AN APPROACH TO

VALIDATION IN CLASSIFICATION

A Thesis

presented by

Mageed Mohamed Abdalla

to the

University of St. Andrews

in application for the degree

of Master of Science

December 1990

ACKNOWLEDGEMENTS

I wich to express my sincere thanks and gratitute to Dr. Gordon for his useful supervision and for the greatest interest he showed in my studies. His many useful suggestions have been the greatest aid towards the successful completion of this work. I would also like to extend my sincere appreciation to Mr. Duncan who helped in the proof reading. Finally I sincerely acknowlege the G.P.C. who provided the financial assistance for my training and studies.

DECLARATION

I Mageed Mohamed Abdalla hereby certify that this thesis has been composed by myself, that it is a record of my own work and that it has not been accepted in partial or complete fulfilment of any other degree or professional qualification.

Signed..... Date.....

In submitting this thesis to the University of St. Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker.

ABSTRACT

One aim of classification is to partition a set of objects into groups of objects. If a computer package is instructed to partition the data into a specified number of groups so as to optimize some criterion of homogeneity, it will always provide a result, but this result may distort the structure in the data. Validation studies assess the results of such classifications. The approach taken in this thesis is to divide the data set into two subsets, classify these subsets separately, and compare the results with those obtained when classifying the complete data set, obtaining a measure of the reliability (R_i) with which each object has been classified and an overall measure of classifiability (M). The work concentrates on investigating partitioning data into three groups using the sum of squared distances criterion. The approach is applied to a range of artificial data sets with known structure, and to two real data sets.

CONTENTS

CHAPTER I

CHAPTER I

INTRODUCTION

1. 1. GENERAL STATEMENT OF CLASSIFICATION

Classification is an important tool in exploratory multivariate data analysis. It is a tool to reduce or summarise large data sets so that they can be more easily handled or analysed. In short, objects sharing a class are similar to one another, and unlike those from other classes. We shall not discuss this at length since it has been treated adequately elsewhere (Gordon, 1981 & 1987; Everitt, 1980), but Cormack's review paper (1971) gives an example of a comprehensive review. Note that classification is a different subject from discrimination because in this situation we assume that there are no pre-determined groups.

1. 2. VALIDATION IN CLUSTER ANALYSIS

Cluster validity is concerned with the objective interpretation of the results provided by clustering algorithms and tries to separate ' true ' structures from artifacts of clustering algorithms (Bailey and Dubes, 1982; Dubes and Jain,

1978, Smith and Dubes,1980). A clustering method, for example sum of squares or single link method, can provide a partition into (e.g.) 3 groups. If a computer package (such as CLUSTAN, (Wishart, 1987) which will be used to analyse the data throughout) is instructed to find 3 groups, it will always provide a solution even if there are no groups in the data. Also, the same data can give different groups when analysed by different clustering methods. How can we be confident that this is a reasonable summary of the data?

To make sure whether these objects fall naturally into these groups, it is thus important to validate the results. In other words if the structure of the data contains groups, it is important to see if these have been successfully discovered in the classification provided. The general approach that is taken here is:

(1) To cluster the complete data set;

(2) To divide the complete data set into two subsets;

(3) To cluster each subset separately, and compare the results with those obtained in (1). If similar results are obtained, we may be more confident that there is genuinely structure in the data. A review of such work is given by Smith and Dubes (1980). The topic of validation is linked with that of stability: here, the idea is to note how the results change if the original data set is slightly changed, before being analysed by the same clustering method (e.g. Strauss et al., 1973; Gordon and De Cata, 1988).

Early work by Sokal and Michener (1958) compared by eye a hierarchical classification of half of the data (odd numbered objects) with the original hierarchical classification.

Smith and Dubes (1980) check the stability of hierarchical classifications using the steps (1)-(2) described above, and compare the results using Goodman and Kruskal's gamma statistic (Hubert, 1974). Our work examines the validation of a single partition of the data.

CHAPTER II

CHAPTER II

METHODOLOGY

2. 1. TYPE OF DATA

The basic data to be analysed by classification methods are commonly presented in one of two different formats, either as a raw data matrix or as a dissimilarity matrix. The raw data matrix is an n*p (n rows, individuals), (p columns, variables) matrix $X = (x_{ik})$ where x_{ik} denotes the value of the k^{th} variable observed for the i^{th} object. Even if some variables are qualitative, we can still define a measure of dissimilarity d_{ij} between the i^{th} and j^{th} individuals. Many classification methods first require the raw data matrix to be transformed into an n*n matrix (n rows & n columns) of pairwise dissimilarities $D = (d_{ij})$ where d_{ij} denotes the dissimilarity between the i^{th} and j^{th} objects. A dissimilarity measure satisfies certain minimum conditions:
$d_{ij} \geq 0$, $d_{ii} = 0$ and $d_{ij} = d_{ji}$ for all i, j belonging to the set of objects. One example is $d_{ij} = \sum_k (x_{ik} - x_{jk})^2$. This is one of the most commonly used measures of dissimilarity, the squared Euclidean distance between points i and j, but there are many other measures, some defined for qualitative data, or a mixture of data types; such measures are presented by Gower (1985).

2. 2. SUM OF SQUARES METHOD

This technique can be used for the classification of objects which can be represented as points in Euclidean space of some number of dimensions. Let x_{ik} (i=1..., n; k=1....., p) denote the k^{th} co-ordinate of the i^{th} point, p_i. The aim is to partition the set of n points into g groups so as to minimize the total within-group sum of squares about the g centroids, i.e. if the centroid of the m^{th} group, which contains the n_m points,

$$P_{m_i} \ (i=1, \ldots . n_m)$$

has co-ordinates

$$Z_{mk} \equiv \frac{1}{n_m} \sum_{i=1}^{n_m} x_{m_i k} \qquad (k=1, \ldots ., p)$$

and if the within-group sum of squares of the m^{th} group is

$$S_m \equiv \sum_{i=1}^{n_m} \sum_{k=1}^{p} \left(x_{m_i k} - Z_{mk} \right)^2$$

then the aim is to find a partition which minimizes

$$S(g) \equiv \sum_{m=1}^{g} S_m$$

There are many different ways of partitioning n objects into m classes (Fortier and Solomon, 1966) and it is not computationally feasible to examine all of them, to see which is the best partition. Fortier and Solomon show that the number is

$$P(n,m) = \left\{ m^n - \sum_{i=1}^{m-1} m_{(m-i)} \quad P(n, i) \right\} / m!$$

In general two main types of algorithm have been used to search for a minimum sum of squares partition: (1) agglomerative, and (2) iterative relocation. The latter will be used here.

In the iterative relocation procedure, we start with an initial partition into g groups and relocate an object from one group into another if this reduces the sum of squares. The initial partition can be either a systematically-chosen computer start or alternatively a random start; both of these methods will be discussed later. This procedure of relocation continues until a local minimum of the sum of squares is reached, in the sense that the relocation of any object will not reduce the value of the sum of squares. Many algorithms of this kind have been proposed (e.g. Forgy,1965; Jancey,1966; MacQueen,1967). Other sum of squares algorithms are reviewed by Gordon and Henderson (1977).

2. 3. CLUSTAN'S CODES & FUSION

All the clustering procedures produce results in the form of a classification with the following content:

Line 1 title

Lines 2, 3, ... C_i, i = 1, 2, ... n

The value of C_i is the code of the cluster to which the i^{th} case belongs (Wishart, 1987). In other words it is the symboled pigeon-hole into which an individual entity can be dropped or located in the classification separated store $ class when CLUSTAN runs.

The fusion merges the closest pair of groups to obtain a partition into k-1 groups from a partition into k groups. In our use of CLUSTAN, we always started with the number of groups that were sought and so no fusion occured, only relocation.

2. 4. METHODOLOGY ILLUSTRATED ON RANDOMLY-GENERATED DATA

Before approaching the analysis of real data, it will prove instructive to analyse randomly-generated data in order to examine the behaviour of the method on data whose structure we know. In this way we can become more familiar with the processes involved and introduce any necessary adjustment theoretically before tackling data drawn from the real world. Thirty points will be generated randomly in two dimensions - the main properties of the procedure should emerge from detailed study of small data sets and having only two dimensions allows the data to be drawn in the plane.

We want to find the best (in the sense of minimum sum of squares) partition into 3 groups, and proceed as follows:

1-Start with a random partition into 3, (fusion minimum 3, codes 1-3) and use CLUSTAN'S iterative relocation routine to search for the 'best' partition into three groups.

2- The process was repeated 5 times, because the starting partition can influence the final partition.

However, the computer 'random start' is a systematic one, i.e. it looks like :

1, 2, 3, 1, 2, 3, 1, 2, 3, ...

This means that at the start of the iterations, CLUSTAN will locate the first object in group 1, the second

object in group 2, the third object in group 3, then the fourth object in group 1 and so on for the whole data set, and so one obtains the same starting partition by using this CLUSTAN option. Therefore, we used a short computer program to provide a different random starting partition, e.g.

1, 1, 2, 1, 3, 2, 1, 2, 3, 3, 3, 2, 1.....

Almost always the same final partitions were found, on all five occasions. When this was not the case we found which of the different results had the smallest sum of squares (using a short computer program). To illustrate our approach, for one of the randomly-generated data sets the final results of the three groups were found to look like this :

A - 8, 23 , 12 , 21 , 11 , 2 , 1 , 20 , 10 , 6 , 17

B - 5, 3 , 7 , 30 , 14 , 18 , 26 , 27 , 13 , 9 , 16

C - 4 , 22 , 15 , 28 , 25 , 24 , 29 , 19

A,B,C is the partition of the complete data set of 30 objects and these objects are illustrated in Figure 1.

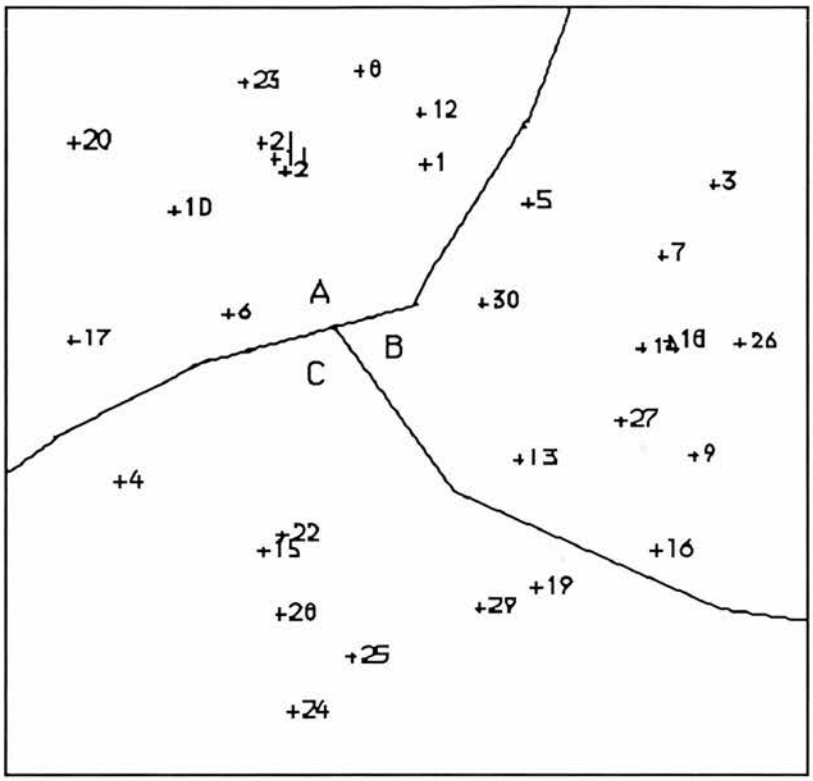

Figure 1
Illustrating the scatter of the
objects in the groups

To discover whether these objects fall naturally into these three groups, the whole data set was divided into two subsets and each subset was clustered separately. We wanted to know if the same results could be obtained for the subsets. If similar results are indicated we are more confident that there is genuinely structure in the data set. It is necessary to make sure that each object is assigned to only one subset. Two different ways of choosing the two sub-sets were used. (1) Matching and

(2) Random selection. These are described in subsections 2. 4. 1 and 2. 4. 2.

2. 4 . 1. CHOOSING SUBSETS BY MATCHING

For matching these 30 objects some short FORTRAN programs were used. The aim is to match the objects in pairs so that the sum of the dissimilarities within the pairs is minimized. In symbols, let d_{ij} denote the dissimilarity between the i^{th} and j^{th} objects and let v_{ij} =1 if objects i and j are matched, and 0 if they are not matched. This problem will be solved as an assignment problem.

$$\min \sum_{ij} d_{ij} * v_{ij}$$

subject to:

$$\sum_i v_{ij} = 1$$

$$\sum_j v_{ij} = 1$$

$$v_{ij} = v_{ji}$$

$$v_{ij} = 0 \text{ or } 1$$

A short computer program was written that called routine HO2BAF from the NAG library to carry out this matching. The output is pairs of matched objects, e.g.

1 - 5

2 - 11

3 - 7

4 - 15

.

.

The first number in each row was chosen to belong to the first sub-set, and the second number was chosen to belong to the second sub-set, as follows:-

1st sub-set

1, 2 ,3 ,4 , 6, 8, 9, 10, 13, 14, 18, 19, 21, 22, 24

2nd sub-set

5, 11, 7, 15, 17, 12, 16, 20, 30, 27, 26, 29, 23, 28, 25

The same classification procedure was carried out on each sub-set. If there is an exact correspondence with the results of analysing the complete data set, one expects to find in the first subset the following:-

1st sub-set groups:-

1: 1, 2, 6, 8, 10, 21

2: 3, 9, 13, 18, 14

3: 4, 19, 22, 24

Table 1

	A	B	C
1	1, 2, 6, 8, 10, 21		
2		3, 9, 13, 18, 14	
3			4, 19, 22, 24

A, B, C, is the partition of the whole data set.
1, 2, 3, is the first sub-set partition.

In the analysis of the other half of the data set, an exact correspondence of results would give the following partition:

2nd sub-set groups :-

I: 11, 17, 12, 20, 23

II: 5, 7, 16, 27, 26, 30

III: 15, 25, 28, 29

Table 2

	A	B	C
I	11,17,12 20,23		
II		5,7,16,30 26,27	
III			15,25,28 29

I,II,III is the second sub-set partition.

Each of the two sub-sets were partitioned into three groups five times. The same three partitions were actually found on each of the five attempts. The results corresponded exactly with those given in Tables 1 and 2 (sometimes the labels of the groups were different).

When the whole data set was compared in pairs, this gives: (Note that one of the matched pairs was divided into different groups in the partition; see Figure 1).

Table 3

	I	II	III
1	(2, 11), (6, 17) (8, 12), (10,20) (21, 23)	(1, 5)	
2		(3, 7), (9, 16) (13, 30), (14, 27) (18, 26)	
3			(4, 15), (22, 28) (24, 25), (19, 29)

The points according to their subsets were then compared to the whole data set as follows:-

Table 4

	A	B	C
1 /	1, 2, 6, 8, 10, 21 *1, 11, 12, 17, 20, 23*		
2 //		3,9,13,14,18 *5, 7, 16, 26, 27, 30*	
3 ///			4, 19, 22, 24 *15, 25, 28, 29*

1, 2, 3, is the first subset partition

/, //, ///, is the second subset partition

Because these objects are not clustered, but randomly generated, the result was not expected to be perfect agreement.

The analysis was repeated using different assignments of each object from a matched pair to a separate subset, with similar disappointing results. It therefore seems that choosing the subsets by matching is probably not helpful.

2. 4. 2. RANDOM SELECTION OF SUBSETS

In this method the 30 objects were randomly divided into two subsets each containing 15 objects using a routine from the NAG library to generate two random subsets. For the data considered earlier, the results were:-

1st subset:-

13, 19, 12, 11, 28,10, 29, 2, 4, 16, 5, 22, 15, 7, 18

2nd subset:-

1, 21, 20, 8, 30, 23, 27, 24, 6, 9, 25, 3, 26, 14, 17

As before, each subset was partitioned into three groups using the CLUSTAN package five times. The same partition was found on each of the five attempts as follows:-

1st subset groups:-

1: 2, 5, 10, 11, 12

2: 4, 22, 15, 28

3: 7, 13, 16, 18, 19, 29

These were then compared with the whole data set partitions as follows:-

Table 5

	A	B	C
1	2, 10, 11 12	5	
2			4, 15, 22 28
3		7, 13, 16 18	19, 29

The same analysis was done for the second subset with the results as follows:-

2nd subset groups:-

I: 1, 6, 8, 17, 20, 21, 23

II: 3, 9, 14, 26, 27, 30

III: 24, 25

When we compared this with the partition of the whole data set we found the following :-

Table 6

	A	B	C
I	1,6, 8, 17 20, 21, 23		
II		3, 9, 14, 26 27, 30	
III			24, 25

2. 4. 3. MEASURES OF AGREEMENT AND RELIABILITY

We want a measure of how closely the two partitions (A, B, C) and (1, 2, 3) agree with one another. In order to achieve this we match the groups in pairs so as to maximize the total numbers of objects in common. This was obtained from the matching A1, B3 and C2, which has twelve objects in common. We

define as a measure of agreement the proportion of the objects that are in common. Here, the Measure of Agreement

M =12/15 = 80%

We are more confident about the results for the 12 points in common and less confident about the other three points: 5, 19 and 29. Similarly the second subsets groups were matched in pairs AI, BII and CIII. Perfect agreement was found this time. Averaging the two measures of agreement, we have a Measure of Agreement for the whole data set:-

M = 1/2 (12/15+15/15) = 27/30 = 90%

We are more confident about the results for 27 of these objects and less confident about the other three: 5, 19 and 29. One can go further in analysing the data of the random divisions into subsets. Four other pairs of subsets were chosen and were treated by the same method of analysis five times for each subset. The results were found to be:

Attempts	M	Non-matched objects
The 2nd	= 1/2(14/15+15/15) = 29/30 = 97%	(4)
The 3rd	= 1/2(13/15+15/15) = 28/30 = 93%	(1, 16)
The 4th	= 1/2(14/15+15/15) = 29/30 = 97%	(13)
The 5th	= 1/2(14/15+14/15) = 28/30 = 93%	(5, 17)

9 not matched out of 5*30=150.

141 in agreement out of 150 in agreement. M =141/150= 94%

For each of the objects we can define a measure of the ' reliability ' R_i with which it is represented in the classification of the whole data set; this is the proportion of times that it appears in common e. g. for these data:

$Object_i$	R_i	%
1	4/5=	80
4	4/5=	80
5	3/5=	60
13	4/5=	80
16	4/5=	80
17	4/5=	80
19	4/5=	80
29	4/5=	80
.	.	.
.	.	.
	5/5=	100

All other objects having reliability 5/5.

M = 141/150= 94%

It can be seen from Figure 1 that the objects that have smaller values of R_i are close to the boundaries between groups. But, sometimes, we expect to be faced with the situation that the matched cells are not unique, e. g.

Table 7

	A	B	C
1	5		29, 13, 12, 10
2	22, 14, 28	1, 21, 18, 30	
3	8	11, 16	

Two different matchings lead to there being nine objects in common:

	1C	(29, 13, 12, 10)	4
(i)	2B	(1, 21, 18, 30)	4
	3A	(8)	1
	or		
	1C	(29, 13, 12, 10)	4
(ii)	2A	(22, 14, 28)	3
	3B	(11, 16)	2

If the maximum matching is not unique, a fraction is added to the numerator of the reliabilities of each of the objects in the matching; here, objects 29, 13, 12, 10, receive a contribution of 1 (1/2+1/2) because they all belong to both optimal matchings, whereas all the other objects receive a

contribution of only 1/2 because they belong to only one of the optimal matchings.

2. 5. SUMMARY

To summarise, one random data set of 30 objects in two dimensions was divided into 3 groups using the sum of squares method implemented by an iterative relocation algorithm. To check whether there are genuine groups in the data, the 30 objects were then divided into two subsets. Two methods of division were employed:

(1) By matching them in pairs with one of each pair being placed in one of the subsets (objects within a pair were placed in different subsets);

(2) By random division into subsets of 15.

In each case, each of the subsets was divided into 3 groups using the sum of squares criterion and the results compared with those obtained by dividing the entire data set into 3. In every case, 5 different attempts were made, so as to try to ensure that the optimal sum of squares partition would be found. Identical results were almost always found. If not, a short computer program checked which of the two or more partitions had the smallest sum of squares .

Comparison yielded firstly a measure of classifiability for the entire data set (M) and secondly a measure

for each object of how readily it could be classified (R_i), taking into account the adjustments for the non-unique matching if necessary. Method (1) matching turned out to be less useful than method (2) random division for obtaining the two subsets. It is now necessary to repeat the whole process for other random data sets (still 30 objects in 2 dimensions) but with a different random ' seed ', and for data sets containing different structure.

CHAPTER III

CHAPTER III

RESULTS & CONCLUSION

3. 1. INTRODUCTION

As was discused in chapter 2 random division was chosen as a method to divide the whole data set into two subsets with 15 objects in each subset. Random division seems to give useful results in checking whether or not there are genuine groups in the data as the first step to validate the results. The aim now is to examine the behaviour of the validation approach when it is dealing with different sorts of data. The application firstly will be on generated data and secondly on real data. The general approach has been already described in chapter 2. In section 3. 2, random data are modified in several different ways to provide three equal-sized groups:

(i) data almost random,

(ii) groups almost separated,

(iii) well-separated groups.

In section 3. 3, a similar procedure was carried out to create four approximately equal-sized groups. Since the size of groups is expected to influence the results, in section 3. 4 a

similar investigation was carried out on data consisting of three groups of different sizes. In section 3. 5 the investigation was carried out on data consisting two equal-sized groups.Some conclusions about how the approach has performed in analysing these artificially-generated data are given in section 3. 6. Finally as previously stated two real data sets are examined in section 3. 7.

3. 2. INVESTIGATION OF THREE

EQUAL-SIZED GROUPS

Five random data sets of thirty points in two dimensions were generated in the manner described in Section 2. 4, with a different "seed" for each data set. Each of these data sets was then analysed in the standard manner: each data set was divided randomly into two subsets of 15 objects, the optimal partition of each subset into three groups was found (from the best of five investigations, the results usually being identical), and then these results were compared with the optimal partition of the complete data set into three groups; as before, each data set was randomly divided into two subsets five times. The results are summarized in the following table:-

Table 8

Measure of agreement for the classification of five random data sets partitioned into three groups.

	No.of Objects (out of 30) in agreement					
Data	Replication Number					
set	1	2	3	4	5	M
1	19	25	19	25	26	114/150=76.0%
2	30	29	27	28	30	144/150=96.0%
3	28	28	25	21	24	126/150=84.0%
4	30	24	25	26	25	130/150=86.7%
5	27	30	26	28	27	138/150=92.0%

Table 8 shows that the data set 1 had a Measure of agreement M = 76.0%, considerably lower than that of the other data sets, so a further study was carried out. Figure 2 shows the configuration of the points. These points that had smaller value of R_i generally lay close to the partition edges of the groups, e.g. points 4, 11, 19 and 28 all had $R_i \leq 0.5$ and the points 1, 3, 5, 7, 16, 20, 22, 24 and 30 all had $R_i \leq 0.7$. The other four data sets had apparently better-separated groups.

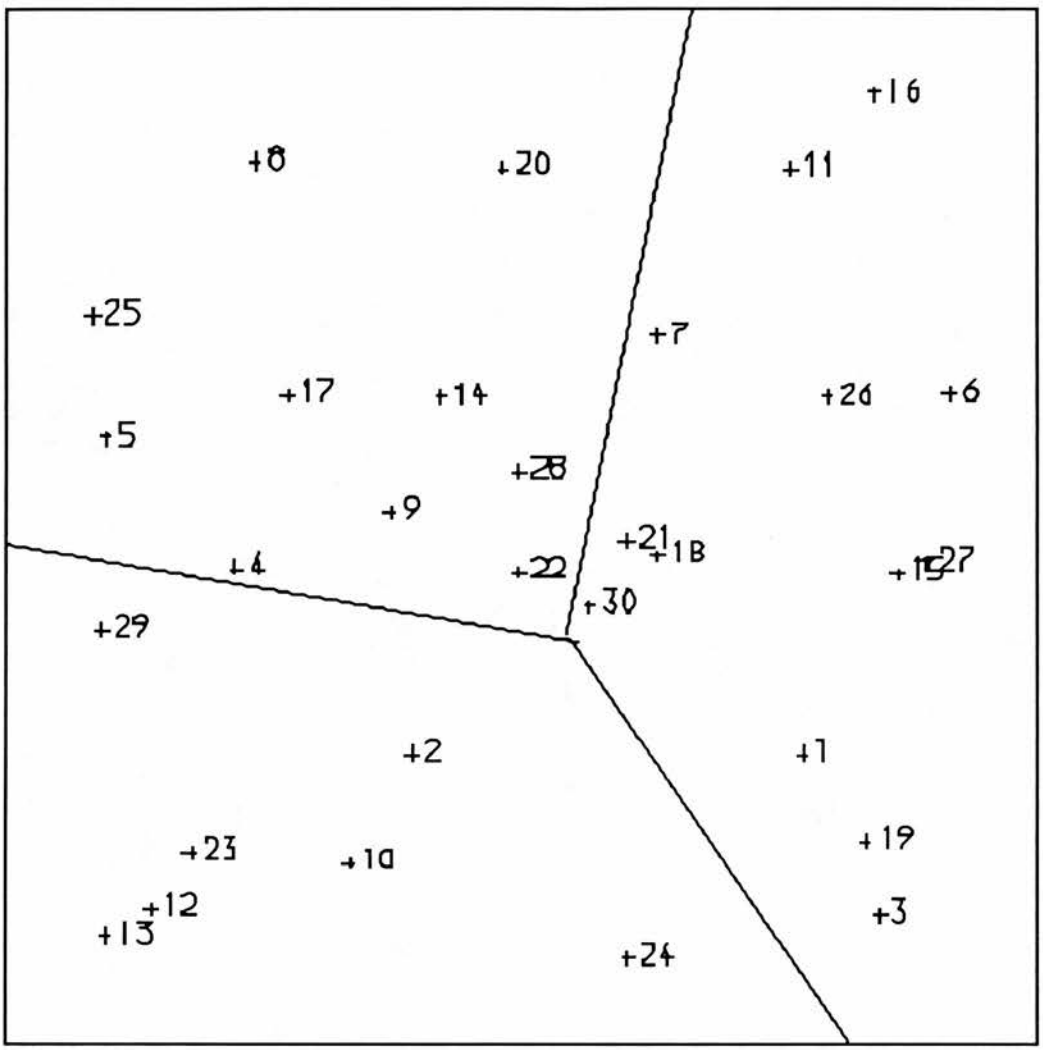

Figure 2

Illustrating the partition of data set 1 into three groups

Each of these five randomly-generated data sets was modified so as to introduce some structure into the data. This was done by moving each point a distance R from the centre (see Figures 3 and 4). The points were moved as follows:

(1) For all points in the first ten (1-10)

$X \longrightarrow X + R\sqrt{3}/2$

$y \longrightarrow y + R/2$

(2) For all points in the second ten (11-20)

$X \longrightarrow X - R\sqrt{3}/2$

$y \longrightarrow y + R/2$

(3) For all points in the third ten (21-30)

$X \longrightarrow X$

$y \longrightarrow y - R$

Three different values of R were chosen, so as to obtain data that were :

(i) almost random,

(ii) groups almost separated,

(iii) well separated groups.

For these data, after some investigation, the three values of R to achieve this were chosen to be 0.4, 0.6, 1.0, respectively. Each of these fifteen data sets was analysed in the standard manner, and the results are given in the following table:-

Table 9

Measure of agreement for the classification of five randomly-generated data sets whose points are perturbed by an amount R so as to provide three groups

	R			
Data set	0.0	0.4	0.6	1.0
1	76.0%	78.7%	89.3%	100%
2	96.0%	74.7%	92.0%	100%
3	84.0%	90.0%	94.7%	100%
4	86.7%	98.7%	98.7%	100%
5	92.0%	86.7%	94.0%	100%

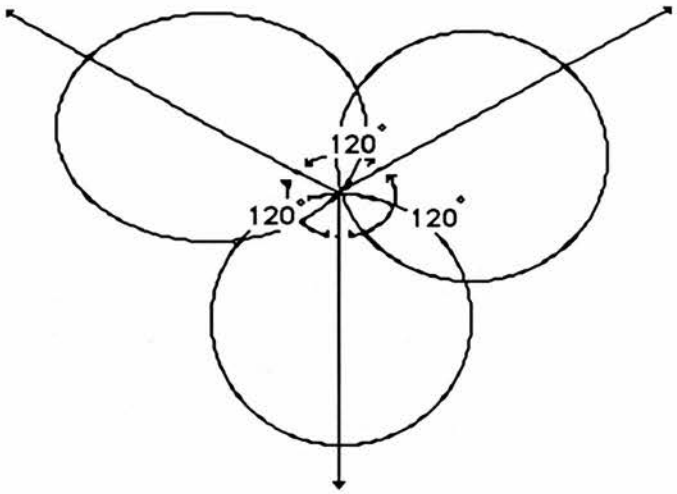

Figure 3

Illustration of how the three groups were obtained from the random data

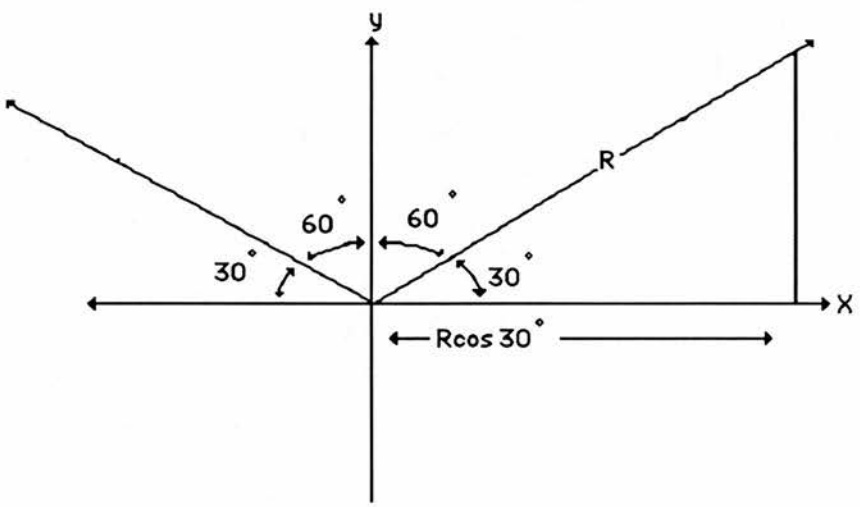

Figure 4

Illustrating the distance (R) that the points were moved to obtain the three groups.

3. 3. INVESTIGATION OF FOUR EQUAL-SIZED GROUPS

Continuing the investigation after the analysis of the three equal-sized groups, the same five data sets which were used in section 3. 2 are used here. Each of these five randomly-generated data sets was modified so as to ensure that it consisted of four groups of 7 or 8 objects each. This was done by moving each point a distance R from the centre (see Figures 5 and 6); the points are moved as follows:-

(1) For all points in the first group (1-7)

$X \longrightarrow X + R/\sqrt{2}$

$y \longrightarrow y + R/\sqrt{2}$

(2) For all points in the second group (8-15)

$X \longrightarrow X + R/\sqrt{2}$

$y \longrightarrow y - R/\sqrt{2}$

(3) For all points in the third group (16-22)

$X \longrightarrow X - R/\sqrt{2}$

$y \longrightarrow y - R/\sqrt{2}$

(4) For all points in the fourth group (23-30)

$X \longrightarrow X - R/\sqrt{2}$

$y \longrightarrow y + R/\sqrt{2}$

Three different values of R were chosen, so as to obtain data corresponding to the three stages stated in the previous section 3. 2; the three values of R were 0.3, 0.5, 1.0. Each of these fifteen data sets was analysed in the standard manner; as before, each data set was randomly divided into two subsets five times, and the results are given in the following table:-

Table 10

Measure of agreement for the classification of five modified data sets whose points are perturbed by an amount R,

so as to provide four groups.

	R			
Data set	0.0	0.3	0.5	1.0
1	76.0%	87.3%	84.0%	83.3%
2	96.0%	87.3%	82.7%	70.7%
3	84.0%	80.7%	82.0%	83.3%
4	86.7%	90.7%	92.0%	88.0%
5	92.0%	89.3%	86.0%	76.6%

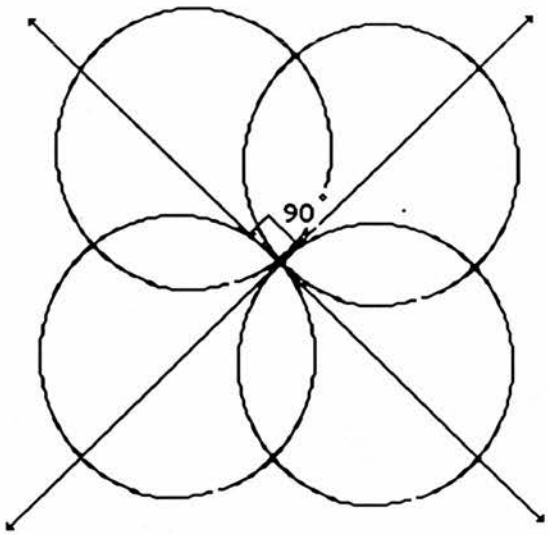

Figure 5

Illustration of how the four groups were obtained from the random data

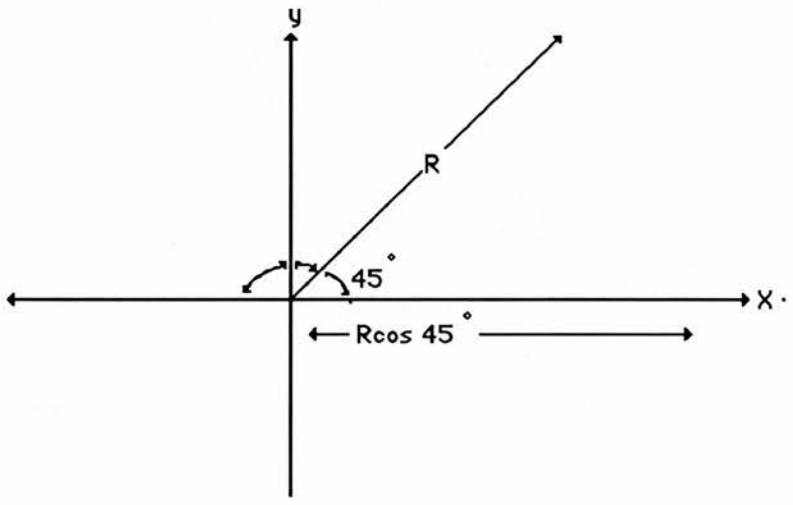

Figure 6

Illustrating the distance (R) that the points were moved to obtain the four groups.

3. 4. INVESTIGATION OF THREE

DIFFERENT-SIZED GROUPS

This time the same five data sets which were used in section 3. 2 are modified so as to introduce some structure into the data; the data were modified to consist of three groups of 5, 10, 15, objects each. This was done by moving each point a distance R from the centre (see Figures 4 and 5). The points are moved as follows:-

(1) For all points in the first five (1-5)

x ----> x + R√(3)/2

y ----> y + R/2

(2) For all points in the second ten (6-15)

x ----> x - R√(3)/2

y ----> y + R/2

(3) For all points in the third fifteen (16-30)

x ----> x

y ----> y - R

The three different values of R were chosen so as to obtain data corresponding to the three stages stated in section 3. 2; the three values of R were 0.3, 0.5, 0.9. Each of these fifteen data sets was analysed in the standard manner; as before, each

data set was randomly divided into two subsets five times, and the results are given in the following table:-

Table 11

Measure of agreement for the classification of the five modified data sets whose points are perturbed by an amount R,

so as to provide three unequal-sized groups.

	R			
Data set	0.0	0.3	0.5	0.9
1	76.0%	89.3%	80.0%	99.3%
2	96.0%	87.3%	95.3%	100%
3	84.0%	86.7%	89.3%	98.0%
4	86.7%	83.3%	90.7%	100%
5	92.0%	95.3%	95.3%	100%

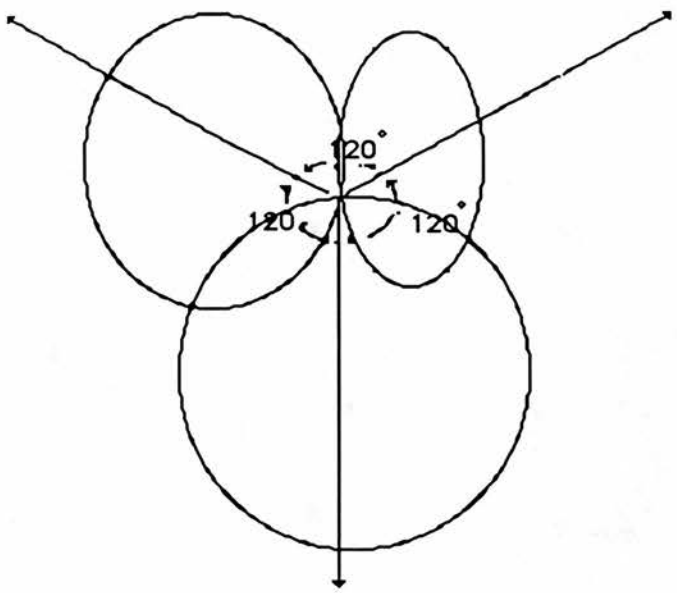

Figure 7

Illustration of how the different sized groups were obtained from the randomly-generated data

3. 5. INVESTIGATION OF TWO

EQUAL-SIZED GROUPS

Finally the same data sets which were used in section 3. 2 are modified to consist of two groups of 15 objects. This was done by moving each point a distance R from the centre. The points are moved as follows:-

(1) For all points in the first fifteen (1-15)

x ----> x + R

y ----> y

(2) For all points in the second fifteen (16-30)

x ----> x - R

y ----> y

The three different values of R were chosen so as to obtain data corresponding to the three stages stated in section 3. 2; the three values of R were 0.3, 0.6, 0.9. Each of these fifteen data sets was analysed in the standard manner; as before, each data set was randomly divided into subsets five times, and the results are given in the following table:-

Table 12

Measure of agreement for the classification of the five modified data sets whose points are perturbed by an amount R, so as to provide two-equal sized groups.

	M			
Data set	0.0	0.3	0.6	0.9
1	76.0%	75.3%	79.3%	72.7%
2	96.0%	86.0%	90.0%	88.7%
3	84.0%	83.3%	86.0 %	84.7%
4	86.7%	84.7%	78.7%	76.7%
5	92.0%	90.0%	88.7%	90.7%

3. 6. CONCLUSION

The results of analysing these artificially-generated data with known structure show that:

(1) When there are three equal-sized well-separated groups in the data, the measure of agreement M is 100%. When there are three groups of unequal size, M is very slightly lower.

(2) As the three groups become less-clearly separated M decreases, but it can still be greater than 75% for random data.

(3) When CLUSTAN is instructed to find the incorrect number of groups M can be lower than 80% even when the groups are well-separated.

(4) Objects given small values of R_i are usually close to the boundaries between groups.

3. 7. INVESTIGATION OF REAL DATA

As a further illustration of the methodology, the analysis of two separate real data sets is described in this section.

3. 7. 1. FISHER IRIS DATA

The classical Fisher Iris data set was examined. This Iris data was collected by E. Anderson in the late 1920_S and early 1930_S, and was published by R A Fisher in 1936. As described in Kendall and Stuart (1976) the data set consists of three types of Iris: Iris Setosa, Iris Versicolor and Iris Virginica. The four measurement variables are sepal length and width and petal length and width, in centimetres. The 150 plants are conveniently arranged in three groups of 50; the order is Iris Setosa (1 - 50), Iris Versicolor (51 - 100) and Iris Virginica (101 - 150). These data have been analysed many times since originally done by Fisher. As stated in Section 2.4 CLUSTAN was restricted to search for the best partition into three groups. The initial partition into three groups was chosen (a) randomly, ten different times, (b) using the three Iris groups. In each case, the same final partition was found, as follows:

A: 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, 44, 45, 46, 47, 48, 49, 50.

B: 51, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 139, 143, 150.

C: 53, 78, 101, 103, 104, 105, 106, 108, 109, 110, 111, 112, 113, 116, 117, 118, 119, 121, 123, 125, 126, 129, 130, 131, 132, 133, 135, 136, 137, 138, 140, 141, 144, 145, 146, 148, 149.

Since we cannot plot points in four dimensions a principal components analysis was used to display the data. Over 97% of the variability is captured in the first two dimensions, so a plot of these should give an accurate summary of the data. The results are shown in Figure 8. In this Figure, the symbols 1, 2 and 3 are used to show which of the Iris species each plant belongs to (1= Iris Setosa; 2 = Iris Versicolor; 3 = Iris Virginica). The boundaries of the groups are shown by the outlines. Note that while all of the Iris Setosa specimens belong to the same group, there is an overlap between the other two species, and there are some differences between the memberships of the other two groups suggested by CLUSTAN and the division between Iris Versicolor and Iris Virginica.

To discover whether these objects fall naturally into these three groups, the whole data set was divided (randomly as stated in subsection 2. 4. 2) into two subsets and each subset was clustered separately. The whole process was repeated ten times. We wanted to know if the same results could be obtained for the subsets. If similar results are indicated we are more confident there is genuinely structure in the data set. Firstly, a measure of classifiability for the entire data (M) was obtained. Secondly, a measure of each plant for how readily it could be classified (R_i); the plants that had values of R_i less than 1 are summarized in the table below.

(1) Measure of classifiability

Objects	R_i	%
51	5/10	50
52	8/10	80
53	7/10	70
55	6/10	60
57	6/10	60
66	7/10	70
73	4/10	40
77	4/10	40
78	4/10	40
84	5/10	50
86	8/10	80
87	7/10	70
114	8/10	80
115	4/10	40

120	5/10	50
122	8/10	80
124	5/10	50
127	3/10	30
128	5/10	50
135	7/10	70
139	7/10	70
143	6/10	60
147	4/10	40
150	7/10	70
	----	---
M	1400/1500	93.3

All other objects have reliability 10/10.

After further investigation of Figure 8 it was found that the objects which have smaller values of R_i are close to the boundaries between the groups. In general we are uncertain with all the results $\leq$ 9/10, but we are very uncertain about all the results $\leq$ 5/10 and these plants are: 51, 73, 77, 84, 115, 120, 124, 127, 128, 147; these objects are close to a boundary between groups. Figure 8 also makes it clear that there is a difference between Fisher's results and CLUSTAN results.

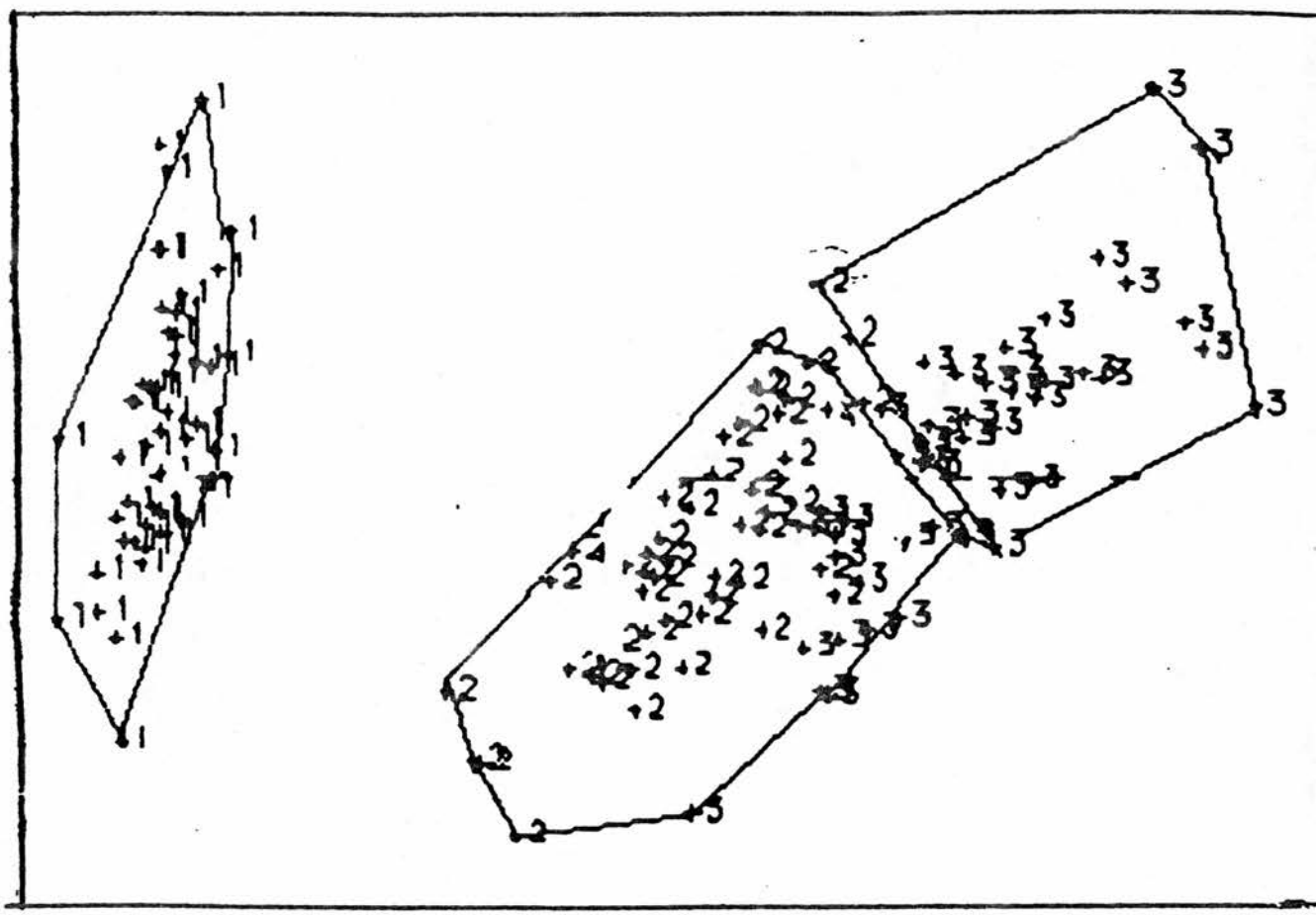

Figure 8

Divisions of the Iris data into three groups,illustrating the difference between CLUSTAN results(shown outlined) and FISHER results (shown by symbols 1, 2 and 3).

3. 7. 2. OLD FAITHFUL GEYSER DATA

Some data on the Old Faithful Geyser were chosen to be the second real data set. As shown in Azzalini and Bowman (1990), the data consist of 299 pairs of measurements, referring to the time interval between the starts of successive eruptions w_t and the duration of the subsequent eruption d_t. The analysis deals with data which were collected continuously from August 1st until August 15th, 1985. To understand the mechanism of the series Azzalini and Bowman considered the joint behaviour of the three variables(w_t, d_t, w_{t+1}). In our study after excluding the duration times which were recorded as L, S or M, the first 150 triples of measurements were chosen; it was believed that the main features of the data would be displayed in this size of data set. Azzalini and Bowman (1990) inspected some plots of the data and argued from these that there were three groups. In their Figure 3 which plots w_t against d_t, they draw boundaries approximately at $w_t = 68$ and $d_t = 3$. To test this, CLUSTAN was restricted to search for the best partition into three groups, the three groups were found to look like this:

A: 2, 5, 7, 9, 12, 16, 26, 28, 30, 32, 34, 36, 38, 40, 42, 48, 51, 53, 57, 63, 65, 67, 69, 71, 73, 75, 77, 79, 82, 86, 88, 90, 92, 94, 102, 113, 114, 116, 118, 120, 127, 129, 131, 135, 137, 139, 142, 144, 146, 148, 150.

B: 1, 4, 11, 15, 18, 19, 20, 21, 22, 23, 24, 25, 44, 45, 46, 50, 55, 56, 59, 60, 61, 62, 83, 84, 85, 96, 97, 98, 99, 100, 101, 104, 105, 106, 107, 108, 110, 111,122, 123, 124, 125, 126, 133, 134, 141.

C: 3, 6, 8, 10, 13, 14, 17, 27, 29, 31, 33, 35, 37, 39, 41, 43, 47, 49, 52, 54, 58, 64, 66, 68, 70, 72, 74, 76, 78, 80, 81, 87, 89, 91, 93, 95, 103, 109, 112, 115, 117, 119, 121, 128, 130, 132, 136, 138, 140, 43, 145, 147, 149.

To enable the data to be plotted and examined visually w_t was plotted against d_t in Figure 8, To examine whether these objects fall naturally into these three groups, the same approach as before was used; a measure of classifiability for the entire data (M) was obtained, and also a measure of how readily each reading could be classified (R_i). In the five attempts the results showed that there was perfect agreement except for three readings (73, 25, 113) as shown below:

(1) Measure of classifiability

Object	R_i	%
25	4/5	80
73	4/5	80
113	4/5	80

All other objects have reliability 5/5. Thus, the

(2) Measure of agreement M = 747/750= 99.6%

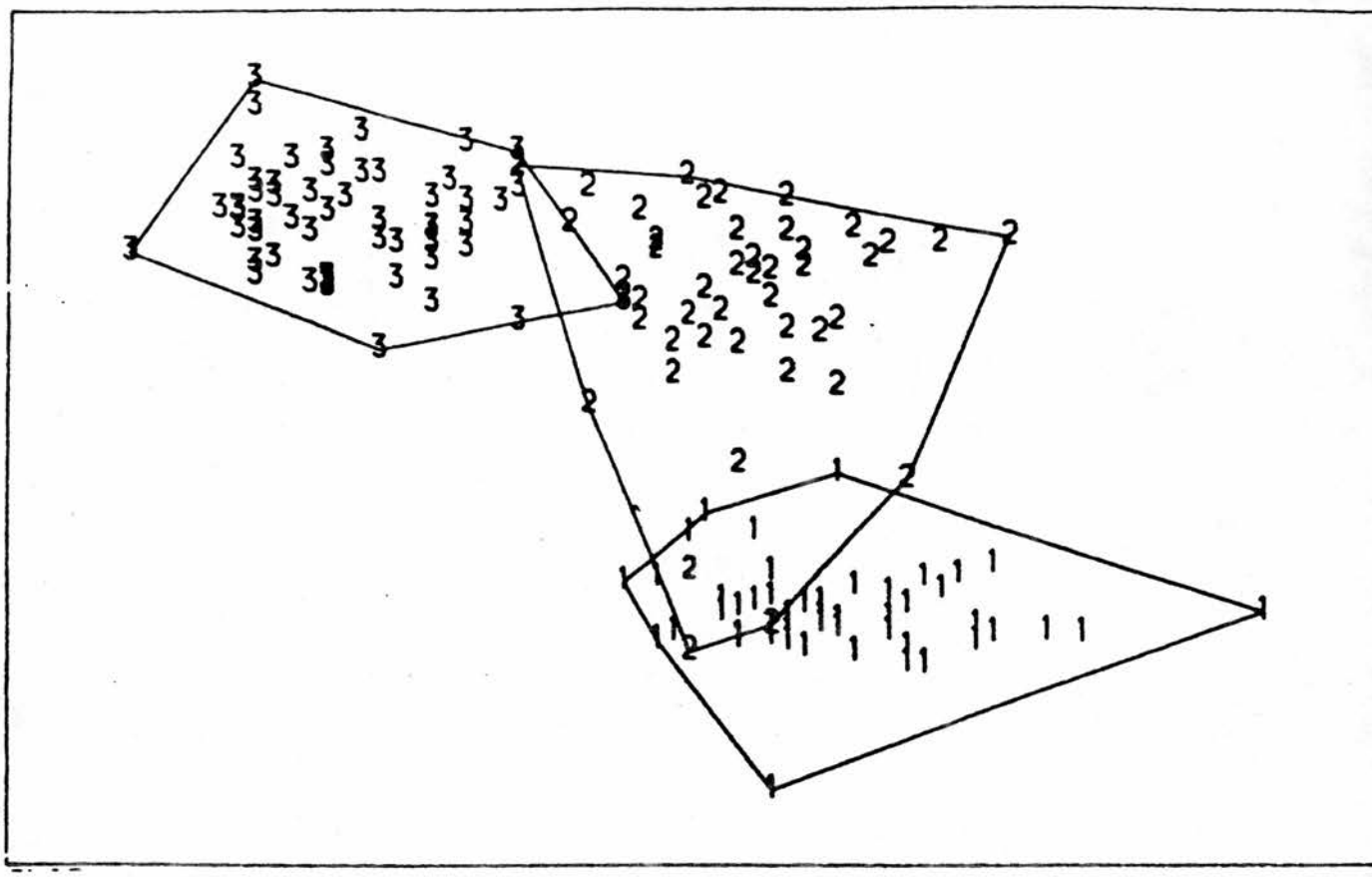

Figure 9

Illustrating the scatter of d_t

against w_t

A comparison of Figure 9 with Figure 10 (Figure 3 of Azzalini and Bowman (1990)) shows that very similar groupings were obtained(M takes the value 90.7%): the dotted lines on Figure 3 of Azzalini and Bowman are drawn at $w_t = 68$ and $d_t = 3$. Figure 11 plots the labels of points whose results appear unusual. However, differences on the third variable explain apparent anomalies in groupings, e.g. points 24, 47, 108 and 122 have higher values of w_{t+1} than their neighbours in the plot.

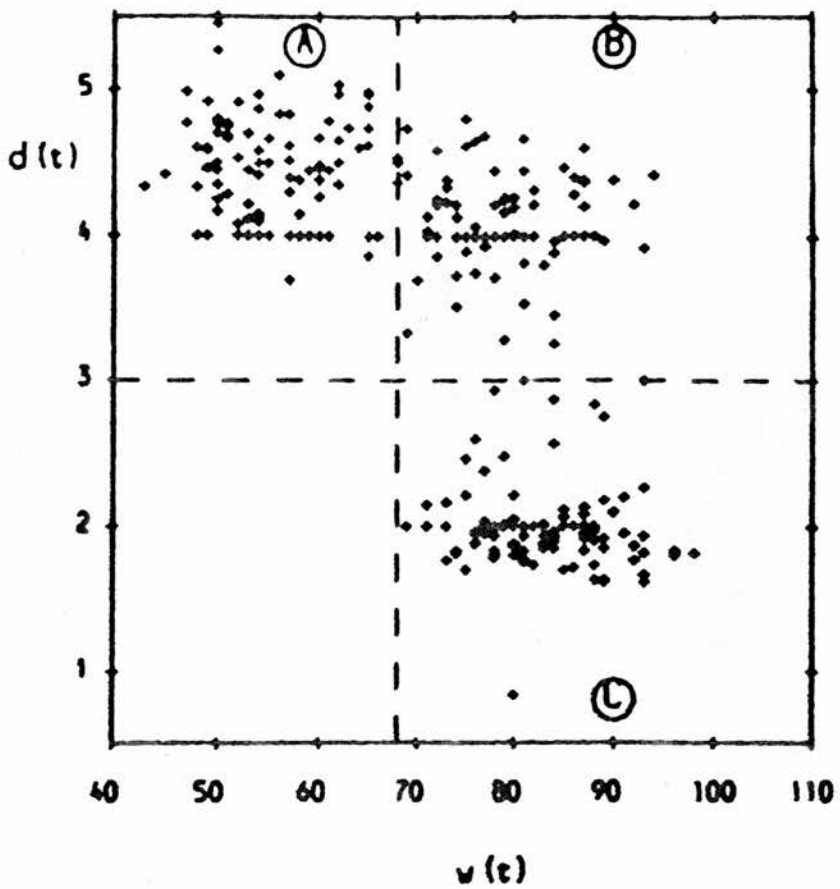

Figure 10

Azzalini and Bowman Figure 3

illustrating the scatter of d_t against w_t

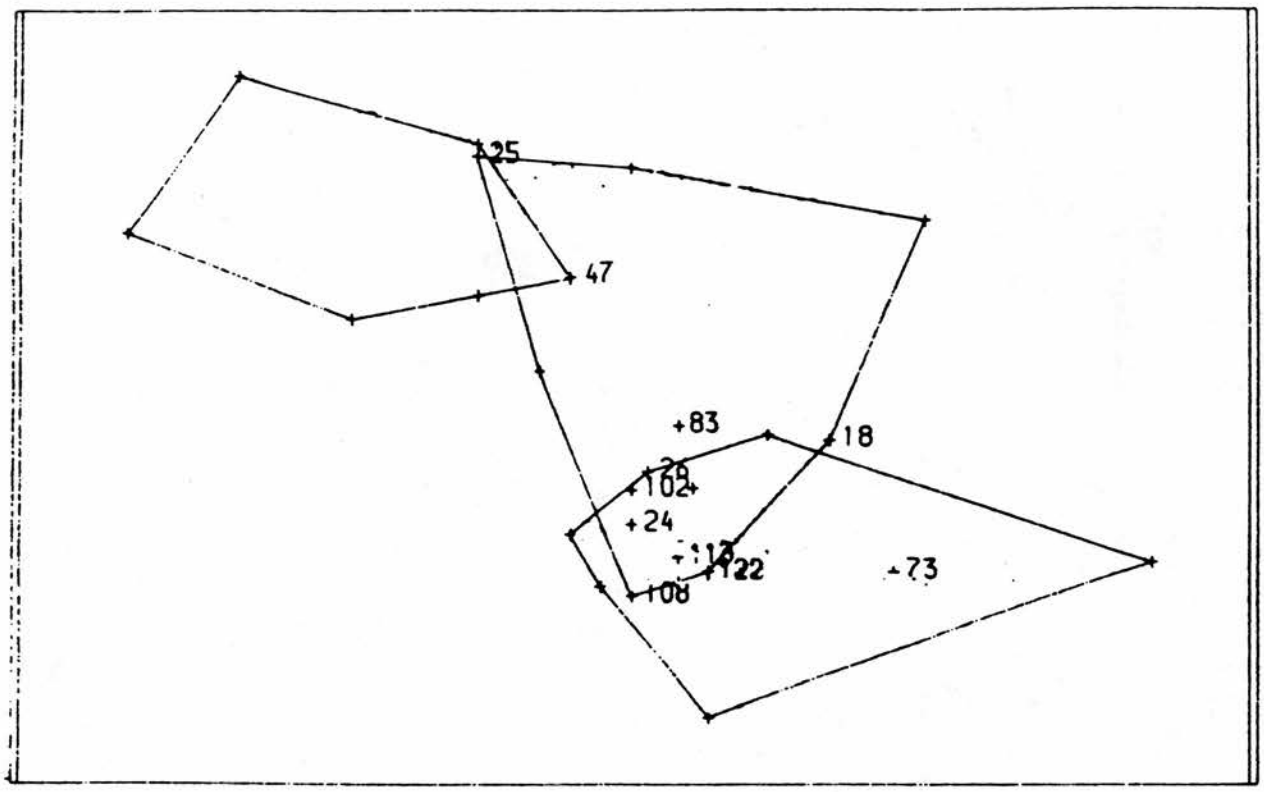

Figure 11

Illustrating the scatter of the overlapping readings

3. 7. 3. CONCLUSION

In the foregoing section, two real data sets were considered. Firstly the analysis was applied on the classical Fisher Iris data set. The principal components analysis was used, it shows that 97% of the variability is captured in the first two dimensions, which enables us to plot the data. The comparison shows that there was a difference between CLUSTAN results and Fisher results for the second and the third species. The disappointing results for the Iris data seem to be because the clustering criterion used was not the most appropriate one for these data. Overall, the results suggest that the method can be useful if a relevant clustering criterion is used.

Secondly, the analysis was applied on some data on the Old Faithful Geyser. Only one hundred and fifty triples of measurements were chosen. The obtained measure of agreement was 99.6%, and the comparison with Azzalini and Bowman (1990) shows very similar groupings. We can be confident that there are genuinely three groups in the data.

REFERENCES

(1) Azzalini, A. and Bowman, A. W. (1990), " A look at some data on the Old Faithful Geyser ". Appl. Statist., 39, pp. 357-365.

(2) Bailey, T.A. and Dubes, R.C. (1982),
" Cluster validity profiles ". Pattern Recognition, 15, pp. 61-83.

(3) Cormack, R.M. (1971), " A review of classification (with Discussion) ". J.R. Statistic Soc., A, 134, pp. 321-367.

(4) Dubes, R. and Jain, A.K. (1978), " Models and methods in cluster validity ". Proc. IEEE conf. on Patt. Recog. and Image Proc., pp. 148-155.

(5) Everitt, B. (1980), " Cluster analysis ". 2nd Edn, Heihemann Educational Books, London.

(6) Fisher, R.A. (1936), " The use of multiple measurements in taxonomic problems ". Annals of Eugenics 7, pp. 179-188.

(7) Forgy, E.W. (1965), " Cluster analysis of multivariate data: efficiency versus interpretability of classifications ". (abstract). Biometrics, 21, pp. 768-769.

(8) Fortier, J.J. and Solomon, H. (1966), " Clustering procedures ". in Proc. Symp. Multiv. Analysis, Dayton, Ohio (P.R. Krishnaiah, ed.) pp. 493-506. New York: Academic Press.

(9) Gordon, A.D. (1981), " Classification: Methods for the exploratory analysis of multivariate data ". London: Chapman & Hall.

(10) Gordon, A.D. (1987), " A review of hierarchical classification ". J.R. Statistic Soc. A, 150, pp.119-137.

(11) Gordon, A.D. and De Cata, A. (1988), " Stability and influence in sum of squares clustering ". Metron, 46, pp. 348-360.

(12) Gordon, A.D. and Henderson, J.T. (1977), " An algorithm for Euclidean sum of squares classification ". Biometrics, 33, pp. 355-362.

(13) Gower, J.C. (1985), " Measures of similarity, dissimilarity, and distance ". Encyclopedia of Statistical Sciences (S.Kotz, N.L.Johnson and C. B. Read, eds), 5, pp. 397-405. New York: Wiley.

(14) Hubert, L. (1974), " Approximate evaluation techniques for the single-link and complete-link hierarchical procedures ". Journal of the American Statistical Association, 69, pp. 698-704.

(15) Kendall, M. and Stuart, A. (1976), " The advanced theory of statistics ". Volume 3 (3rd Edn) . London:Griffin.

(16) Jancey, R.C. (1966), " Multidimensional group analysis ". Austral. J. Botany, 14, pp. 127-130.

(17) MacQueen, J. B. (1967), " Some methods for classification and analysis of multivariate observations ". Proceedings of 5th Berkeley Symposium, 1, pp. 281-297.

(18) Smith, S.P. and Dubes, R. (1980), " Stability of hierarchical clustering ". Pattern Recognition, 12, pp.177-187.

(19) Sokal, R.R. and Michener, C.D. (1958), " A statistical method for evaluating systematic relationships ". The University of Kansas Science Bulletin, 28, pp. 1409-1437.

(20) Strauss, J.S., Bartko, J. J. and Carpenter Jr., W. J. (1973), " The use of clustering techniques for classification of psychiatric patients ". British J. of Psychiatry, 122, pp. 531-540.

(21) Wishart. D. (1987), " CLUSTAN User Manual (Fourth Edition) ". Computing Laboratory, University of St. Andrews.