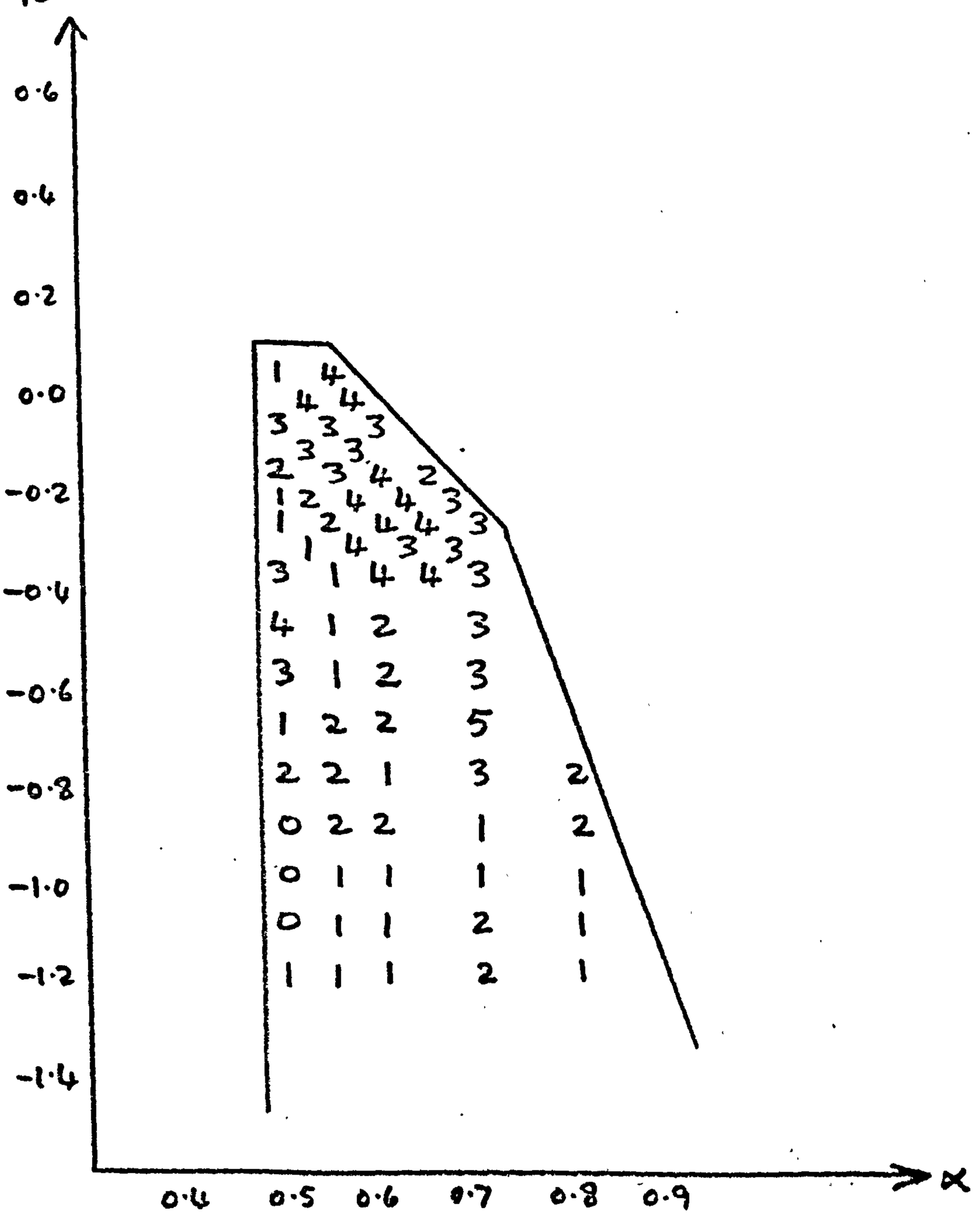
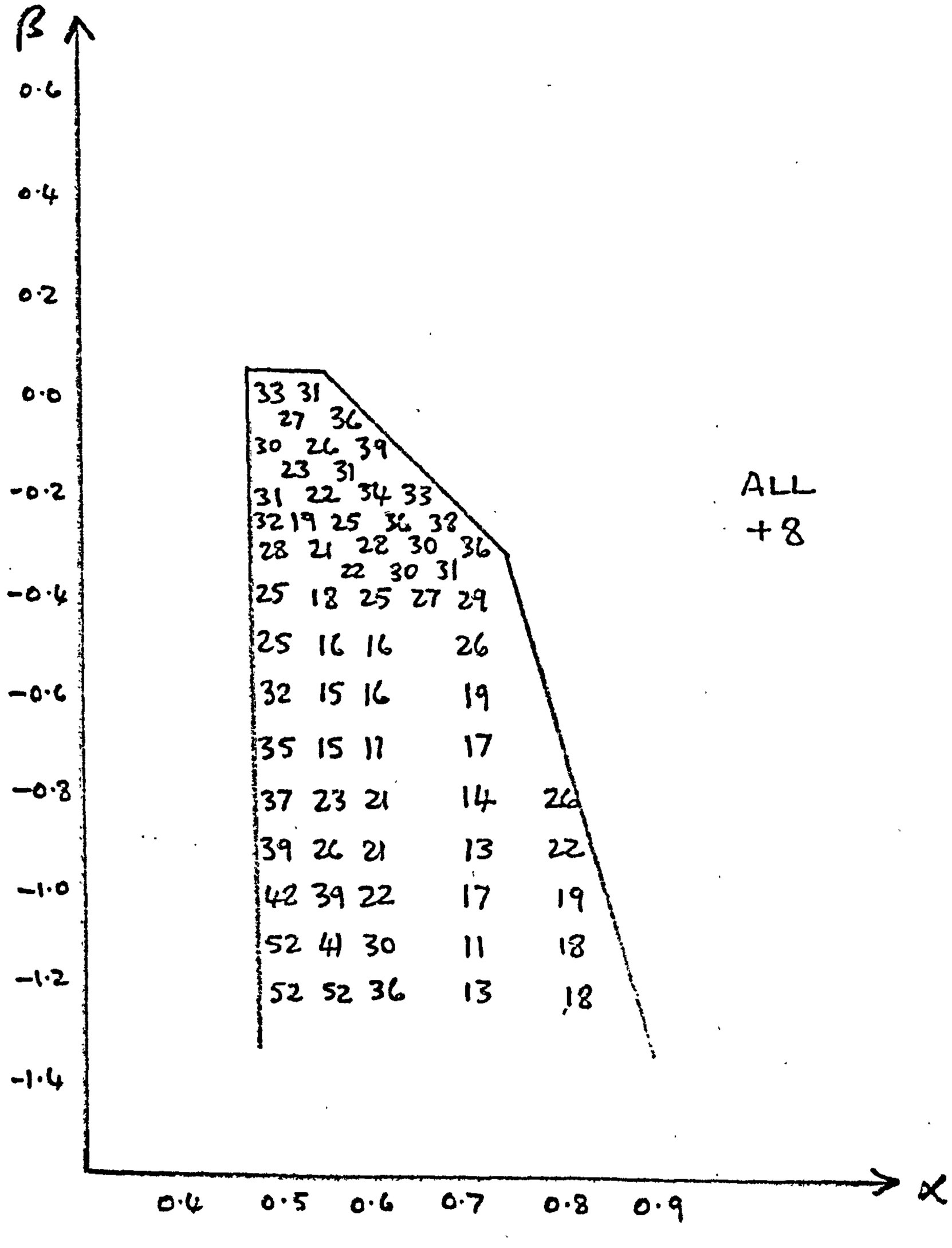


TABLE 16 STRAGGLY TEST SUCCESSES
ROUND HYPOTHESIS



ALL
+8

TABLE 17 ROUND TEST FAILURES

FULL HYPOTHESIS

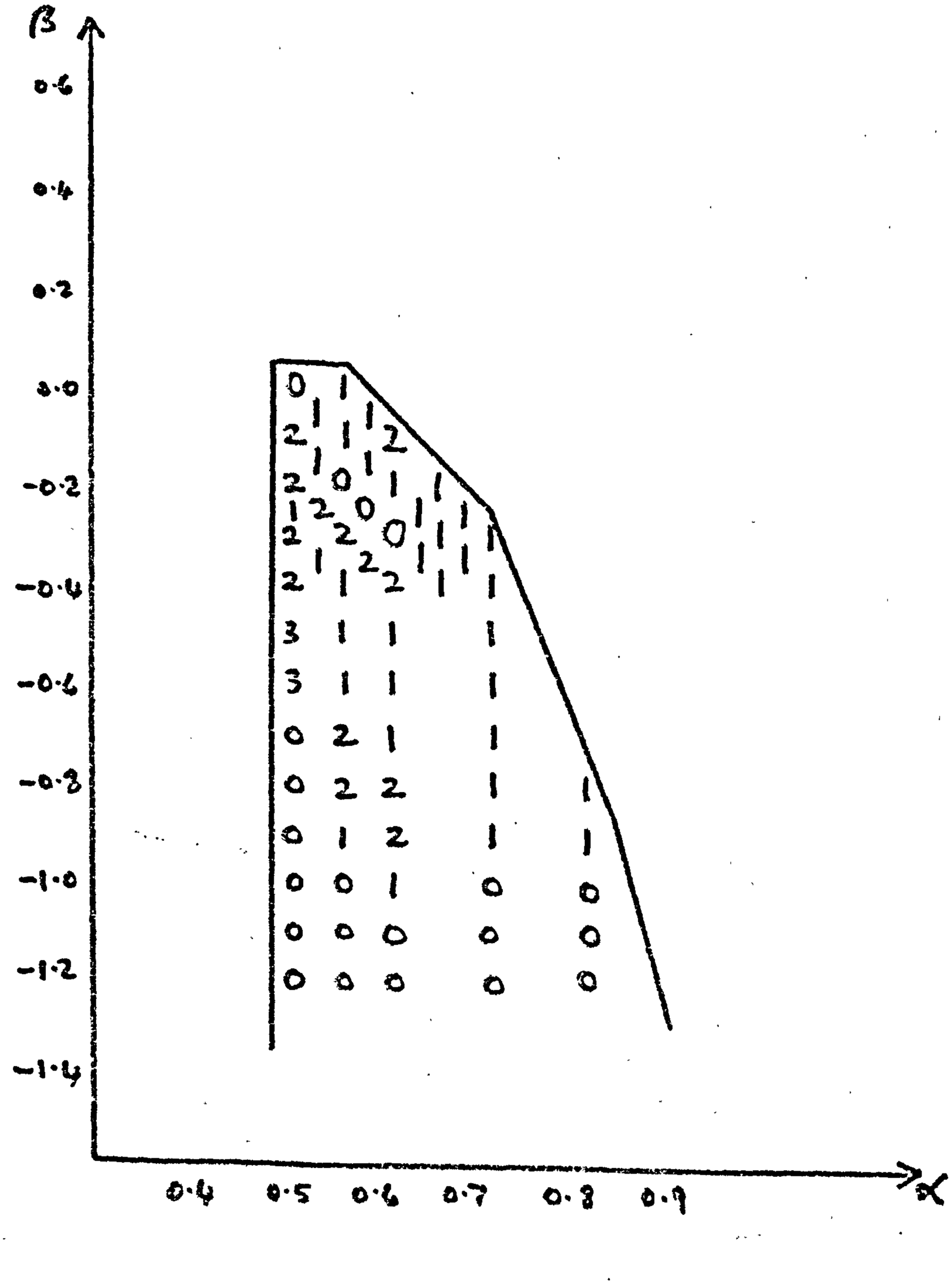


TABLE 18 STRAGGLY TEST SUCCESSES

FULL HYPOTHESIS

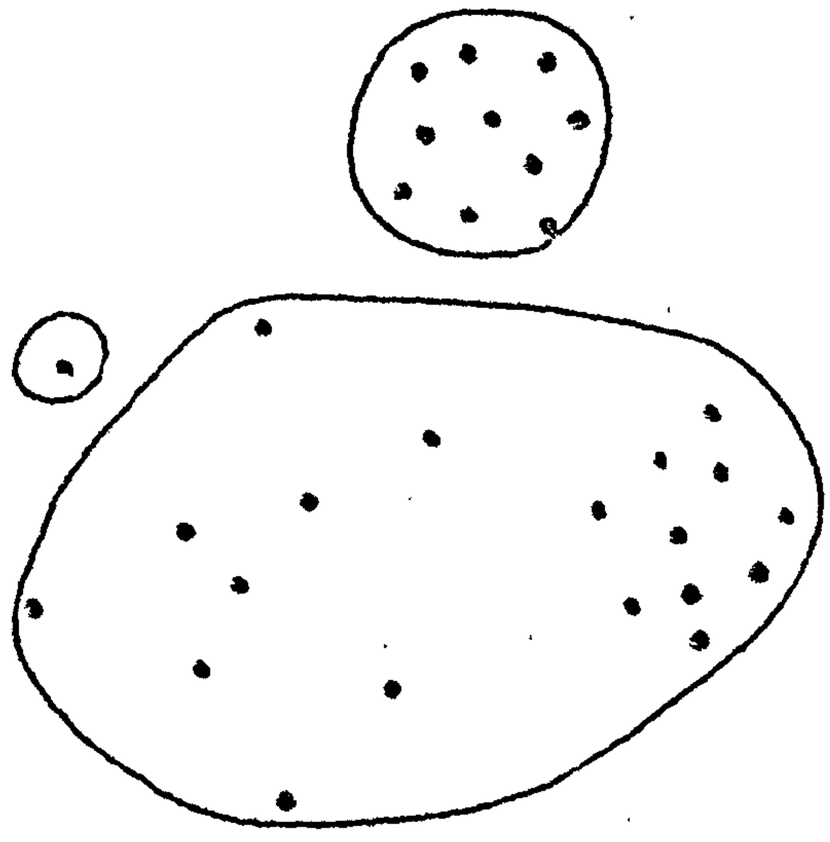
From these values it can be seen that the values of $\alpha = 0.6$, $\beta = -0.7$; $\alpha = 0.7$, $\beta = -1.1$; and $\alpha = 0.7$, $\beta = -1.2$ appear to fare best. These do better than any of the methods discussed so far on the round tests and only slightly worse on the straggly tests. These results show that further investigation of other values with lower β may be worthwhile.

So far we have mainly dealt with methods which have had limited success with our straggly group tests. We now examine our Klink method, which links a point to its nearest neighbour if the summed similarity to its nearest K neighbours is below a certain level. We also examine the Nucleus method which was also designed for all shape groups - here nearest neighbour groups are allowed to overlap by up to j objects before merging. The single link method is a limiting case of both methods and we will include it in our table of results.

METHOD	64 ROUND TESTS Level of success			32 STRAGGLY TESTS Level of success		
	1	2	3	1	2	3
KLINK K=4	40	16	12	15	3	2
KLINK K=3	44	24	11	20	3	2
KLINK K=2	53	17	16	25	9	7
NEAREST NBR	53	23	15	27	16	13
NUCLEUS K=1	47	20	7	20	8	5
NUCLEUS K=2	37	13	6	17	6	2
NUCLEUS K=3	23	17	7	12	5	0
NUCLEUS K=4	13	9	5	9	3	0

The general impression from the above table is that nearest neighbour performs better than the other methods, and that the two methods do best with the lower parameter values. Klink with K=2 performs almost as well as nearest neighbour. However the above results mask the fact that the Klink and Nucleus methods succeed in some cases where nearest neighbour fails, and in fact none of the eight methods in the table are dominated by any of the others. As the Nucleus method with K=3 and 4 performed very badly, even with straggly groups, these can be eliminated from detailed discussions.

Klink 2 has the most correct groupings after nearest neighbour. It succeeded on two round tests where nearest neighbour failed. One of these was test 27, where the failure was as follows:

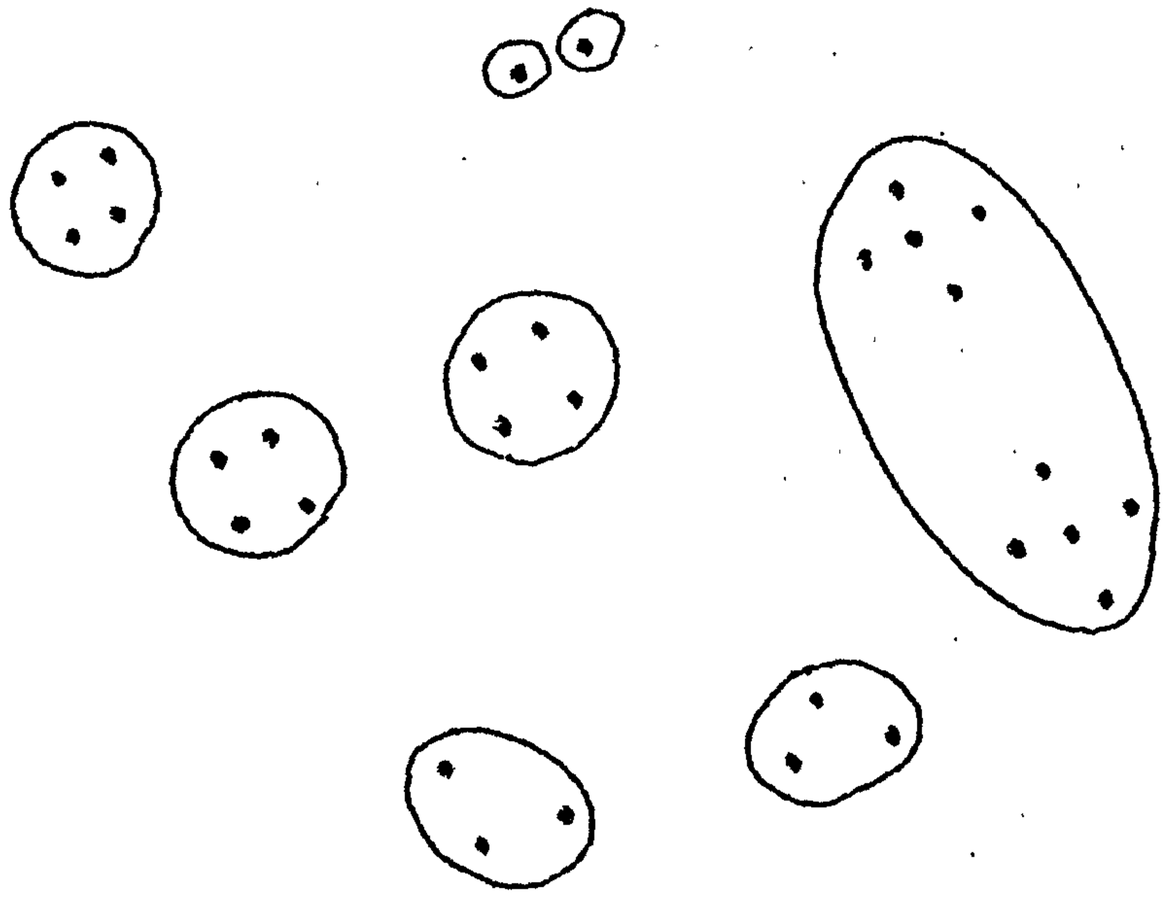


TEST 27

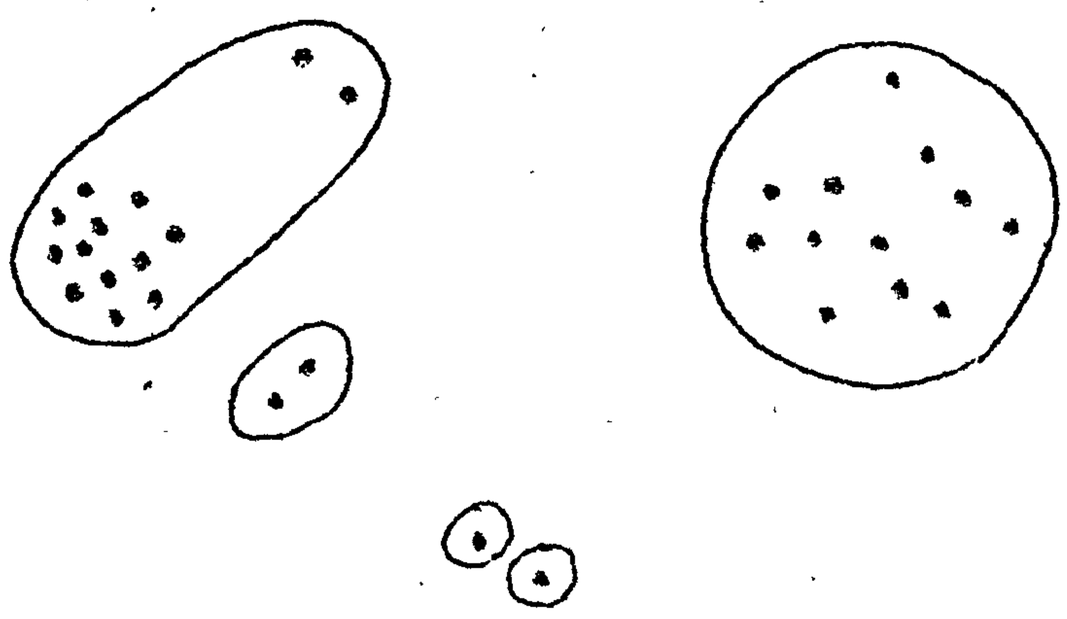
KLINK 2

The other case was very similar, and both cases where outliers were found when groups chained together earlier than they should. Thus Klink 2 has less of a tendency to chain. Nucleus 1 also succeeded in two cases where chaining had

caused nearest neighbour to fail, including test 27 above. However Klink 2 failed in two cases where nearest neighbour succeeded. All the other Klink and Nucleus methods also failed these. Here are the results with the Klink 2 method:



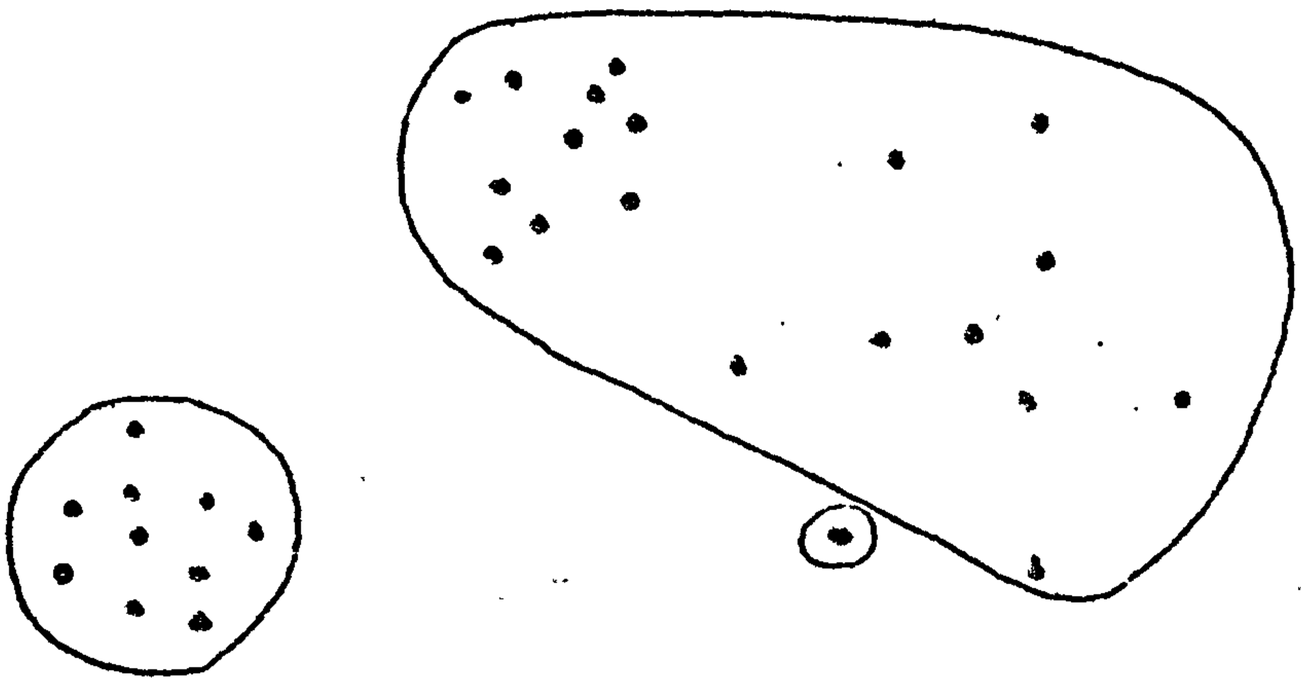
TEST 63
KLINK 2



TEST 60
KLINK 2

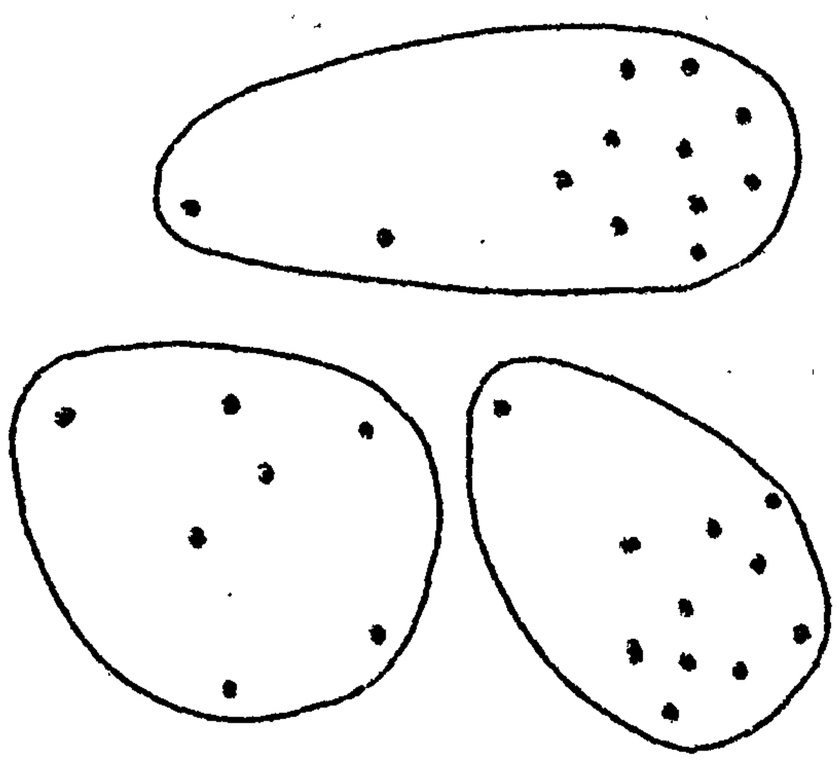
There is clearly difficulty in joining the two point groups. This is not surprising since the Klink method requires the similarity with k other points before nearest neighbours join, and the Nucleus method requires overlap of k points.

Klink 3 succeeds in two cases where Klink 2 fails. They are also both cases where nearest neighbour fails, and Klink 4 succeeds. We illustrate with one of these two - test 29:



TEST 29
KLINK 3,4

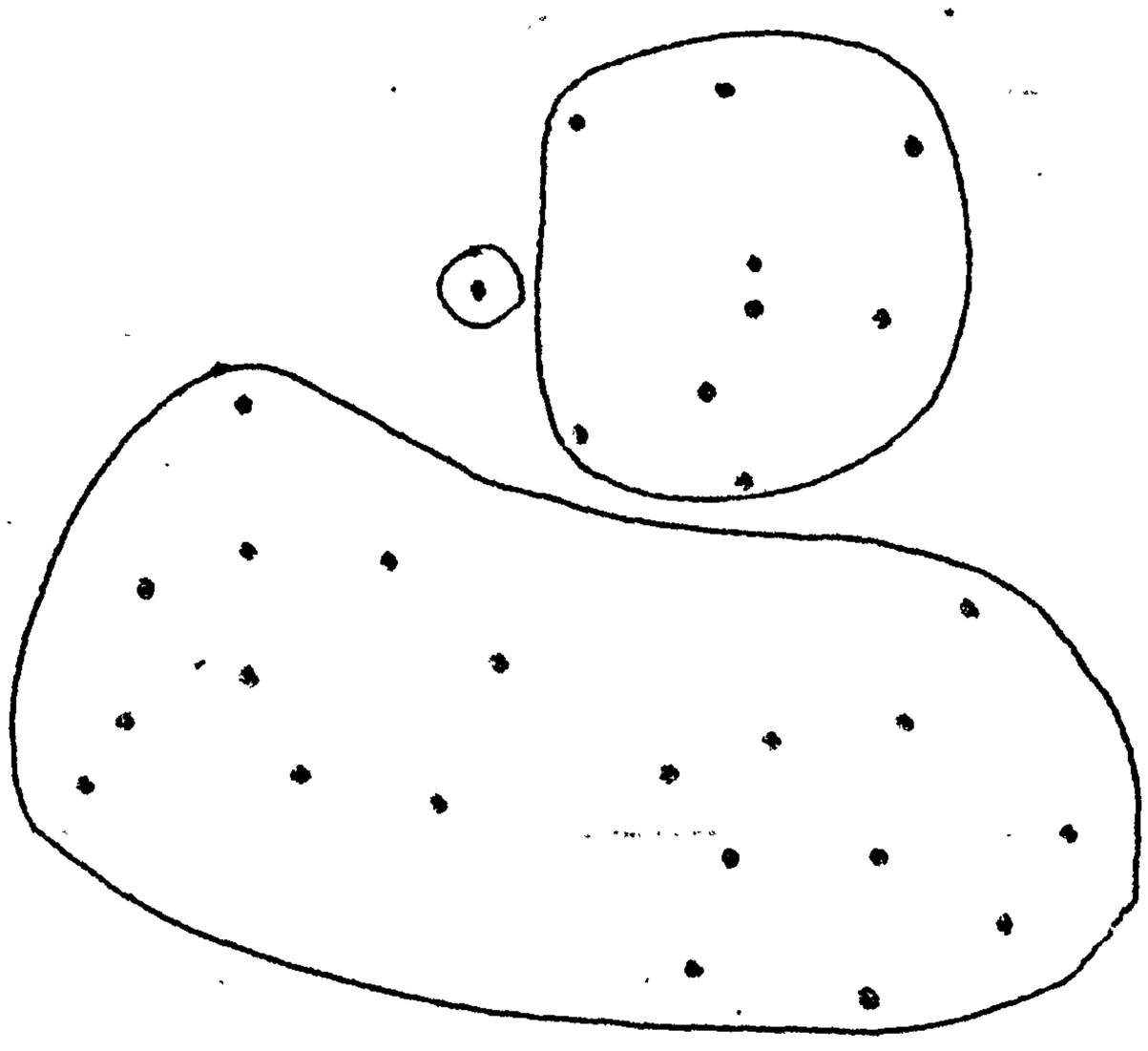
Here chaining is still present, thus with Klink, k is a parameter which allows a degree of chaining which decreases as k increases. Klink 3 and 4 failed in other tests where outliers or small groups were present - thus one penalty of decreasing chaining is the difficulty of detecting small groups. This accounts for all the failures of Klink 3 where Klink 2 succeeds, apart from test 27 which we have just discussed here the Klink 3 solution is:



TEST 27
KLINK 3

This is apparently caused by the different densities of the clusters, giving a high linkage value of outliers of the less dense group to the dense clusters. The additional failures of Klink 4 can be explained by these two difficulties outlined above.

We have covered the successes of Nucleus 1 over nearest neighbour already. The method failed more than Klink 2 with very small groups, and also two other tests where nearest neighbour and all the Klink methods succeeded. One of these is shown below - test 18:



TEST 18
NUCLEUS 1

The cause of this is similar to that which causes difficulty with small groups - outlying point will be emphasized. Nucleus 2 failed on more tests, but all for the reasons explained above.

With the straggly group tests a similar pattern emerged - the Klink method, however, was more sensitive to K, with a success rate falling from 78% with K=2, to 47% with K=4. On these tests Klink 3 dominated Klink 4. The reasons for failure were as in the round groups case, and the Klink method again tended to do better than Nucleus.

We can now consider the methods on the null hypotheses. Firstly supposing equal variance:

ROUND TESTS

		SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO SIGNIF.	WRONG CLUSTERS AND SIGNIF.
KLINK	K=4	16	10	19	14	5
	K=3	24	10	16	10	4
	K=2	17	18	6	18	5
NEAREST NBR		23	15	11	14	1
NUCLEUS	K=1	20	12	10	14	8
	K=2	13	7	14	17	13
	K=3	17	0	21	6	20
	K=4	9	0	38	4	13

STRAGGLY TESTS

KLINK	K=4	3	8	11	4	6
	K=3	3	7	7	10	5
	K=2	9	5	4	11	3
NEAREST NBR		16	8	4	3	1
NUCLEUS	K=1	8	6	7	6	5
	K=2	6	2	9	9	6
	K=3	5	2	14	5	6
	K=4	3	1	17	5	6

Nearest neighbour appears to be easily the best, and the Klink method does better than the Nucleus method. All of the methods have a large number of results in the two right hand columns, indicating the difficulty of a null hypothesis in such cases where only a few values from the similarity matrix determine the clustering at each stage.

We now move on to the full null hypothesis results.

ROUND TESTS

		SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO SIGNIF.	WRONG CLUSTERS AND SIGNIF.
KLINK	K=4	12	23	24	5	0
	K=3	11	31	20	2	0
	K=2	16	29	9	10	0
NEAREST NBR		15	32	11	5	1
NUCLEUS	K=1	7	33	13	6	5
	K=2	6	26	24	5	3
	K=3	7	14	35	2	6
	K=4	5	7	48	1	3

STRAGGLY TESTS

KLINK	K=4	2	13	15	1	1
	K=3	2	16	10	2	2
	K=2	7	12	5	6	2
NEAREST NBR		13	11	5	3	0
NUCLEUS	K=1	5	11	11	3	1
	K=2	2	13	14	2	1
	K=3	0	11	20	1	0
	K=4	0	8	23	1	0

Nearest neighbour again appears best, but is matched by Klink 2, at least on round groups. The Klink methods again do better than the Nucleus, but on straggly groups only Klink 2 and Nucleus 1 performed anywhere near as good as nearest neighbour.

Although we have discussed the parameters in these cases by their numerical values, they are related to the number of objects under study. Another point worth mentioning is that although the Nucleus method did not perform as well as the others, it has also the provision for overlapping groups.

We can summarize our conclusions with these methods as follows:

1. The methods Klink 2 and 3 and Nucleus 1 all perform well on straggly groups, but in general not as well as nearest neighbour.

2. The methods reduce the adverse chaining effects of nearest neighbour, at the expense of difficulty in finding small groups, and in cases where groups are of very different densities.

3. In general the Klink method outperforms Nucleus, and in particular Klink 2 was the best method tested, although this value may increase slightly with n.

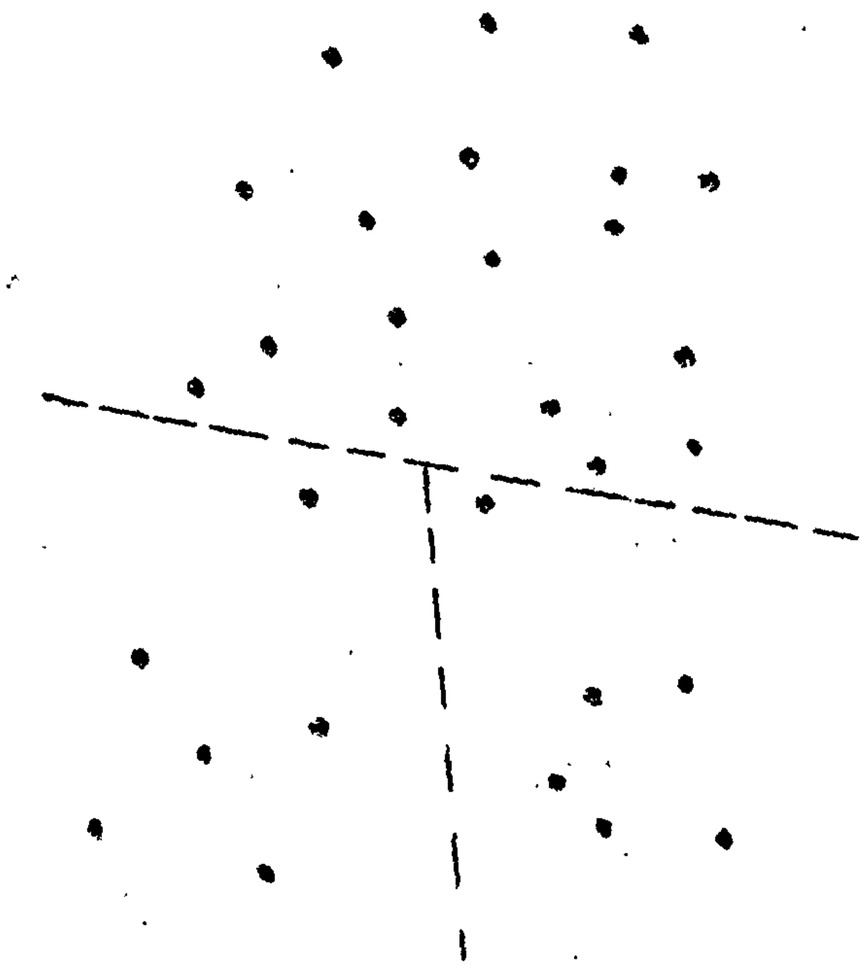
Dissimilarity analysis was the only divisive method to be fully analysed, and is the last hierarchic method for discussion. The method succeeded on 52 of the round tests and 7 of the straggly tests, which means it was outperformed by most of the methods discussed so far. It did, however,

have success where other methods failed. Divisive methods also have the advantage that they produce solutions with a small number of groups, early in the analysis. These are normally those in which one is primarily interested, they can thus be terminated at a particular level, and can hence be very fast.

The method correctly processed test 44 which all the hierarchic round group methods failed, and was successful on several (e.g. tests 3, 23, 55) which all the straggly group methods discussed so far failed. From examination of these tests it can be seen that dissimilarity analysis does not have the failings where groups are of different densities.

The method failed on several tests which had been passed by most of the other methods. We illustrate with tests 15 and 33:

TEST 15 DISSIMILARITY ANALYSIS



TEST 33 DISSIMILARITY ANALYSIS

The problem occurs with groups of large area, which have points at a large distance from their centre, and are assessed as more similar to a smaller group.

No satisfactory null hypothesis could be created for dissimilarity analysis. The problem is not as straightforward as that of agglomerative methods since operations are executed on separate groups of points. Several criteria were investigated to see if they were satisfactory rules for distinguishing between grouped and random data.

Suppose a group splits into two parts with objective function equal to a and these two groups split at levels a_1 and a_2 then for a case where two groups are present, the following parameters were investigated:

$$k_1 = \frac{a}{\max(a_1, a_2)}$$

$$k_2 = \frac{a}{a_1 a_2}$$

However even in very obviously clustered cases there was little difference from the random tests.

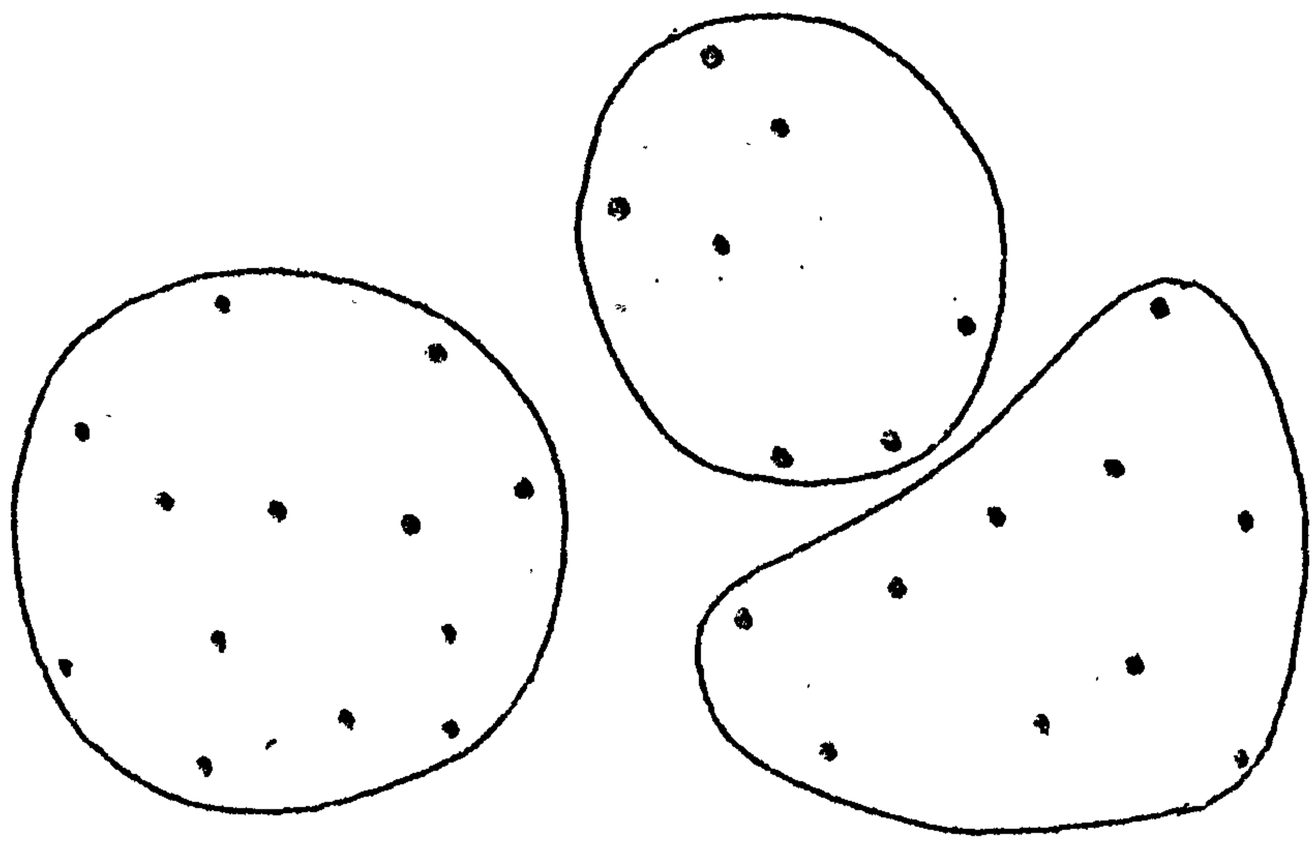
There were three iterative relocation methods in our investigation - Beale's method, group average relocation, and the neighbourhood method. The results were as follows:

	64 ROUND TESTS Level of success			32 STRAGGLY TESTS Level of success		
	1	2	3	1	2	3
Beale's	59	49	34	11	1	1
Group Av. Reloc.	46	36	31	12	3	1

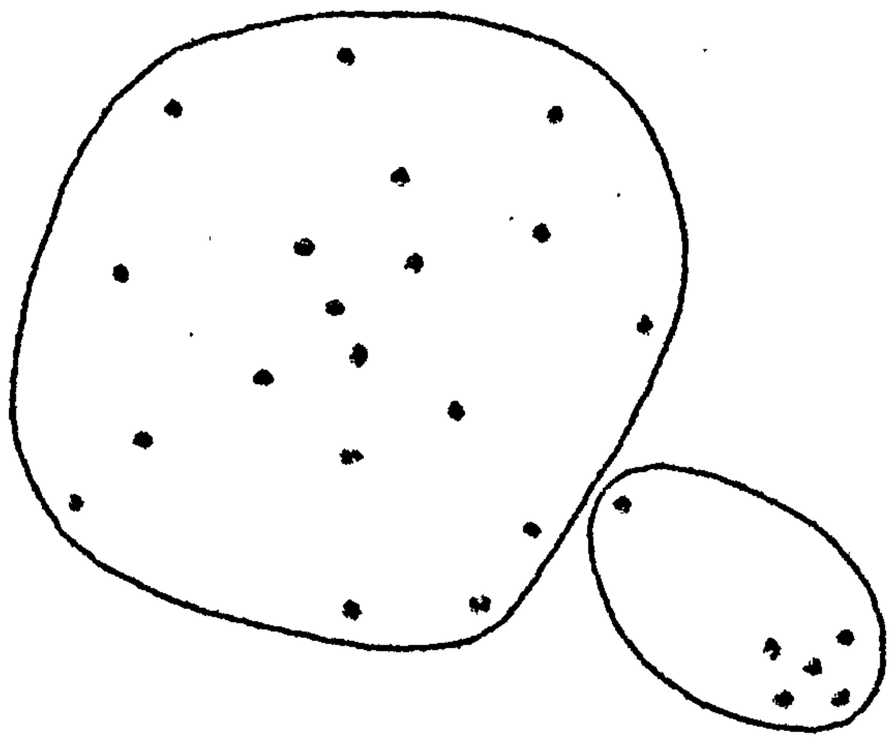
The neighbourhood method null hypotheses were unsatisfactory, only one result was found to be significant. The method, however, was very successful at finding the groups:

	K = 1	K = 2	K = 3	K = 4	K = 5
64 Round Tests	63	63	64	63	63
32 Straggly Tests	13	14	14	13	14

Beale's method was surprisingly dominated by Ward's method on the round group tests. There were two cases where Ward's method succeeded and Beale's failed. These were:



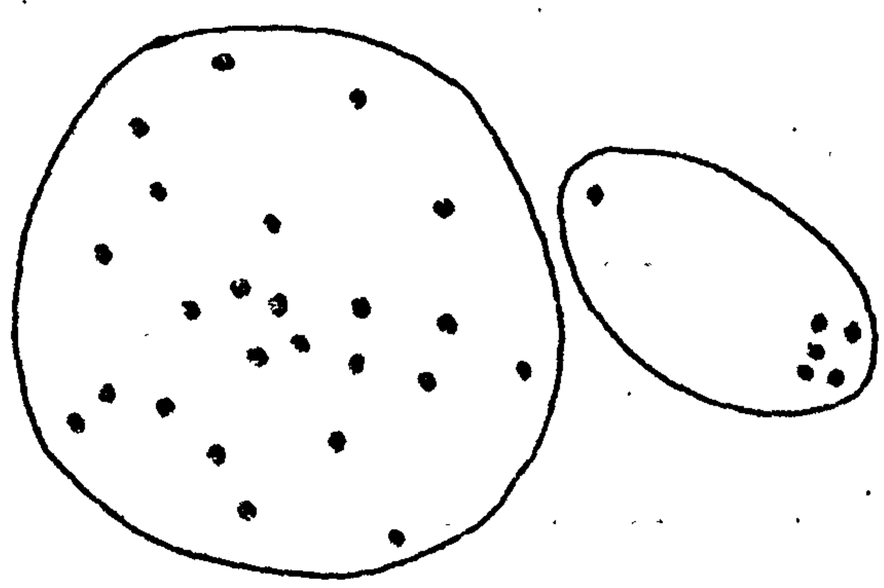
TEST 49
BEALES



TEST 43
BEALES

In both these cases the objective function is higher than for the expected groupings - this is a defect of the within-group error sum of squares as an objective function. This defect occurs when small groups exist.

The group average relocation method performed badly with round groups - only achieving about a 70% success rate. The fault lies with the group average criterion. A typical result is that of test 15:



TEST 15
GROUP AVERAGE
RELOCATION

Here the optimum for the particular objective function has been found, but it is clearly not the 'best' result.

The results with the null hypotheses assuming dimensions of equal variances were as follows:

ROUND TESTS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
Beale's	49	6	0	4	5
G.A.R.	36	8	7	3	10

STRAGGLY TESTS

Beale's	1	5	12	5	9
G.A.R.	3	5	12	4	8

The results of Beale's method are similar to those with the best hierarchical methods. The corresponding results with the full null hypotheses are:

ROUND TESTS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
Beale's	34	25	4	0	1
G.A.R.	31	15	12	0	6

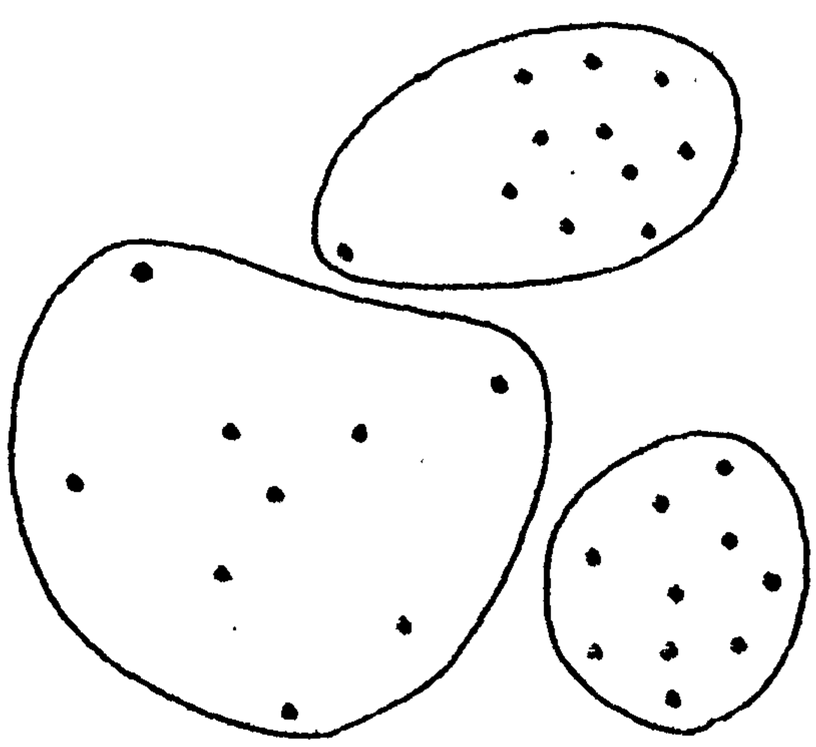
STRAGGLY TESTS

Beale's	1	9	20	1	1
G.A.R.	1	8	18	3	2

Beale's method, on the above results, dominates all but the extended flexible method on round groups, and is less liable to give the wrong result on all 96 tests, than any method.

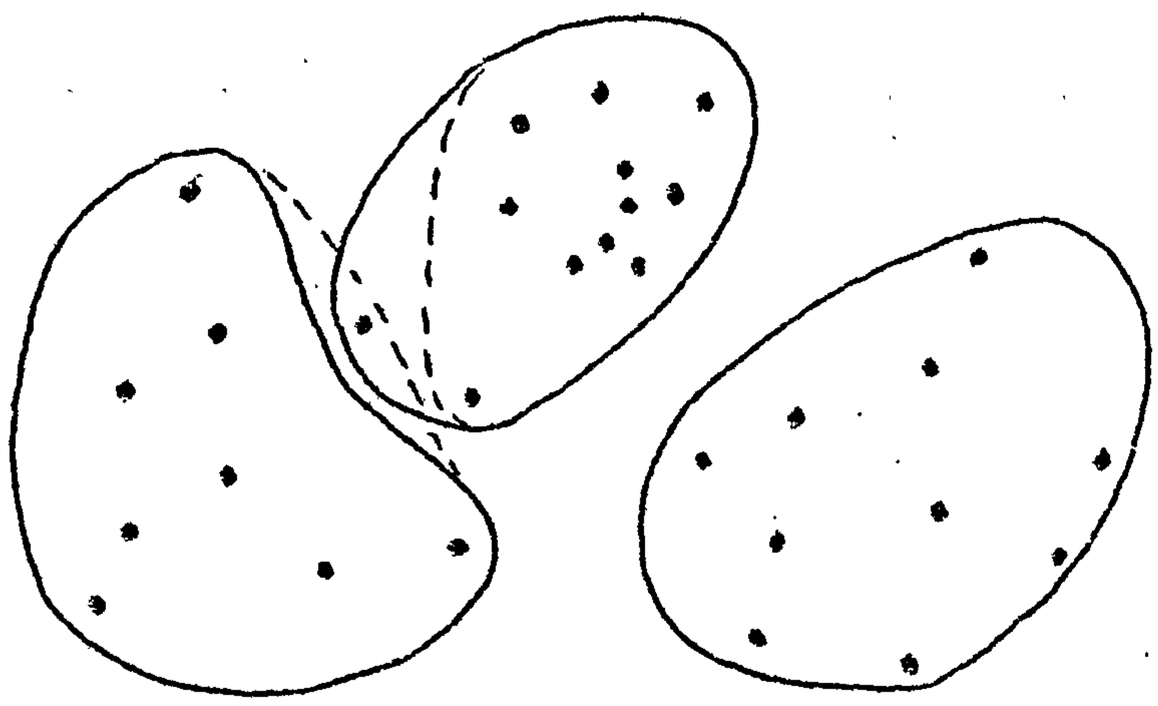
The group average relocation method also did well on these results. Thus we can suggest that relocation methods do well under such a null hypothesis. This is probably due to the fact that a global optimum is more often found, and one is comparing optima, and not results from a series of optimum moves which are often non-global optima, as in hierarchical methods.

The neighbourhood method had surprisingly low success with straggly groups, although excellent results were achieved with round groups. With K=3 there were no failures, and with the other parameters one failure each. The failure with K=1 was as follows:



TEST 28
NEIGHBOURHOOD
K=1

And those with K=2, 4, 5 were:



TEST 23
NEIGHBOURHOOD
——K=2
- - - K=4,5

It can be seen that the method has a slight tendency to 'lose' outlying members of large groups to denser groups nearby. On straggly groups the method had success generally in the same cases as the hierarchical methods for round groups.

Unfortunately, as stated earlier the null hypotheses had negligible discriminating power.

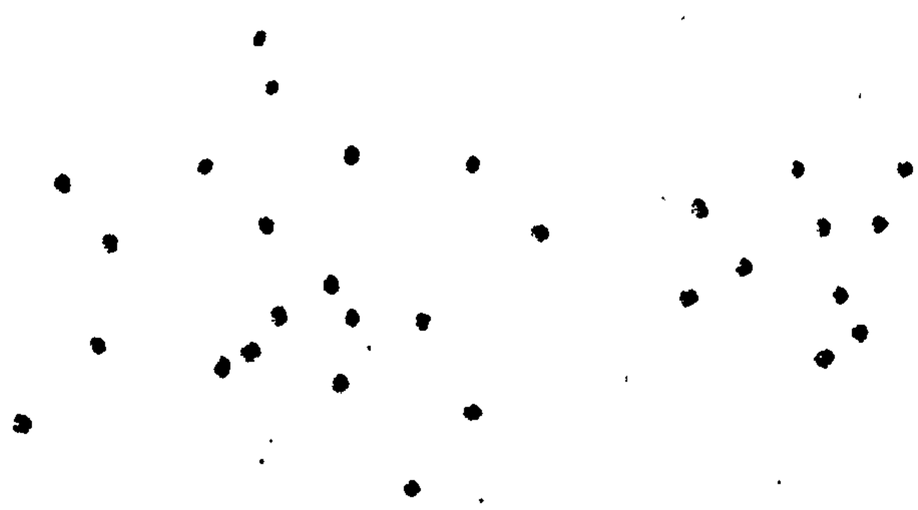
Mode analysis is a method designed for groups of any shape, and it achieved good results with these tests. We give the results, together with nearest neighbour and Klink 2 - the best two methods for straggly groups analysed so far. (Nearest neighbour is a limiting case of mode analysis.)

	64 ROUND TESTS	32 STRAGGLY TESTS
KLINK 2	53	25
NEAREST NEIGHBOUR	53	27
MODE K=1	57	25
K=2	51	21
K=3	35	16
K=4	22	11
K=5	10	8
K=6	5	7

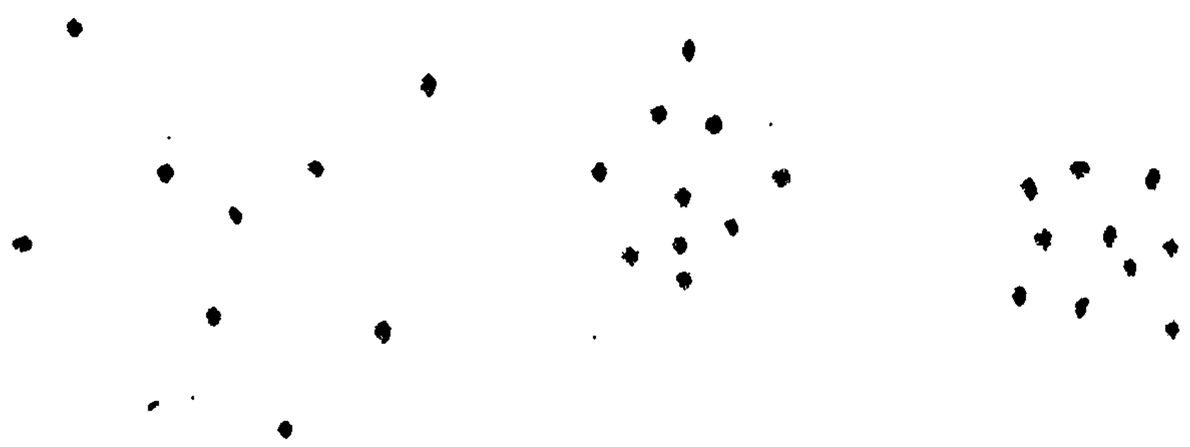
It can be seen that there is a sharp drop in achievement after Mode 2, and that Mode 1 and 2 do about as well as the Nearest Neighbour and Klink 2 methods. On the round group tests the mode results dominate each other in the order of K, i.e. Mode 1 dominates Mode 2, etc. With the other test set, the same order relations were true except for three tests - one where only Mode 4 succeeded, one where 4 and 5 succeeded, and one which only 1 and 6 succeeded on.

All the Mode methods failed on tests 60, 63 and 64 which all contained outliers. However the method had less difficulties with small groups than the Klink and Nucleus

methods, and with Mode 1 has slightly less trouble with groups of different density. There were however a few cases that were correctly grouped by most of the Klink and Nucleus methods but on which the Mode methods (2-6) all failed. Here are two examples:

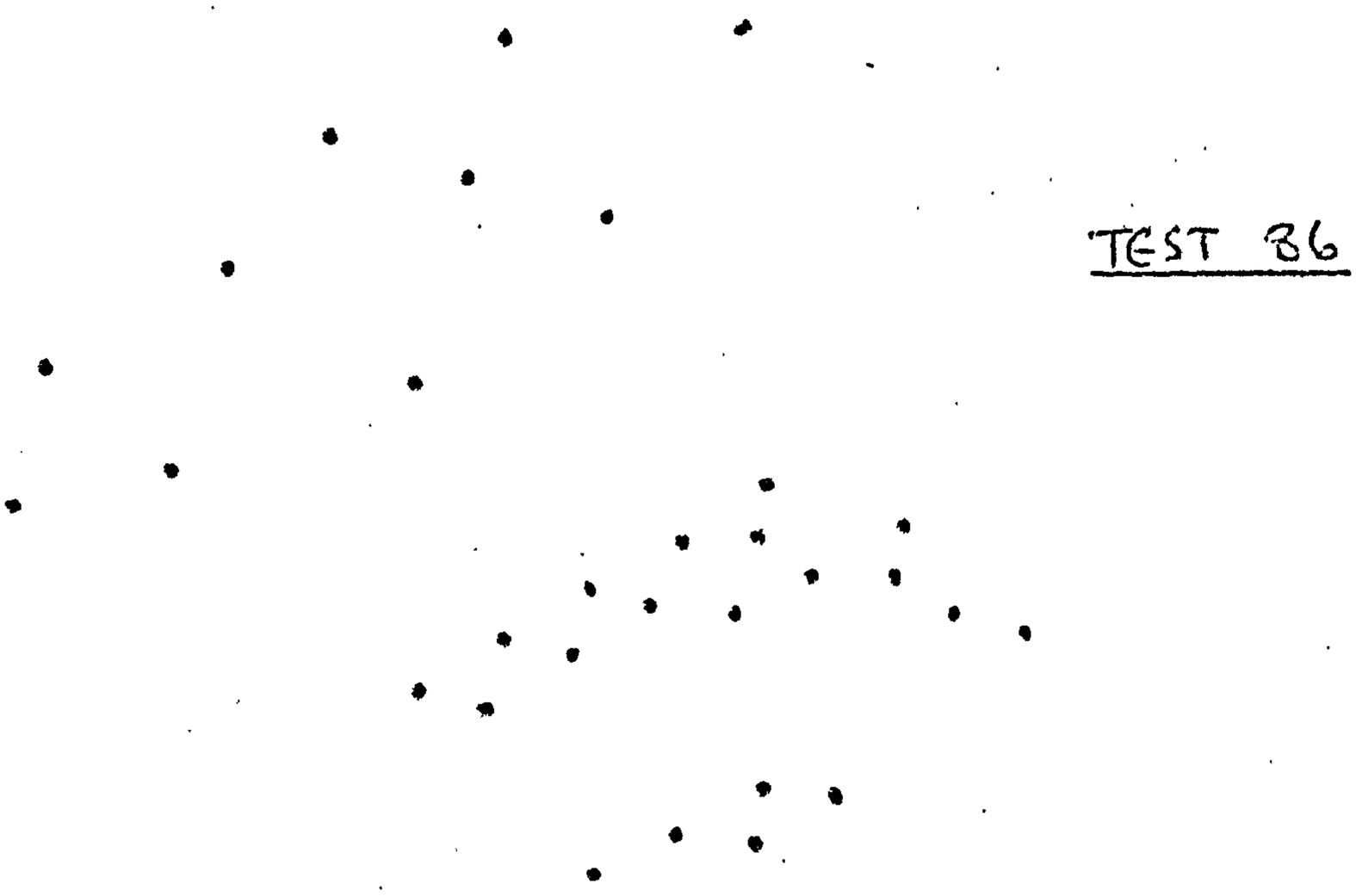


TEST 12



TEST 31

In the first example (test 12) Mode 4, 5 and 6 produced no groupings at all, and with Mode 2 and 3 the two groups formed, but joined before the outliers members had joined their groups. In the second example the two dense groups join before the other group is properly formed. Mode 1 correctly grouped the above tests, and on the straggly tests succeeded on a test which no other method succeeded on - test 86:



TEST 36

The mode method also could not identify outliers with straggly groups.

A difference between Mode analysis and other methods is that the method can output the result that the data is unimodal. Thus one would like random data to give this result, so that if the method gave groups, this would mean they were statistically significant.

With K=6 none of the sixty random tests were significant. With K=5, three of the random tests produced groupings - thus if any data of 30 objects produced groupings with K=5 we could be 95% certain that they were significant. With K=4, nine tests produced groups - this was thought to be too many in which to base a reasonable test. The groupings that were produced were examined and it was found that often the grouping came about because a single point became a separate group at level a_1 which very soon (say at level a_2) joined

the other group. Thus from investigating these groupings, the null hypothesis that no group exists for which $a_2/a_1 > 1.1$, was used. With this, only four of the random tests were significant - the equivalent of a $93\frac{1}{3}\%$ level significance test.

With K=3, 24 of the 60 tests produced groupings, and 22 were significant under the above rule. By further investigation the following null hypothesis was arrived at - that no grouping exists for which $a_2/a_1 > 1.2$, at a stage in the analysis where over 80% of the objects belong to modes. This produces a fairly weak hypothesis with 8 of the random tests significant. The same hypothesis with K=2 gives 29 tests significant - no major improvement was found which would give a good hypothesis.

It could be argued that one should only use random unimodal data for such a test, in which case less random groups are significant, but this would depend on the exact null hypothesis required for a particular investigation. The above figures hardly varied at all between random tests with dimensions of equal variances, and the other tests.

The results under these tests were as follows:

ROUND TESTS

	SUCCESS	RIGHT CLUSTERS NOT SIGNIF.	WRONG CLUSTERS NOT SIGNIF.	RIGHT CLUSTERS WRONG NO. SIGNIF.	WRONG CLUSTERS AND SIGNIF.
MODE 3	34	1	10	0	19
4	21	1	16	0	26
5	10	0	26	0	28
6	5	0	35	0	24

STRAGGLY TESTS

MODE 3	14	2	6	0	10
4	11	0	12	0	9
5	7	1	20	0	4
6	7	0	21	0	4

The results are not too encouraging, and in some cases more incorrect groupings are found significant than correct groupings.

Our condensation model method had three parameters, to which we gave two or three set values.

The method involves gravitating all points together slowly, gradually forming clusters. When points are within a distance d of each other they are amalgamated. S was a measure of the speed with which the points were allowed to move. c was a parameter which damped motion between close points to avoid overshooting.

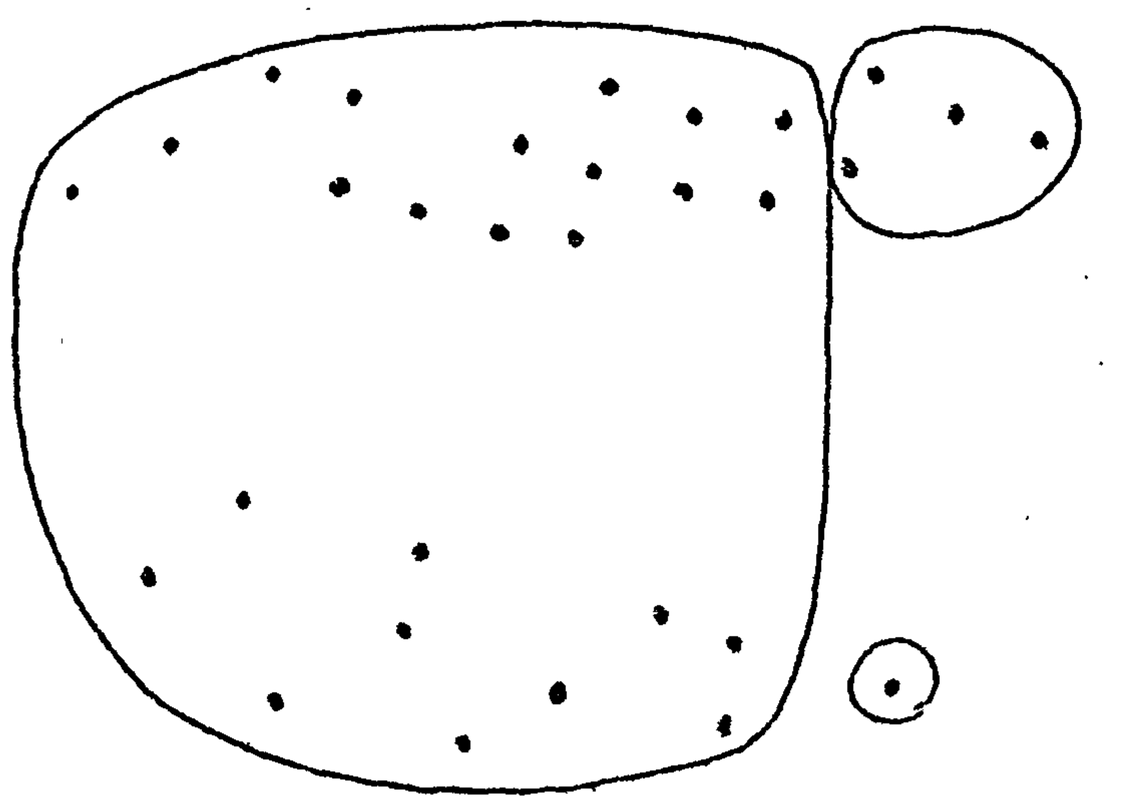
The results with the method were very encouraging as shown below:

			64 ROUND TESTS			32 STRAGGLY TESTS		
			Level of success			Level of success		
d	S	c	1	2	3	1	2	3
0.04	0.004	0.2	64	25	23	15	4	3
		0.3	63	12	12	11	1	1
		0.4	56	11	10	10	0	0
	0.006	0.2	64	26	16	15	5	4
		0.3	63	15	15	11	1	1
		0.4	57	12	12	10	0	0
0.06	0.004	0.2	64	30	23	13	4	4
		0.3	63	16	16	11	1	1
		0.4	59	7	7	11	1	1
	0.006	0.2	64	19	17	14	5	3
		0.3	62	20	16	11	2	1
		0.4	59	8	8	11	0	0
0.08	0.004	0.2	64	32	30	13	1	1
		0.3	64	22	22	11	1	1
		0.4	60	9	9	12	1	1
	0.006	0.2	64	32	31	14	4	2
		0.3	63	24	24	11	1	1
		0.4	60	10	10	11	1	1

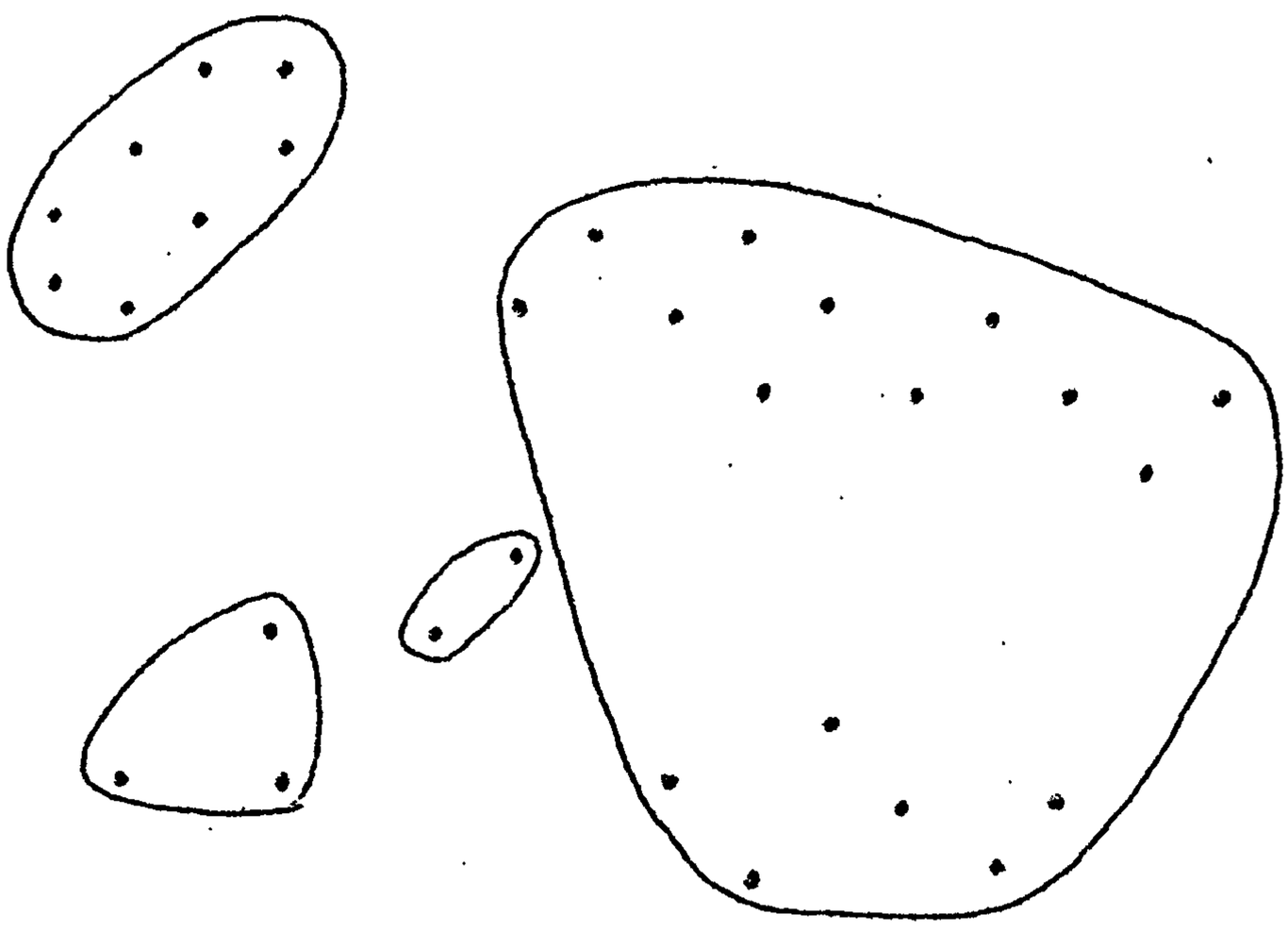
The parameter c appears to be the most sensitive on the above values and c=0.2 dominates nearly all the results with c=0.3 or c=0.4, and all the c=0.2 values achieve maximum success with the round groups. The results under changes in d and S are less sensitive, but more contradictory between the round and straggly tests. The round tests are performed better by higher values of d and lower values of S, but the straggly tests are passed by lower values of d and higher S. The results with the null hypotheses are as good as the better hierarchical methods, but there is some evidence that

the first differences of the objective function may form a better hypothesis with this method.

With c=0.2 the methods succeeded on all the round tests, so we look at failures on the straggly tests. The method tended to succeed on those tests which the hierarchical methods successfully completed. However two tests - 102 and 106 which most hierarchical methods succeeded on, gave the following results:



TEST 102
CONDENSATION



TEST 106
CONDENSATION

The groups are shown for d=0.04, S=0.004 and c=0.2, but are the same as those found by most of the other parameter values. Test 66 shows that the method has difficulty with long parallel groups, and the same effect is present in test 70 where two long groups have prematurely joined. The results are encouraging however, and possibly by the use of a more exact analogy to the motion of masses in a viscous medium, might yield better results.

The results of most of the methods on each test are shown in Tables 6, 7 and 8. The conclusions from these results, consideration of the computer times for each method, and our results in the light of other researches, will be discussed in the next section.

C.6 CONCLUSIONS

General Conclusions

Here we shall first list the conclusions drawn from our analysis of the results which we have discussed in the preceding section.

1. Nearest neighbour can find groups of any shape, but has difficulty where 'freak' points exist between clusters and where groups are of different densities. Nearest neighbour is particularly good at identifying outliers.

2. Furthest neighbour is inferior to median, centroid, group average, weighted average and Ward's method, having a tendency to form round groups of equal area.

3. Group average is generally more inaccurate than centroid. The cause of this inaccuracy is the replacing of a group by a single point at the centre, which means that groups covering large areas tend to 'lose' points to groups of smaller area. The median method had very similar results.

4. Weighted average had the difficulty that groups are given the weight of a single point, and so the centre of the cluster tends to drift away from the densest part of the cluster. This gives a tendency for forming clusters of equal numbers of points.

5. Of the methods analysed which form dendrograms, centroid and Ward's were the best methods with round groups. They were also able to find more straggly groups than expected. Centroid had some difficulty when groups of widely differing density were present, and Ward's method had difficulty identifying outliers.

6. Values of α and β with the extended flexible method were affected by chaining and pairing in some cases. Several parameter values correctly grouped all the 64 round tests, and had the highest success of any method with the full null hypothesis. With straggly groups, slightly less success was achieved than in the methods discussed so far. The best parameter values were $\alpha = 0.6$, $\beta = -0.7$; $\alpha = 0.7$, $\beta = -1.1$ and $\alpha = 0.7$, $\beta = -1.2$.

7. The Klink and Nucleus methods both reduced the chaining effects of nearest neighbour, but had difficulty identifying small groups and outliers. Klink was generally better than Nucleus, and Klink 2 was the best method.

8. Dissimilarity analysis did not do well. It had successes in cases where groups of differing densities were present, where other methods had difficulty, but had a tendency to allocate outlying members of large clusters to smaller nearby clusters.

9. Beale's method was surprisingly not as successful as Ward's; this is due to the use of error sum of squares as an objective function, which can misallocate peripheral points of large clusters to nearby small clusters. Beale's method was, however, more successful than any of the dendrogram forming methods under the full null hypothesis.

10. Group average relocation failed because of the use of group average as an objective function, the method had bad performance, although with the full null hypothesis, it worked well.

11. The neighbourhood method was very successful with round groups, and as good as the best round group method with straggly groups.

12. Mode analysis had similar failings to Klink and Nucleus, but to a lesser degree. Mode 1 was the best method, and one which outperformed nearest neighbour.

13. The condensation model method worked very well on round groups and as well as the best hierarchical methods on straggly groups. There was some difficulty with elongated parallel groups. The parameter c=0.2 was better than 0.3 or 0.4, but the results on the other two parameters were difficult to interpret.

14. Of all the methods which form a dendrogram the only method to succeed on all the round tests was the condensation method. The other methods which succeeded on all these were the extended flexible, and the neighbourhood method.

15. Nearest neighbour succeeded most of the 32 straggly tests, then came Mode analysis and Klink.

16. With the null hypotheses tests, the extended flexible method performed best, with Beale's, Ward's and the Centroid methods next best.

17. Among the methods discussed there is a clear distinction between methods which search specifically for round groups and those which seek 'natural' clusters.

Comparison with other Researches

How do our results compare with other people's? From the chart on page 220 one can see that our results are at variance with the few studies that have been previously attempted. The main difference between our results and the other studies is the high position in our tests of centroid,

and the low position of furthest neighbour. Cunningham (1972) and Strauss (1971) are the previous two major works in this field. Cunningham's study was on six data sets (four of which contained a large number of ties), and was a comparison of hierarchies and the original data, which may well give different results from our experiments which have been tests of clustering, rather than hierarchic representation. Strauss' results show greater correspondence with our results, but as shown by our tests, there are only slight differences in performance between the methods Strauss used, and with only a few tests his results could easily show a different ordering.

Our results coincide with those of Sneath in the table on page 228 - results which are based on long experience with the four methods he discusses.

Computer Considerations

The computer times for the methods depend on the algorithm used, and the expertness of the programming. Having no claim to such expertise, general conclusions will be drawn as to the time taken for each method, and no actual times will be compared.

Dissimilarity analysis was probably the fastest method in our study, and had the advantage that one could stop groups below a certain size being further divided and so shorten the computation further.

All the methods which can be performed by Lance and Williams' algorithm are very fast, and all take approximately the same time (but with those dependent on group size -

group average, centroid and Ward's method having very slightly slower times). Nearest neighbour and furthest neighbour could be executed faster by simplification of this algorithm, and nearest neighbour can be speeded up even more by a different algorithm. Both these neighbour methods can be performed manually fairly easily up to about 50 objects.

The condensation method was the next fastest, due to the way in which a large number of points merge at early stages in the algorithm, and so the effective data set size is reduced for most of the calculation.

The group average relocation method and Beale's method are probably the next fastest, although much slower than the methods above - due to the search for improvement at each stage. The methods have the advantage that if one begins with a random or stratified grouping, after a few fusions and their associate reallocations, the grouping would be near optimal. This obviously reduces computer time, but with a greater chance of error. A more accurate approach would be to 'switch off' the relocation procedure (in which case the methods revert to group average and Ward's, respectively) until the number of groups has been reduced.

The other four methods are based on neighbourhoods and not on similarity to a group centre, and hence take longer computationally - in the probable order of Klink, mode analysis, Nucleus method, neighbourhood method, with the fastest first.

The computer storage requirements of cluster analysis have been discussed in general terms earlier. Most of the above methods can be programmed to store either the data matrix, or the lower off-diagonal distance matrix, whichever is smaller. Nearest neighbour has the advantage that it can be programmed with much less storage than this (see the algorithm by Sibson 1973).

Decision Rules for Choice of Method

No one method dominated all others, and so the decision of what method to employ in a specific case cannot be simply answered. The most important consideration in deciding the approach to be taken is the problem itself, including the data which is available, and the type of question one is investigating. This may preclude the use of certain methods.

In order to facilitate our discussion we have tabulated (Table 19) the results of the best methods on both sets of tests at the first and third success levels (i.e. the groups found, and the groups significant with the full null hypothesis). An indication is also given of the computer times involved. The ordering of the methods represents little but the author's prejudices. (The values of the best point on Lance and Williams' flexible line is included along with their recommended point, for completeness).

RESULTS

Method	64 Round Tests		32 Straggly Tests		Computer Time
	Groups found	Groups signif.	Groups found	Groups signif.	
α=0.6, β=-0.7	62	45	8	1	LOW
α=0.7, β=-1.1	64	45	9	0	LOW
α=0.7, β=-1.2	64	43	9	0	LOW
d=0.04, S=0.004, c=0.2	64	23	15	3	LOW
d=0.08, S=0.006, c=0.2	64	31	14	2	LOW
NEIGHBOURHOOD, K=3	64	*	14	*	HIGH
NEIGHBOURHOOD, K=2, 5	63	*	14	*	HIGH
NEIGHBOURHOOD, K=1, 4	63	*	13	*	HIGH
BEALE'S	59	34	11	1	MEDIUM
WARD'S	61	33	12	0	LOW
CENTROID	61	32	14	1	LOW
FLEXIBLE (β=-0.15)	61	25	10	1	LOW
MODE, K=1	57	*	25	*	HIGH
KLINK, K=2	53	16	25	7	HIGH
MEDIAN	58	30	10	1	LOW
WEIGHTED AVERAGE	58	29	11	0	LOW
FLEXIBLE (β=-0.25)	59	20	11	1	LOW
GROUP AVERAGE	57	25	12	0	LOW
NEAREST NEIGHBOUR	53	15	29	13	LOW
MODE, K=2	51	*	21	*	HIGH
NUCLEUS, K=1	47	7	20	5	HIGH
DISSIMILARITY ANALYSIS	53	*	7	*	LOW
FURTHEST NEIGHBOUR	50	19	5	0	LOW
GROUP AV. RELOCATION	46	31	12	1	MEDIUM
KLINK, K=3	44	11	20	2	HIGH

* indicates that no satisfactory null hypothesis was found

TABLE 19

A strategy which can be applied to most problems is an initial nearest neighbour clustering to identify possible outliers. These can be eliminated from the data set, and considered independently. There are some instances where this would not be applicable, for example if groups had to be of a minimum size, but the approach is normally satisfactory even in these cases as useful additional knowledge. If outliers are removed from the data then certain methods become more accurate - these are Ward's, Beale's, and the neighbourhood type methods - Nucleus, Neighbourhood Method, Klink and Mode Analysis. Normally, the type of dendrogram produced by nearest neighbour, especially in the latter stages is a set of individual points gradually joining a large group. This is a normal type of dendrogram produced by excessive chaining. It is possible however that more than one group may persist until higher levels in the dendrogram - this would indicate clearly separate groups. The data can then be divided into these groups and each analysed separately.

An important consideration in considering what type of information we wish to extract from our data is whether a dendrogram is required. It may for example be necessary in biology or botany where clustering is required at several levels - subspecies, species, genus, etc. The question of how good a dendrogram represents data is somewhat different from how good clusters represent certain data, but the indications from our studies are that Ward's method is probably better than the other methods which can be performed

by Lance and Williams' algorithm. Centroid is as good as Ward's method, but can give rise to reversals, which may or may not be acceptable in a particular study - this is a question for user control. Other methods which can produce dendrograms, and which are worth consideration, are the condensation method, and Klink, with a value of about $\frac{n}{15}$, where n is the number of objects under investigation.

If there is restriction on the amount of computer time available then Klink cannot be considered. Another consideration is the difficulty of round and straggly groups. Groups may be desired to have properties such as high similarity with all other members of the group, in which case we need only consider round groups, and use either Ward's method or the condensation method (or both). There may also be other factors in the data or required result which preclude methods for straggly groups. Also if one has strong reasons for expecting, or requiring, groups which may not be roughly hyperspherical, then one is limited to the Klink method, which may be precluded by computer cost - in which case one is left to intuitive reasoning with the nearest neighbour method (and possibly in conjunction with the Ward's method or condensation method solution).

If one has no specific reason for wanting or suspecting groups to be any particular shape, or one is using cluster analysis purely for exploratory purposes, and if computer time were not restricted, then the use of all three methods mentioned above, coupled with interpretation by someone with

knowledge of the data, would be the recommended approach. If computer time was limited then one could either rely on Ward's or condensation alone, remembering their fair success with straggly groups, or use them in conjunction with the nearest neighbour solution.

The use of prior knowledge, replicating the data set, obtaining other data, splitting the data in two halves, and other tests for robustness of clusters will be of obvious advantage. The construction of null hypotheses can be a useful approach, but as can be seen in the table, is mainly of assistance with confirming the existence of round groups, although as shown in our results significance tests with nearest neighbour may be of some value.

If a dendrogram is not required then all the methods we have discussed are open to us, but there may be certain instances which necessitate the use of another particular subset of methods. If we have an open choice then the set of methods recommended, covering all types which may be required by consideration of group shape or computation time, are - Extended flexible, Condensation, Neighbourhood, Beale's, Ward's, Centroid, Mode, Klink and Nearest Neighbour.

Our main differentiation between these nine methods are on consideration of the group shape and time taken. If we consider that the round group methods are necessary or sufficiently good enough, then we have a choice of the first six methods listed above. Of these only the neighbourhood method is precluded if computation limits exist. This

leaves five methods, of which extended flexible is probably best if one intends to use the null hypothesis, and condensation if one wishes to have a higher chance of finding groups of any shape. Most of these methods are fairly fast, and if it were not too expensive then a good strategy would be to use two or three methods of different types - for instance Beale's, extended flexible and condensation. Similarly if the neighbourhood method is not precluded then this can also be used in conjunction with other methods.

If we need a method which will detect straggly groups accurately then we can use Mode analysis or Klink. It would be preferable to use both, but if only one could be used then Mode has the advantage of slightly better accuracy, while Klink has a null hypothesis and is a little faster. If these methods are too slow and expensive, one is left to an interpretation of a nearest neighbour solution in the light of a round group methods result (condensation or centroid do best on straggly groups).

If the shape of groups present in the data is unsuspected, and unrestricted by the problem, then we would select a straggly group method and a round group method from the list of nine methods above and interpret them together. If we were limited to the faster methods we would have to use one of the round group methods which was best at straggly groups - condensation, Ward or centroid and interpret them with the nearest neighbour solution.

The decision tree of the above discussion is shown in Figure 29. It is important however to see this tree as a

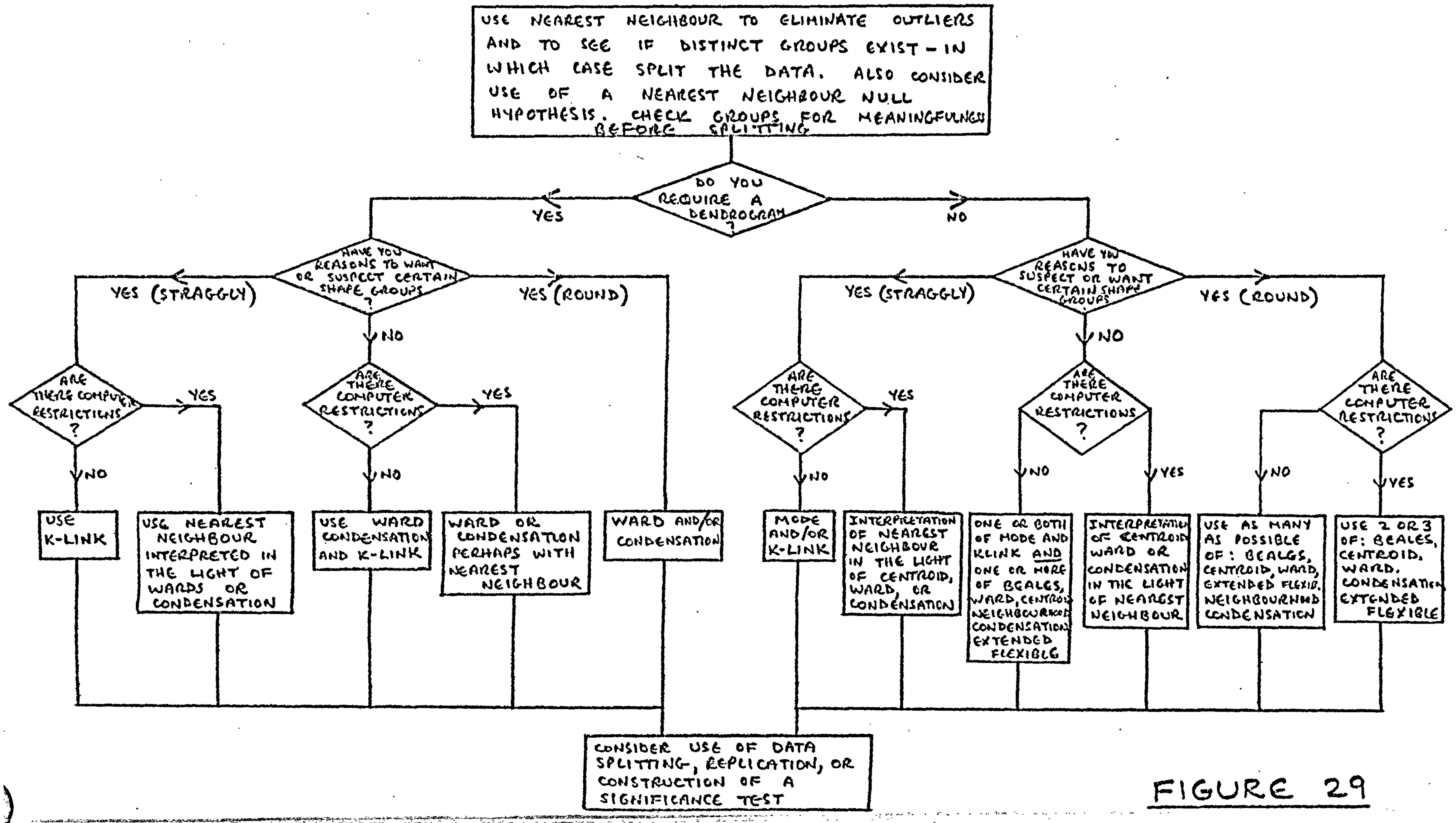

USE NEAREST NEIGHBOUR TO ELIMINATE OUTLIERS AND TO SEE IF DISTINCT GROUPS EXIST - IN WHICH CASE SPLIT THE DATA. ALSO CONSIDER USE OF A NEAREST NEIGHBOUR NULL HYPOTHESIS. CHECK GROUPS FOR MEANINGFULNESS BEFORE SPLITTING
DO YOU REQUIRE A DENDROGRAM ?
YES
NO
HAVE YOU REASONS TO WANT OR SUSPECT CERTAIN SHAPE GROUPS ?
YES (STRAGGLY)
YES (ROUND)
NO
ARE THERE COMPUTER RESTRICTIONS ?
YES
NO
USE K-LINK
USE NEAREST NEIGHBOUR INTERPRETED IN THE LIGHT OF WARDS OR CONDENSATION
USE WARD CONDENSATION AND K-LINK
WARD OR CONDENSATION PERHAPS WITH NEAREST NEIGHBOUR
WARD AND/OR CONDENSATION
HAVE YOU REASONS TO SUSPECT OR WANT CERTAIN SHAPE GROUPS
YES (STRAGGLY)
YES (ROUND)
NO
ARE THERE COMPUTER RESTRICTIONS ?
YES
NO
MODE AND/OR K-LINK
INTERPRETATION OF NEAREST NEIGHBOUR IN THE LIGHT OF CENTROID, WARD, OR CONDENSATION
ONE OR BOTH OF MODE AND KLINK AND ONE OR MORE OF BEALGS, WARD, CENTROID NEIGHBOURHOOD CONDENSATION EXTENDED FLEXIBLE
INTERPRETATION OF CENTROID WARD OR CONDENSATION IN THE LIGHT OF NEAREST NEIGHBOUR
USE AS MANY AS POSSIBLE OF: BEALGS, CENTROID, WARD, EXTENDED FLEXIB. NEIGHBOURHOOD CONDENSATION
USE 2 OR 3 OF: BEALES, CENTROID, WARD, CONDENSATION EXTENDED FLEXIBLE
CONSIDER USE OF DATA SPLITTING-, REPLICATION, OR CONSTRUCTION OF A SIGNIFICANCE TEST

FIGURE 29

general layout and to realize that each problem must be analysed to investigate the requirements of the method(s) to be employed. The validation of results is also a very important stage of the approach which depends largely on the problem. The defects of each method, as shown in the previous section, should be borne in mind, and play a part in the decision process. For example the choice between group average and weighted average can largely be decided from consideration of the problem and type of result wanted.

The Use of Null Hypotheses

We show the null hypotheses for six of the best methods in our study in Figures 30-35. These are all for thirty points in two dimensions. Figures 36 and 37 show for one of the methods the effect of a change in dimensionality, and the effect of an increase in number of points, for comparison with Figure 38 which is based on ten sets of normally distributed data - 30 points in 2 dimensions. The three sets were all normalized to unit variance.

The main points of interest from these diagrams is the similarity in the shape of the curves, and the close resemblance between the graphs obtained from 30 points in 2 dimensions, and 30 points in 4 dimensions. The graph of 60 points in 2 dimensions is somewhat steeper than the other graphs. Thus one can suppose a family of curves of similar shape becoming more steep as the number of points increases and less dependent on the number of dimensions.

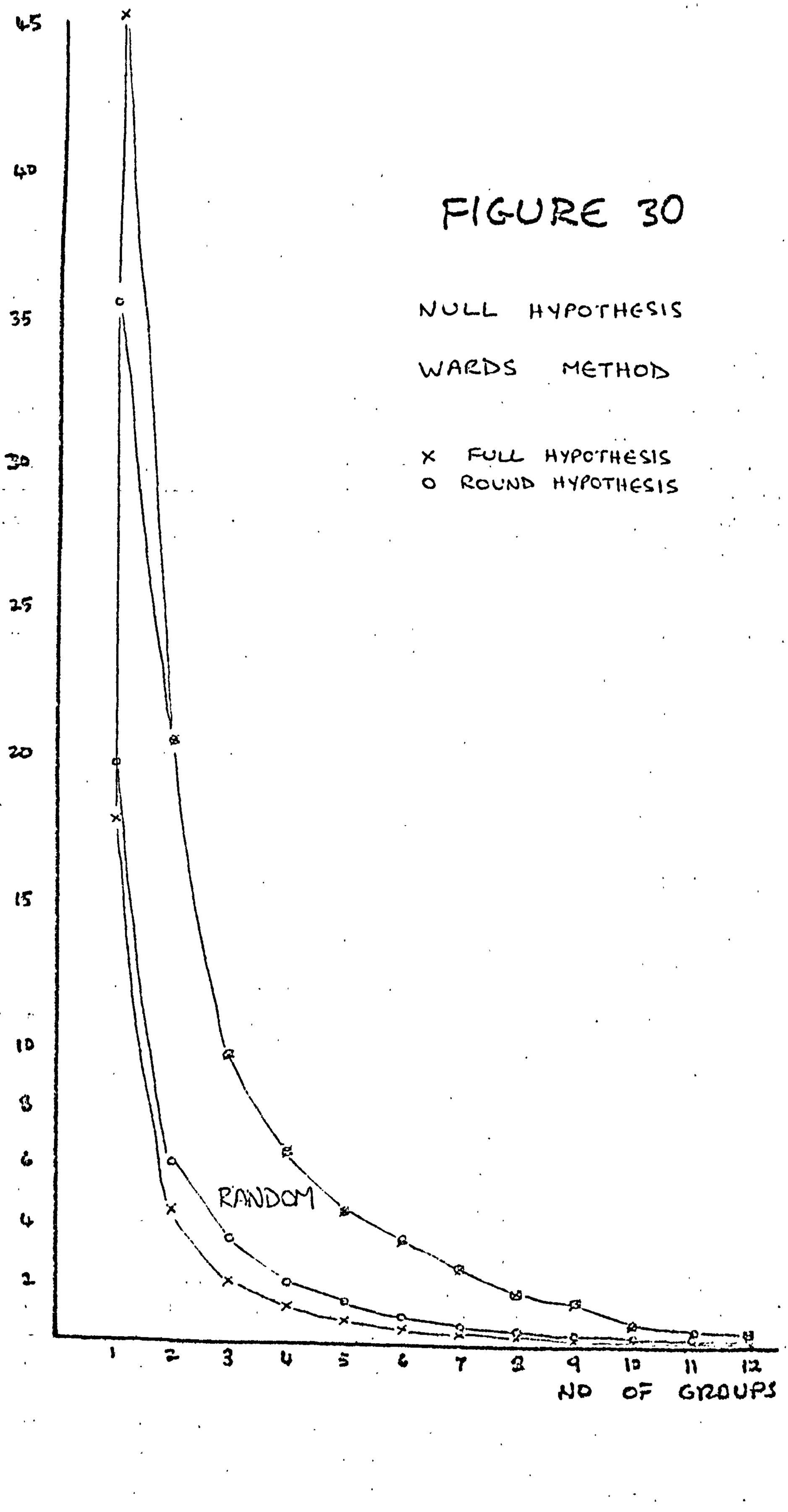


FIGURE 30

NULL HYPOTHESIS

WARDS METHOD

X FULL HYPOTHESIS
O ROUND HYPOTHESIS

FIGURE 31

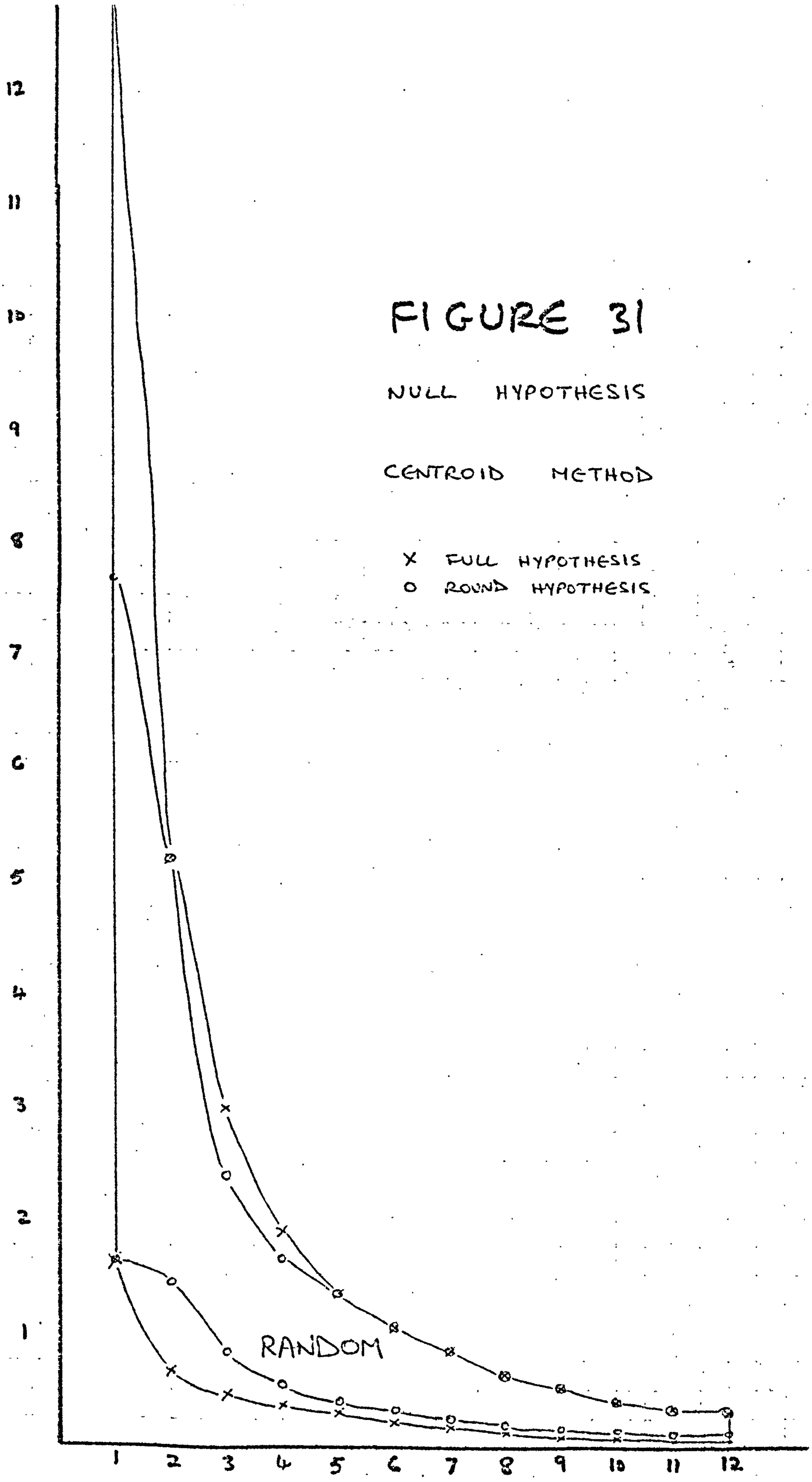
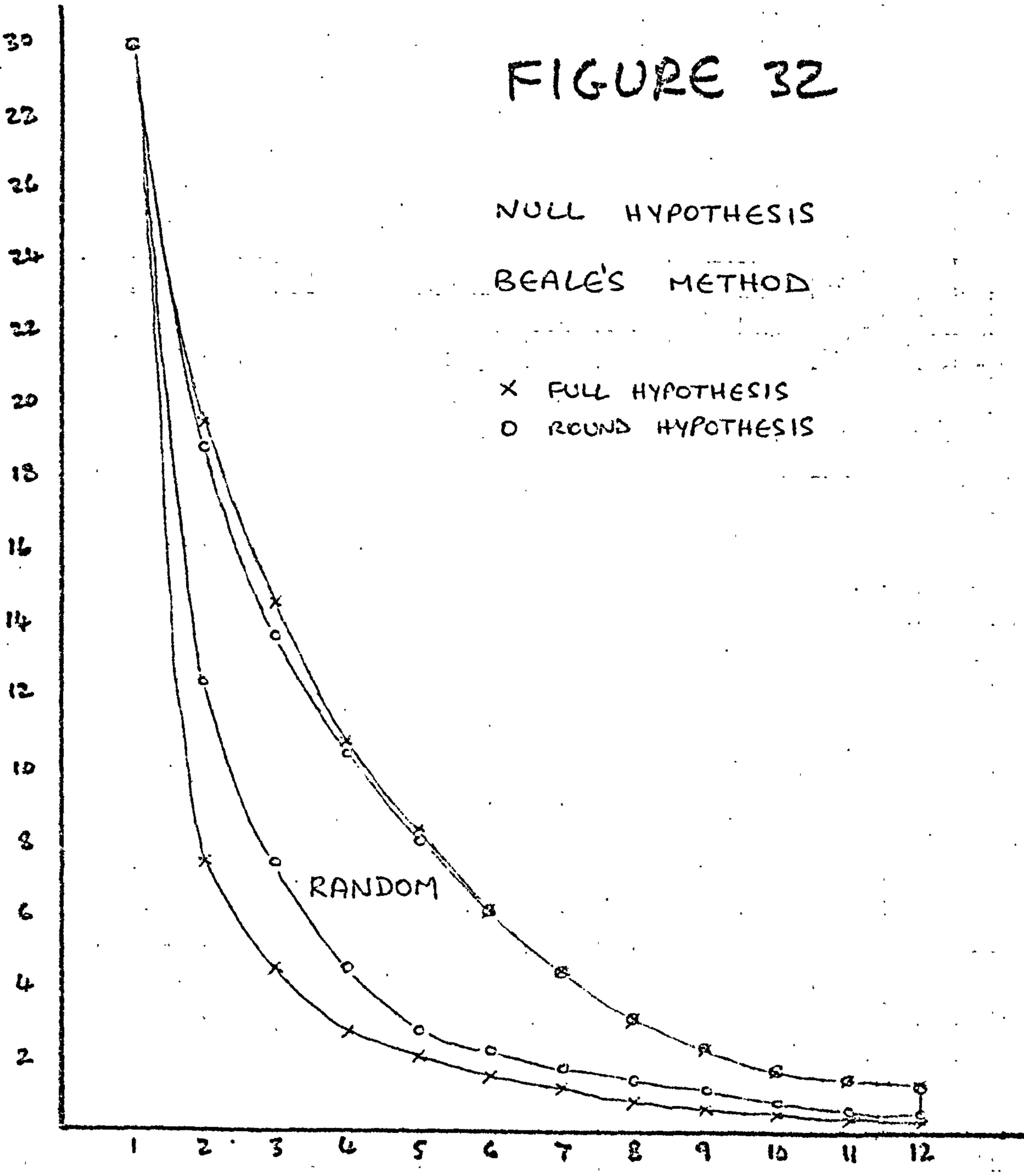

NULL HYPOTHESIS
CENTROID METHOD
X FULL HYPOTHESIS
O ROUND HYPOTHESIS
RANDOM
1
2
3
4
5
6
7
8
9
10
11
12

FIGURE 32



NULL HYPOTHESIS

BEALE'S METHOD

× FULL HYPOTHESIS
O ROUND HYPOTHESIS

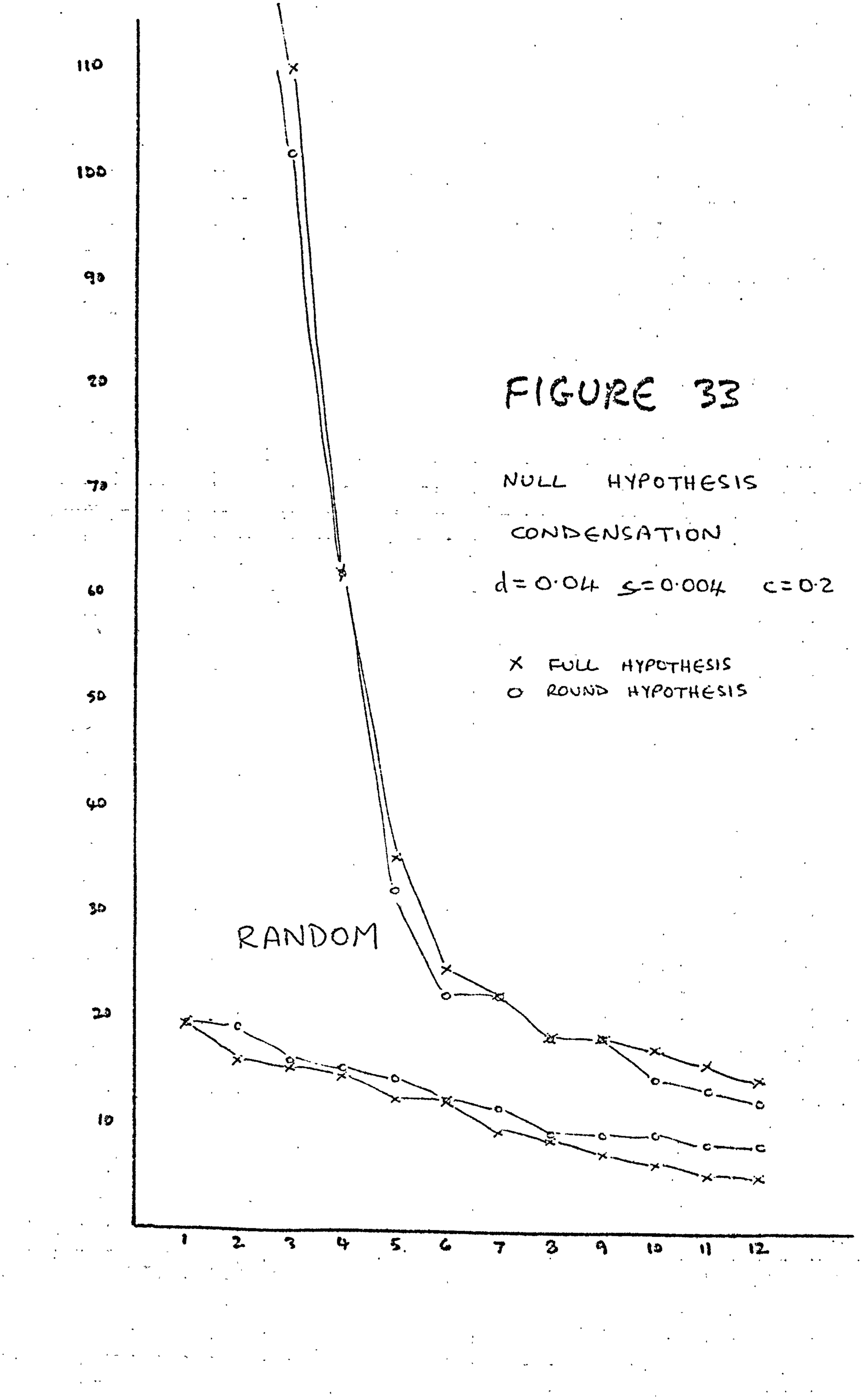


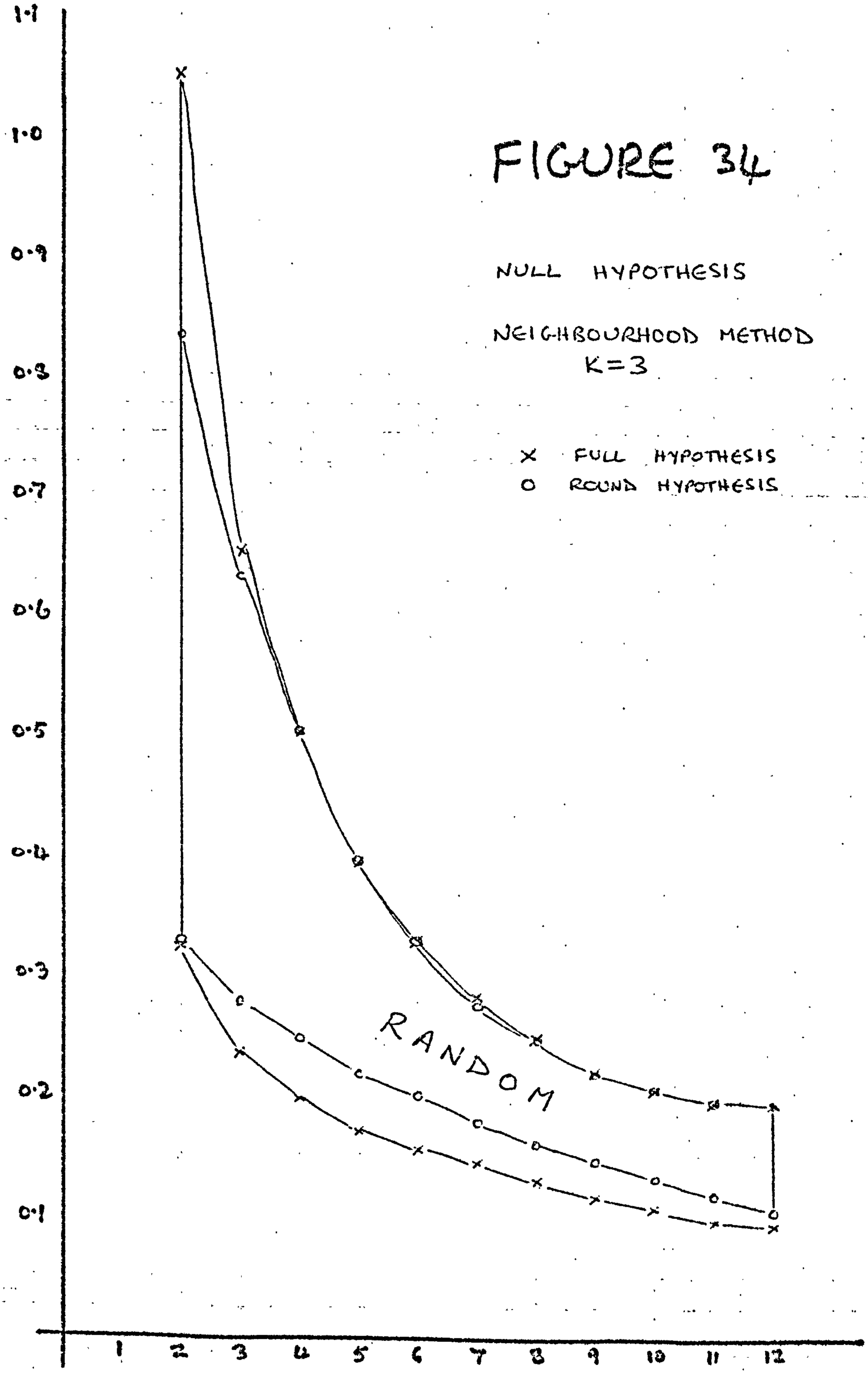
FIGURE 33

NULL HYPOTHESIS

CONDENSATION

$d = 0.04$ $s = 0.004$ $c = 0.2$

FIGURE 34



NULL HYPOTHESIS

NEIGHBOURHOOD METHOD
K=3

X FULL HYPOTHESIS
O ROUND HYPOTHESIS

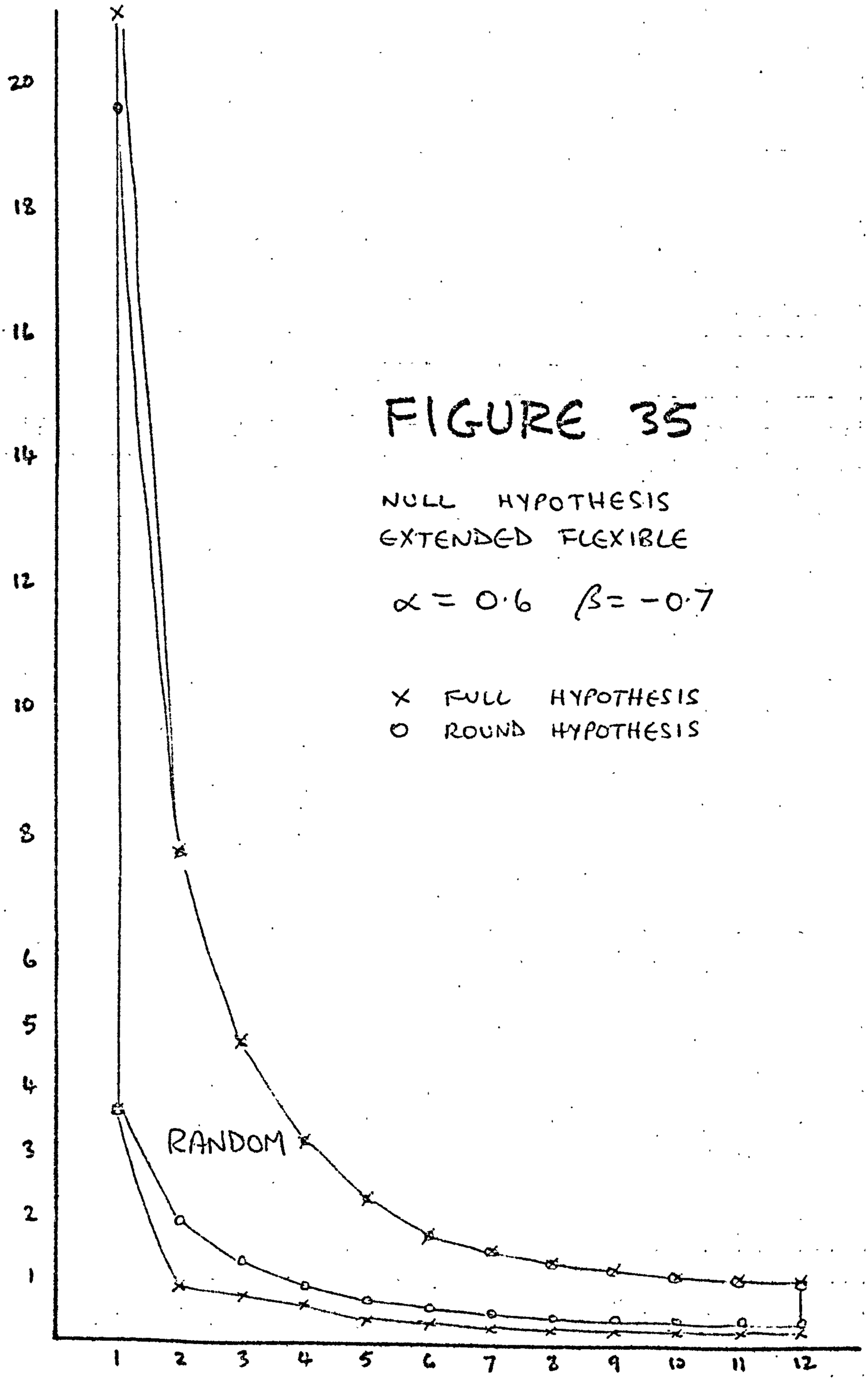


FIGURE 35

NULL HYPOTHESIS
EXTENDED FLEXIBLE

$\alpha = 0.6 \quad \beta = -0.7$

X FULL HYPOTHESIS
O ROUND HYPOTHESIS

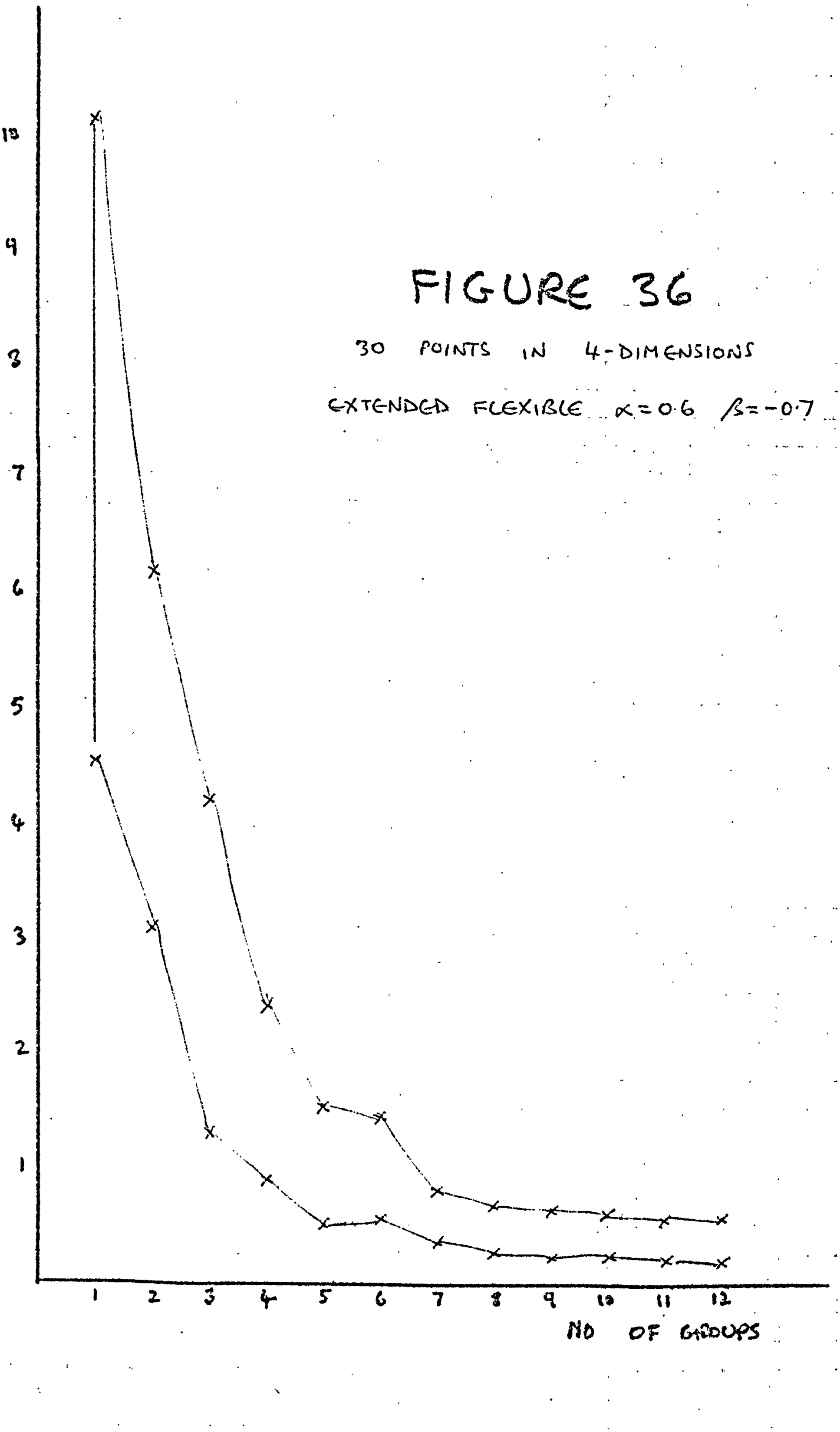


FIGURE 36

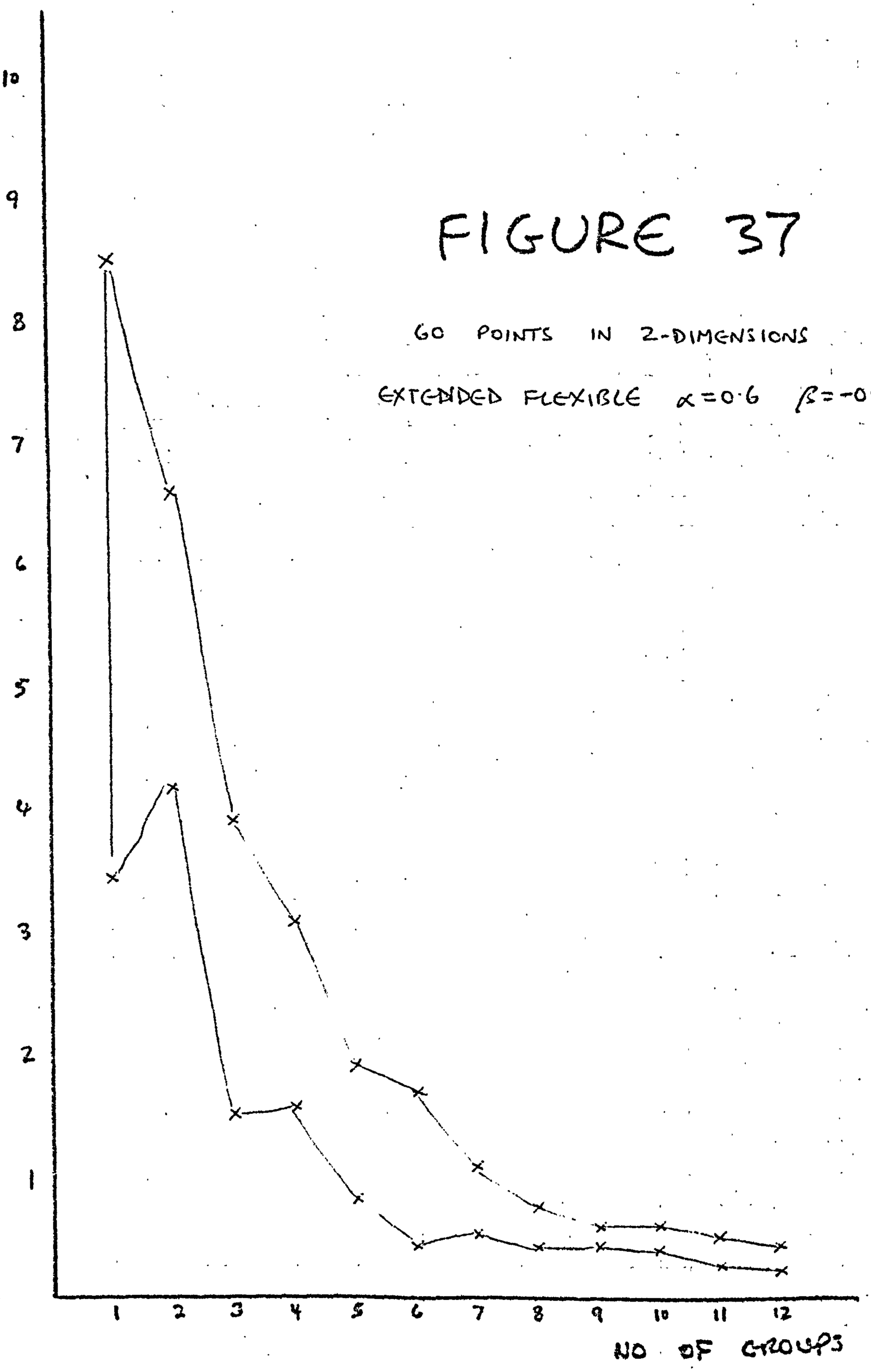
30 POINTS IN 4-DIMENSIONS

EXTENDED FLEXIBLE $\alpha = 0.6$ $\beta = -0.7$

FIGURE 37

GO POINTS IN 2-DIMENSIONS

EXTENDED FLEXIBLE $\alpha = 0.6$ $\beta = -0.7$



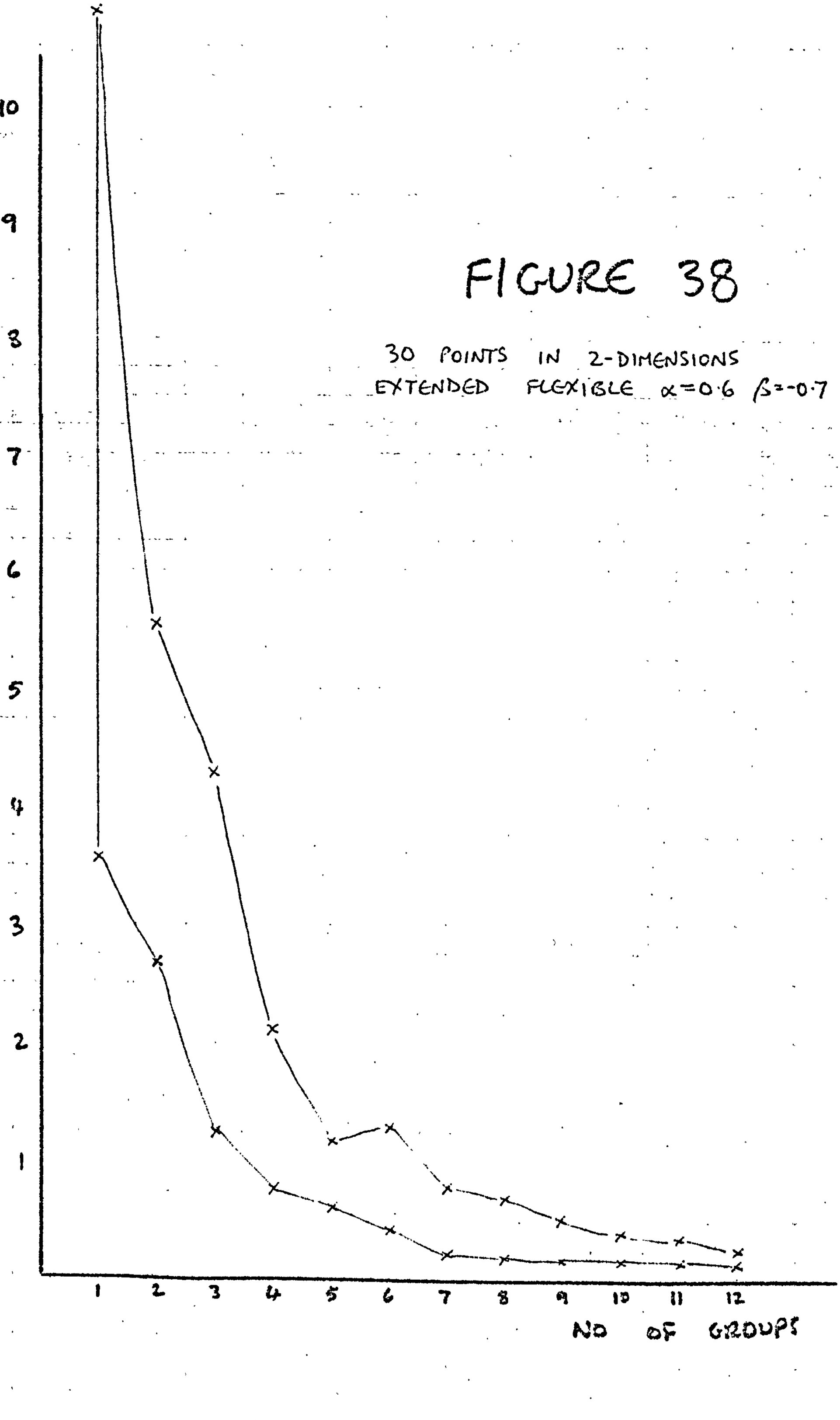


FIGURE 38

30 POINTS IN 2-DIMENSIONS
EXTENDED FLEXIBLE $\alpha = 0.6$ $\beta = -0.7$

The recommended use of null hypotheses as a practical significance test is to generate random numbers with the same number of points and the same variance in each direction (as shown by a principal components analysis), as the data under investigation, using the method of clustering to be used in the study.

Areas for Further Research

The results above and of the previous section indicate several areas which would be worth investigation. These are discussed below:

1. The extended flexible results indicated that, whilst the probably best values of α and β had been investigated, values of β less than -1.2 should be further considered.

2. The condensation method also showed good results and the parameters of the method need further refinements to determine best values. Other gravitational models would also be worth investigation - such as an inverse cube law, or a more physically exact version of the movement of masses in a viscous medium.

3. The effect on these methods of different similarity measures is one of the areas of cluster analysis which needs greater research. It is hoped that the comparison of method study in this research, being freer of dependence on the similarity measure used than some other studies, forms a basis which reduces the number of methods to be further analysed with different similarity measures.

4. Some recent work on the examination of the distance matrix for evidence of clustering (see Inglis and Johnson 1970, Hills 1969 and Kruskal 1972), seems to be a possible direction that could lead to a method of detecting whether round or other shape groups were present, and this could be a useful line for further research.

5. Of the methods we have examined for some we were unable to devise suitable null hypotheses and for others the null hypothesis was not completely satisfactory, and in this area there is a need for further work. Here again the studies in this work have identified methods which are worth investigating more in this direction.

Large Data Sets

One particular aspect of cluster analysis which is of particular interest in some sciences, is the extension of the use of the methods to large data sets. Only the nearest neighbour method can be programmed to take up store of the order n (where n is the number of objects under study), and most methods need storage of the order n^2, or by increasing computation time, storage of n*m (where m is the number of variables). This latter strategy will normally increase the number of objects which can be analysed, especially if the number of variables is reduced by principal component analysis, but there still exist restrictions on the size of n.

Our proposed strategy for dealing with large values of n is explained in one of the case studies in section E.2. The key to the problem as we see it is the fact that nearest neighbour can be programmed efficiently for several thousand objects. If the maximum number of objects that can be

analysed by a certain method on a certain computer is n then generally one can, by employing the following two strategies, achieve an effective increase in n of up to 10%. The strategies are:

1. Use of nearest neighbour to identify outliers which can be eliminated from the data set.

2. The identification of almost identical groups or pairs of objects (also by nearest neighbour); all but one from each group can be eliminated. The group representative may be taken as the group average, or more simply as one of the original group members, chosen at random. This may call for a change in weight of this group representative, depending on the cluster method used.

The strategy developed in section E.2 depends on a further property of nearest neighbour - the fact that it identifies the densest regions of objects.

Constraints

Another area of development in clustering is the use of constraints in cases where one requires a certain structure in the output clusters. These can easily be incorporated into most of the programs which we have dealt with. With the Lance and Williams algorithm methods, each pair of groups which would have been joined at a particular stage are tested to see if their fusion violates preset conditions, such as contiguity in geographical studies. Constraints can be set allowing only certain elements to join, groups to be limited to a certain number of elements or to a certain area group, etc.

The use of side conditions is of particular importance in operational research, where in order to adapt general methods to particular problems it is necessary to incorporate the constraints of a real world to reach a useful solution. Such problems are difficult to discuss in generalities, and it is one of our aims to show cluster analysis as an adaptable approach rather than a set technique, so we illustrate the use of constraints in problems by case studies given in section E.3.

N-Dimensional Data

A particular problem that can occur in certain instances is that of objects which are themselves matrices. Here the difficulty lies in calculating the similarity between the objects. We have mentioned the problem in section B.5, and there we came to the conclusion that the question is a difficult one, but one that should be resolved by the analyst in each case, and no overall solution can be put forward.

This particular problem occurs in our case study in section E.2 where for a set of companies we had three dimensional data - a breakdown of employees by age, sex, and occupation. In this example several possible resolutions of the data are discussed, and a preferred method is analysed.