

University of Bath



**PHD**

**Multidimensional scaling: A simulation study and applications in politics, ethnology, taxonomy and nutrition.**

Osmond, Clive

*Award date:*  
1982

*Awarding institution:*  
University of Bath

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

**Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 22. May. 2019

MULTIDIMENSIONAL SCALING:

A SIMULATION STUDY AND APPLICATIONS

IN POLITICS, ETHNOLOGY, TAXONOMY AND NUTRITION

submitted by Clive Osmond

for the degree of Ph.D.

of the University of Bath

1982

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

... *Clive Osmond* ...

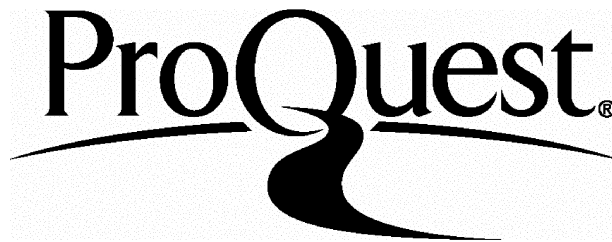
ProQuest Number: U333128

All rights reserved

INFORMATION TO ALL USERS

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

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



ProQuest U333128

Published by ProQuest LLC(2015). Copyright of the Dissertation is held by the Author.

All rights reserved.

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

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

UNIVERSITY OF BATH		
LIBRARY		
22	14 JUN 1982	
PDD		

# C O N T E N T S

	<u>PAGE</u>
Acknowledgements .. .. .	i
Summary .. .. .	ii
Chapter 1: FOUNDATIONS: AN INTRODUCTION TO TECHNIQUES AND THEORETICAL RESULTS .. .. .	1
1 Introduction .. .. .	2
2 Procrustes Analysis .. .. .	3
3 Classical Scaling .. .. .	6
4 Ordinal Scaling .. .. .	9
5 Least Squares Scaling .. .. .	15
6 Example .. .. .	17
7 Preprocessing the Dissimilarity Matrix .. .. .	24
8 Single-Link Clustering .. .. .	26
9 Partition Likelihood Clustering .. .. .	27
10 Principal Component Analysis .. .. .	30
11 The Generation of Dissimilarities .. .. .	31
12 Choice of Dimensionality in Classical Scaling .. .. .	36
13 Perturbational Analysis of Procrustes Statistics .. .. .	40
14 References .. .. .	43
Chapter 2: PREVIOUS SIMULATION STUDIES OF MULTIDIMENSIONAL SCALING .. .. .	47
1 Introduction .. .. .	48
2 The Founding Fathers: Shepard and Kruskal .. .. .	51
3 Robustness of Ordinal Scaling .. .. .	53
4 Random Rankings .. .. .	62
5 The Choice of Starting Configuration .. .. .	64
6 The Relative Importance of the Small, Medium and Large Dissimilarities .. .. .	65
7 The Contribution of Lingoes and Roskam .. .. .	67
8 The Recent Contribution of Shepard .. .. .	69
9 A Summary and our Approach .. .. .	74
10 References .. .. .	84
Chapter 3: FOUR SIMULATION STUDIES OF MULTIDIMENSIONAL SCALING	
1 Euclidean Models for the Generation of Similarities .. .. .	88
2 Simulation Studies of Classical Scaling .. .. .	102
3 Comparison of Scaling Methods .. .. .	123
4 Simulations on Scaling Subsets of Similarities .. .. .	128
5 Procrustes Statistics Arising from Slightly Different Configurations .. .. .	136
6 Procrustes Statistics Arising from Slightly Different Squared Distance Matrices .. .. .	144
7 References .. .. .	148

/Continued ...

	<u>PAGE</u>
Chapter 4: AN APPLICATION IN BRITISH POLITICAL HISTORY ..	149
1 Introduction: Motivation for the Project .. ..	150
2 The Extent of Earlier Statistical Analysis of Voting .. .. .	154
3 Acquiring and Assembling the Data .. ..	157
4 A Preliminary Approach .. .. .	168
5 The Measurement of Dissimilarity .. ..	176
6 Definitions of Groups Used in the Analyses ..	178
7 Results Obtained by Single-Link Clustering ..	180
8 Results Obtained by Ordinal Scaling .. ..	183
9 Results Obtained by Least Squares Scaling ..	198
10 Economies in Use of Similarities .. ..	201
11 Conclusions .. .. .	204
12 References .. .. .	206
Chapter 5: AN APPLICATION IN LINGUISTICS AND ETHNOLOGY ..	208
1 Introduction .. .. .	209
2 Materials and Methods .. .. .	210
3 Results .. .. .	219
4 Conclusions .. .. .	229
5 References .. .. .	232
Chapter 6: AN APPLICATION IN TAXONOMY .. .. .	233
1 Introduction .. .. .	234
2 Preliminaries: Selection and Measurement of Specimens .. .. .	239
3 The Measurement of Similarity .. .. .	245
4 Results Obtained from Single-Link Clustering ..	250
5 Results Obtained from Ordinal Scaling .. ..	252
6 Results Obtained by Partition Likelihood Clustering	255
7 Discussion and Conclusions .. .. .	263
8 References .. .. .	265
Chapter 7: AN APPLICATION IN NUTRITION .. .. .	266
1 Introduction to the Data .. .. .	267
2 Results Obtained by Multidimensional Scaling ..	272
3 References .. .. .	281
Appendix: Index to Figures and Tables .. .. .	282

ACKNOWLEDGEMENTS

The appearance of this thesis owes much to the help and encouragement received from many sources:

First and foremost my supervisor, Professor Robin Sibson, who introduced me to the subject area and has continually guided and encouraged when I have struggled.

My wife, Margaret, who has typed the thesis with expertise and care, and been a source of strength, persuading me to press on with the writing.

The collaborators with whom I worked to produce the applications described in the last four chapters: Andrew Baring, John Dawes, Michael Nelson and especially Valerie Cromwell for her consistent interest.

Other staff and students at the University of Bath: Adrian Bowyer, Peter Green, Wyn Lotwick, Bernard Silverman, Steve Langron, Graeme Thomson and the Computing Service.

The Medical Research Council Environmental Epidemiology Unit which has kindly provided facilities for typing the thesis.

The Social Science Research Council which provided financial support during my research studentship.

Other friends who have shown discretion in asking about my progress, balanced with persistence to keep me at it.

To them all I give my grateful thanks.

## SUMMARY

This thesis has three sections. Section one contains two chapters, the first describing those techniques used later, principally multidimensional scaling, procrustes fitting and cluster analysis. Least squares scaling, preprocessing the dissimilarity matrix and clustering by maximum likelihood partition are less known. The second chapter reviews simulation studies previously published in multidimensional scaling literature.

Section two contains one chapter detailing four simulation studies in multidimensional scaling. The first considers the robustness of classical scaling in the presence of error in the dissimilarity matrix. Four probabilistic models generating euclidean-distance-like dissimilarity functions are proposed, which reflect some of the ways dissimilarities actually arise, and allow dependence between dissimilarities to be studied. Next we compare how well various scaling methods reconstruct specific configurations, given the same dissimilarity matrix. Properties of preprocessing the matrix and least squares scaling are demonstrated. Thirdly we describe a study, designed to measure the redundancy in a dissimilarity matrix, which justifies subsequent use of scaling with missing data. Finally we determine the robustness of approximations to procrustes statistics obtained from perturbational analysis of classical scaling by Sibson (1979).

Section three contains four applications chapters. Firstly multidimensional scaling is applied to data concerning the voting behaviour of M.P.s in 1861. This large data set requires special handling, some dissimilarity values being best treated as unknown. The results identify both unusual and regular voting behaviour.



The second application is in ethnology. Dissimilarity values derived from phonetic differences between languages are used to derive their genetic origin. The techniques, especially clustering by maximum likelihood partition, reproduce known relationships satisfactorily and suggest others. The third example uses morphological and meristic parameters to generate dissimilarities between specimens of the fish species *Colisa*. Here the aim is taxonomic. Finally we consider dietary changes across Britain through time to identify regional and temporal differences.

---

Reference

SIBSON, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41, pp. 217-229.

C H A P T E R   O N E

FOUNDATIONS: AN INTRODUCTION TO TECHNIQUES AND THEORETICAL RESULTS

	<u>PAGE</u>
1.1    Introduction            .. .. .	2
1.2    Procrustes Analysis    .. .. .	3
1.3    Classical Scaling        .. .. .	6
1.4    Ordinal Scaling         .. .. .	9
1.5    Least Squares Scaling .. .. .	15
1.6    Example                 .. .. .	17
1.7    Preprocessing the Dissimilarity Matrix .. .. .	24
1.8    Single-Link Clustering    .. .. .	26
1.9    Partition Likelihood Clustering .. .. .	27
1.10   Principal Component Analysis    .. .. .	30
1.11   The Generation of Dissimilarities    .. .. .	31
1.12   Choice of Dimensionality in Classical Scaling .. .. .	36
1.13   Perturbational Analysis of Procrustes Statistics	40
1.14   References             .. .. .	43

## 1.1 Introduction

The aim of this chapter is to specify systematically the range of techniques used in the subsequent simulation studies and applications reported in this thesis, in order that a notation may be established, that appropriate recognition be given to former work and that following chapters may refer to a unified treatment of these topics. The last two sections of the chapter refer to theoretical results which are examined in this thesis to determine their usefulness in practice. We also give a simple demonstration of the scaling techniques as they are applied to reconstructing the position of 48 British towns from their distances apart by road.

## 1.2 Procrustes Analysis

The technique of procrustes analysis is well established and there are many references in the literature, principal ones being Mosier (1939), Green (1952), Ahmavaara (1957), Hurley and Cattell (1962), Cliff (1966), Schönemann (1966, 1968), Gruvaeus (1970), Schönemann and Carroll (1970), Gower (1971<sup>b</sup>, 1975), Krzanowski (1971), Kristof and Wingersky (1971) and Sibson (1978). Its early applications were in factor analysis, but recently its relevance to multidimensional scaling has been recognised.

The problem that it deals with is conceptually very straightforward. Given two configurations of points in a space of  $K$  dimensions, with a preassigned correspondence between the points of the two configurations, how similar are the configurations? To answer this question we match the configurations under a specified group of transformations, the group being chosen to be appropriate in the context of the analysis at hand. Some possible groups are the Euclidean group  $E(K)$ , the similarity group  $S(K)$ , the affine group  $A(K)$  and the special Euclidean group  $SE(K)$ . Their properties may be summarised as follows:

	$SE(K)$	$E(K)$	$S(K)$	$A(K)$
Translation	✓	✓	✓	✓
Rotation	✓	✓	✓	✓
Reflection	X	✓	✓	✓
Dilatation	X	X	✓	✓
Shear	X	X	X	✓

We only consider the groups  $E(K)$  and  $S(K)$ . The matching process is procrustes analysis, the residual sum of squared

distances between points, which is minimised, is the procrustes statistic.

A K-dimensional configuration of N points is represented by a K x N matrix Y, where the ordering of the columns labels the points. Thus the sum of squared distances between two configurations Y and Z is defined by

$$G(Y,Z) = \text{tr} (Y - Z)^T(Y - Z) \quad (1.2.1)$$

When we allow matching under the Euclidean group E(K) we obtain

$$G_E(Y,Z) = \inf \{G(Y,\phi Z) : \phi \in E(K)\} \quad (1.2.2)$$

as the procrustes statistic.

Correspondingly

$$G_S(Y,Z) = \inf \{G(Y,\phi Z) : \phi \in S(K)\} \quad (1.2.3)$$

is the procrustes statistic obtained by matching from the similarity group S(K).

The algebra associated with matching under E(K) and S(K) is most conveniently presented in Sibson (1978), in which the author combines accuracy with simplicity, two features that are often missing in earlier work. We summarise the results.

Let  $Y_0$  and  $Z_0$  be the configurations Y and Z translated to have centroid at origin. Then

$$G_E(Y,Z) = \text{tr} Y_0 Y_0^T + \text{tr} Z_0 Z_0^T - 2 \text{tr} (Z_0 Y_0^T Y_0 Z_0^T)^{\frac{1}{2}} \quad (1.2.4)$$

$$G_S(Y,Z) = \frac{\text{tr } Y_0 Y_0^T - \{ \text{tr}(Z_0 Y_0^T Y_0 Z_0^T) \}^{\frac{1}{2}}}{\text{tr}(Z_0 Z_0^T)} \quad (1.2.5)$$

This last form allows the construction of a symmetric, scale-free standardisation of the procrustes statistic by division by  $\text{tr } Y_0 Y_0^T$ . We thus define

$$\gamma_S(Y,Z) = 1 - \frac{\{ \text{tr}(Z_0 Y_0^T Y_0 Z_0^T) \}^{\frac{1}{2}}}{(\text{tr } Z_0 Z_0^T)(\text{tr } Y_0 Y_0^T)} \quad (1.2.6)$$

The two steps of matching under translation and dilation are computationally straightforward. Matching under orthogonal transformation requires an eigenvalue/vector calculation for a symmetric matrix, and can be conveniently solved by using the NAG subroutine FO2ABF.

### 1.3 Classical Scaling

Classical scaling is an algebraic technique for reconstructing a configuration of points from its interpoint distances. The appropriate algebra first appeared in Young and Householder (1938), but it was Torgerson (1952, 1958) who developed the statistical application. Classical scaling was independently derived by Gower (1966) who describes it as principal coordinates analysis. A succinct account of the algebra can be found in Sibson (1979), which we follow in order to establish a notation and terminology.

As in the previous section, a  $K$ -dimensional configuration of  $N$  points is represented by a  $K \times N$  matrix  $Y$ , where the ordering of the columns labels the points. A configuration  $Y$  has its centroid at the origin if and only if  $Y\underline{1}_N = \underline{0}_K$ , where  $\underline{1}_N$  is the  $N$ -vector of 1's and  $\underline{0}_K$  is the  $K$ -vector of 0's. We define the inner product matrix  $b(Y) = (b_{ij})$  as the  $N \times N$  matrix  $Y^T Y$ . This is the matrix of inner products of the coordinate vectors of the points in the configuration.  $b(Y)$  is symmetric, positive-semidefinite, and has the same rank as  $Y$ . The centroid at origin condition  $Y\underline{1}_N = \underline{0}_K$  is equivalent to  $b(Y)\underline{1}_N = \underline{0}_N$ . We define the squared distance matrix  $e(Y) = (e_{ij})$  by

$$e_{ij} = b_{ii} + b_{jj} - 2b_{ij} \quad (1.3.1)$$

which is the familiar 'cosine rule'.

Thus we have defined a progression

$$Y \rightarrow b(Y) \rightarrow e(Y).$$

The aim of classical scaling is to invert this procedure, and again two simple steps are possible. Firstly, given a matrix E of squared euclidean distances, the linear transformation

$$B = q(E) = -\frac{1}{2} \left\{ I - \frac{1}{N} \frac{1 \cdot 1^T}{N} \right\} E \left\{ I - \frac{1}{N} \frac{1 \cdot 1^T}{N} \right\} \quad (1.3.2)$$

produces a corresponding inner product matrix which satisfies the centroid-at-origin condition  $B \underline{1}_N = \underline{0}_N$ . Secondly we recover Y from B as follows. Let  $\underline{e}_1, \dots, \underline{e}_K, \underline{e}_{K+1}, \dots, \underline{e}_N$  be an orthonormal basis of eigenvectors of B with corresponding eigenvalues

$$\lambda_1 > \dots > \lambda_K > 0 = \lambda_{K+1} = \dots = \lambda_N.$$

$$\text{Thus } B = \sum_{k=1}^K \lambda_k \underline{e}_k \underline{e}_k^T.$$

Then  $Y^T = \{\sqrt{\lambda_1} \underline{e}_1, \dots, \sqrt{\lambda_K} \underline{e}_K\}$  defines a configuration Y that agrees with all of the squared distances given in E, and is represented relative to principal axes.

In an application we derive a symmetric matrix of positive distances or dissimilarities for our starting point. From this we can obtain the matrix of squared distances, E.  $q(E)$  will then be symmetric and we can extract its eigenvalues and eigenvectors. When we desire a K-dimensional solution configuration, we use the K largest positive eigenvalues and their associated eigenvectors in an attempt to produce a configuration whose squared interpoint distance matrix is an approximation to the matrix E. Inevitably there will be inaccuracy, the level of which is determined by the extent to which there are substantial positive eigenvalues beyond the Kth and by the number and size of the negative eigenvalues. Often we shall refer to the eigenvalue spectrum. Problems of determining the appropriate dimensionality in a particular application are discussed in Section 1.12.



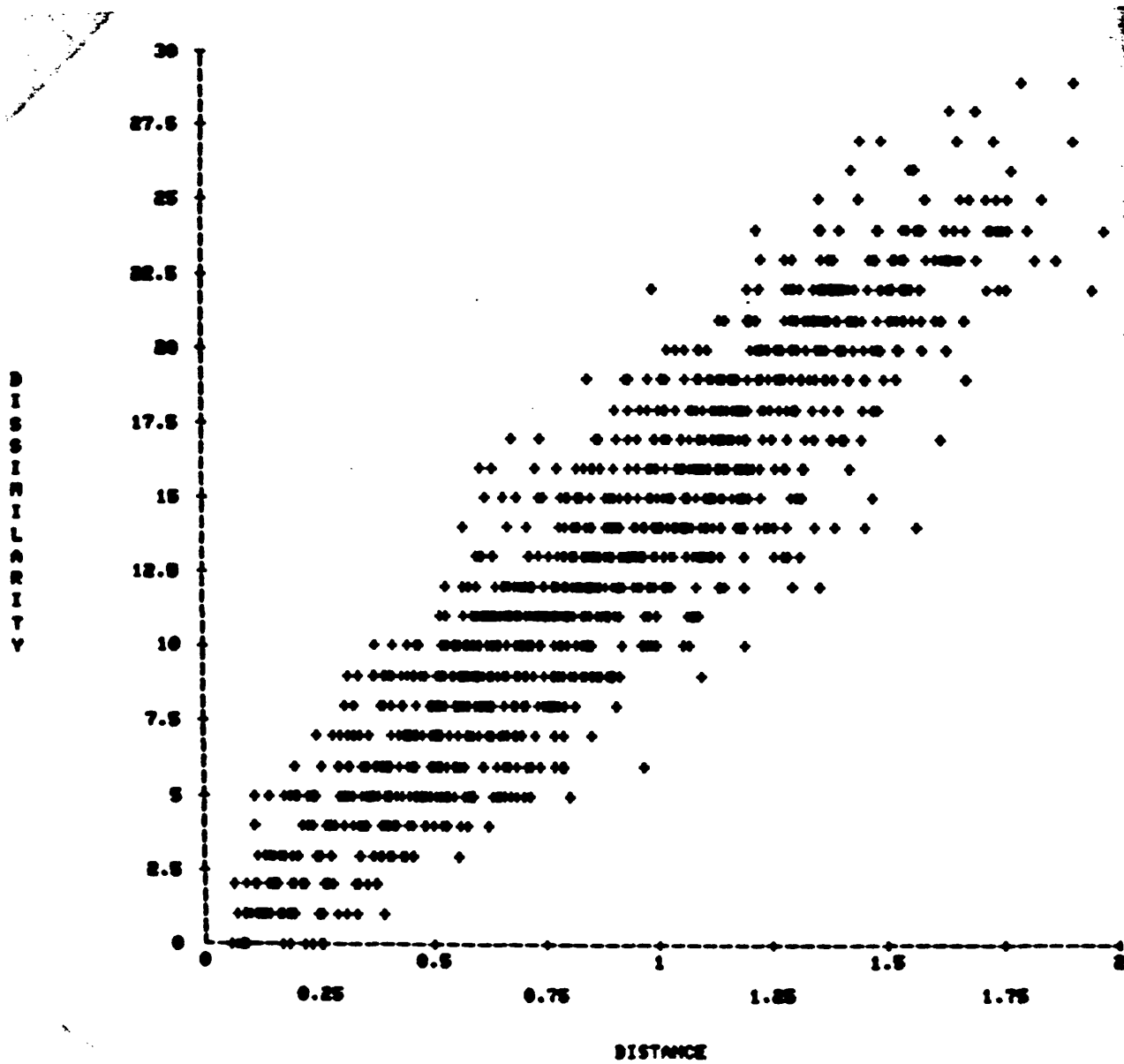
An extension of ordinary classical scaling has been introduced by Critchley (1978, 1980). Rather than using just the dissimilarities themselves to derive the squared distance matrix  $E$ , he extends this step by allowing a monotonic increasing function of the dissimilarities with a variable parameter. The selection of a final configuration then requires optimisation over the range of parameter values. We do not pursue this approach.

#### 1.4 Ordinal Scaling

Ordinal scaling is the name we give to the method of non-metric multidimensional scaling described by Kruskal (1964 a, b). Classical scaling depends upon the assumption that observed dissimilarities between pairs of objects are Euclidean, or at least nearly-Euclidean. Only then can we hope to derive a configuration of points that will successfully approximate the dissimilarity matrix. This is a point that we consider in greater depth in Chapter 3. However it is clear that this assumption is far too restrictive for many sets of data. Shepard (1962 a, b) suggested that the minimum sensible assumption that we should make of our data was that the ordinal properties of the dissimilarities alone, and not their algebraic values, should be considered significant in any method constructed for their analysis. This idea is of course equivalent to the assumption that the given data represent some arbitrary order-preserving transformation of a set of true Euclidean distances. For realism it is necessary to include the possibility of error added to the monotone transformation so that we must seek a monotone function best fitting the data in some sense. That this has been done is a tribute to the work of Kruskal, who recognised that a bridge between monotonic functions, defined on ordered pairs of points, and interpoint distances could be found by using the efficient techniques of least squares monotone regression. An example of a graph showing dissimilarities between objects plotted against their interpoint distances in a configuration is given in Fig. 1.4.1. The dissimilarity function takes integral values in the range 0 to 30 inclusive, and the greatest interpoint distance is just under 2.0. Kruskal showed how to relate such a set of distances and the ordering imposed upon the dissimilarity values.

FIG. 1.4.1 A Plot of Dissimilarity Values Against Configuration

Distance



We now describe his solution to the problem and simultaneously develop our own terminology.

Let there be  $N$  objects and let us seek a configuration lying in  $K$  dimensions. Let the dissimilarity matrix  $\Delta = (\delta_{ij})$  be known so that we can order the dissimilarity values as

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \dots \leq \delta_{i_M j_M}$$

where  $M = N^2$  is the total number of elements in  $\Delta$ . Let our current configuration be  $Y = (\underline{y}^{(1)}, \dots, \underline{y}^{(N)})$ , so that we can derive the interpoint distance matrix  $D = (d_{ij})$  as

$$d_{ij} = \left\{ \sum_{k=1}^K (y_k^{(i)} - y_k^{(j)})^2 \right\}^{\frac{1}{2}} \quad (1.4.2)$$

Then we can define the least squares monotone regression

$\hat{D} = (\hat{d}_{ij})$  of  $D$  on  $\Delta$  to be the matrix with the properties that

- (i)  $\delta_{ij} < \delta_{kl} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{kl} \quad 1 \leq i, j, k, l \leq N$
- (ii)  $\sum_{i=1}^N \sum_{j=1}^N (d_{ij} - \hat{d}_{ij})^2$  is a minimum, where  $Y$  is fixed, but  $\hat{D}$  varies.

We may now define

$$S^*(Y) = \sum_{i=1}^N \sum_{j=1}^N (d_{ij} - \hat{d}_{ij})^2, \quad (1.4.3)$$

$$T^*(Y) = \sum_{i=1}^N \sum_{j=1}^N d_{ij}^2 \quad (1.4.4)$$

and  $S(Y) = \sqrt{\frac{S^*(Y)}{T^*(Y)}} \quad (1.4.5)$

Then  $S(Y)$  can be used as a measure of departure from a perfect fit to the monotonicity hypothesis and is called the stress of the configuration. An optimal configuration minimises  $S(Y)$ . The stress function is invariant under transformations to the configuration by elements of the similarity group.

Elegant and efficient algorithms are available for the least squares monotone regression, and the other computational problem, namely how to minimise stress, is made practicable by the existence of a continuous first derivative of the step function. Thus the method of steepest descent may be used, although the selection of step size for successive iterations requires considerable attention to make the method efficient, and Kruskal's original recommendations are far from obvious since they are based upon considerable empirical experience. McGinley (1977) has explored other minimisation algorithms but none seem better than steepest descent. Thus an iterative procedure may be established. We generate an initial configuration and then successively compute  $D$ ,  $\hat{D}$ ,  $S$  and the derivatives of  $S$  with respect to the  $NK$  coordinate values, until some convergence criterion is satisfied.

Many refinements may be made. Often there will only be dissimilarity values in the subtriangle of the matrix, which is assumed to be symmetric with zeros on the diagonal. In this case we take  $M = \frac{1}{2}N(N - 1)$  and the various ranges of summation may be modified accordingly. Likewise modifications may be introduced to allow for missing values, and even more generally weights may be assigned to the elements of the dissimilarity matrix. The ordinary least squares monotone regression programs may easily be amended to deal with weighting.

Many different metrics are available for use in the calculation

of the interpoint distances, apart from the usual Euclidean distance. For example, much has been made of the use of the family of Minkowski metrics.

Our present definition of the least squares monotone regression is based on a complete, global ordering of the dissimilarity values, and is such that discrepancies between regression values corresponding to equal dissimilarity values are ignored. Two straightforward adjustments permit refinements to these formulations. Firstly we may treat merely local orderings of the dissimilarity values, making comparisons only for those dissimilarity values which are derived from pairs of objects with at least one object in common. Thus they must have a common endpoint. Secondly we may adopt what is referred to as the 'secondary treatment of tied values' which imposes an extra condition on the definition of the least squares monotone regression, namely

$$(iii) \quad \delta_{ij} = \delta_{kl} \Rightarrow \hat{d}_{ij} = \hat{d}_{kl} \quad 1 \leq i, j, k, l \leq N$$

Several normalisation factors other than  $T^*(Y)$  have been tried in the definition of the stress function.

The steepest descent technique is prone to converge to what are merely local optima of the stress function. We shall investigate this in Chapter 3 and show that a sensible choice of initial configuration is most important. It is common to seek the optimal solution by using several randomly generated initial configurations, and by using the configuration obtained by classical scaling, possibly after the dissimilarity values have been transformed under some distributional assumption, as explained in Section 1.7. An alternative method is to use the principal axis solution projected

down from higher dimensional ordinal scaling solutions.

There are certain usages common to the rest of this thesis. Throughout, we use the primary treatment of ties in which ties are an expression of ignorance and can be broken without charge. We use the global ordering of dissimilarity values, and we use the original normalisation of stress by  $T^*(Y)$ . We are very wary of the use of non-Euclidean distance measures, for a point in a non-Euclidean space has the extra structure related to its position relative to the coordinate axes. Shepard (1974) makes this point well, and we take his warning and use only Euclidean distance. Reported stress values have been obtained by using the variety of available initial configuration techniques, so that we are confident that local minima do not vitiate our study.

### 1.5 Least Squares Scaling

Least squares scaling is a method that is not commonly used for scaling dissimilarity data. Certainly there are few references in the multidimensional scaling literature. This is perhaps surprising, because the method has some intuitive appeal and simplicity. The aim is to find a  $K \times N$  configuration matrix,  $Y$ , which minimises

$$\sum_{i=1}^N \sum_{j=1}^N w_{ij} (d_{ij} - \delta_{ij})^2 \quad (1.5.1)$$

where  $D = (d_{ij})$  is the achieved inter-point distance matrix of  $Y$

$\Delta = (\delta_{ij})$  is the known dissimilarity matrix

and  $W = (w_{ij})$  is a constant matrix of weights.

We may relate this to a specific statistical model: if the errors by which the  $\delta_{ij}$  differ from the  $d_{ij}$  are  $N(0, 1/w_{ij})$  and independent, then we are carrying out maximum likelihood estimation.

The method has a mixture of the properties of ordinal and classical scaling. It is similar to ordinal scaling in that the user has to define the dimensionality of his solution configuration, in that an initial configuration must be provided, often from classical scaling, and in that there is the need to minimise an objective function by an iterative procedure. The similarity with classical scaling lies in the significance attached to the actual dissimilarity values. The optimisation problem appears to be considerably better behaved than that arising from ordinal scaling. The Fletcher-Reeves algorithm for function minimisation by conjugate gradients seems to handle it most successfully.

We have discovered five references to least squares scaling.



In a paper written to examine the appropriateness of the 'minimum sensible assumption' of Shepard, namely that only ordinal properties of dissimilarities should be considered, Spaeth and Guthery (1969) claim to discover theoretical and practical shortcomings in the assumption, which lead them to mention the least squares criterion as a possible alternative. However there is no indication that they have considered any actual method based upon it. Sammon (1969) implemented the special case of least squares scaling in which the weights are given by the inverses of the corresponding dissimilarity values, that is  $w_{ij} = 1/\delta_{ij}$ , a method that he called non-linear mapping. Anderson (1971) also considers the idea, but gives no indication of having implemented a practical method. Chang and Lee (1973) adapt the steepest descent algorithm used by Sammon to operate only on pairs of objects at a time. Bloxom (1978) discusses a related but more complicated least squares method requiring a special computational algorithm.

Least squares scaling is particularly well suited to the minority of applications in which it is appropriate to assume that the observed dissimilarities differ from the true interpoint distances by errors that are independent. Such cases do arise, for example, in photogrammetry and surveying. However we shall demonstrate that the method can also be effective in more general applications.

## 1.6 Example

A readily available dissimilarity matrix may be found in the pages of the yearly A.A. Handbook, where road distances between all pairs of 48 British towns are presented in a triangular array corresponding to the subdiagonal of the matrix. These road distances are approximations to the true distances between the towns measured "as the crow flies", but there will be an error term added to each true distance whose magnitude will depend upon how direct the route may be, and upon the true distance. The true configuration of the towns is readily available from maps, and may be digitised very accurately so that we can use procrustes analysis to compare the true configuration with the configurations generated by classical scaling, ordinal scaling and least squares scaling. Thus we have a simple example of all the techniques that have been introduced so far.

### (a) Classical Scaling

We would expect the two-dimensional solution configuration to be quite accurate because the dissimilarities are close to linearly related with the distance in the configuration. This is indeed the case. The eigenvalue spectrum is given in Table 1.6.1. The percentage of the sum of the nine most positive eigenvalues attributable to each of those eigenvalues individually may be seen to be:

74, 13, 4, 3, 2, 1, 1, 1, 1.

If we did not know that the underlying configuration was two-dimensional we would suspect either a one- or a two-dimensional solution. The third eigenvalue is not much greater than that of the remainder and is not much greater than the magnitude of the most negative eigenvalue, indicating that it contains noise. Also the trace criterion (see Section 1.12) would suggest that no more than three dimensions are appropriate, the third being debatable. Those towns which have large

TABLE 1.6.1

<u>EIGENVALUE</u> <u>NO.</u>	<u>EIGENVALUE</u>	<u>SUM OF VALS.</u>
1	1.278E+06	1.278E+06
2	2.207E+05	1.499E+06
3	7.594E+04	1.574E+06
4	5.569E+04	1.630E+06
5	3.061E+04	1.661E+06
6	2.547E+04	1.686E+06
7	1.483E+04	1.701E+06
8	1.002E+04	1.711E+06
9	9.033E+03	1.720E+06
10	6.054E+03	1.726E+06
11	4.678E+03	1.731E+06
12	3.216E+03	1.734E+06
13	2.186E+03	1.736E+06
14	1.878E+03	1.738E+06
15	1.561E+03	1.740E+06
16	1.229E+03	1.741E+06
17	1.197E+03	1.742E+06
18	1.044E+03	1.743E+06
19	7.666E+02	1.744E+06
20	5.405E+02	1.744E+06
21	4.985E+02	1.745E+06
22	3.196E+02	1.745E+06
23	2.940E+02	1.746E+06
24	2.067E+02	1.746E+06
25	1.869E+02	1.746E+06
26	6.534E+01	1.746E+06
27	1.270E-10	1.746E+06
28	-5.901E+01	1.746E+06
29	-9.729E+01	1.746E+06
30	-2.250E+02	1.746E+06
31	-3.472E+02	1.745E+06
32	-4.283E+02	1.745E+06
33	-5.493E+02	1.744E+06
34	-7.640E+02	1.744E+06
35	-8.384E+02	1.743E+06
36	-9.998E+02	1.742E+06
37	-1.276E+03	1.740E+06
38	-1.693E+03	1.739E+06
39	-2.101E+03	1.737E+06
40	-2.616E+03	1.734E+06
41	-2.980E+03	1.731E+06
42	-3.483E+03	1.728E+06
43	-3.892E+03	1.724E+06
44	-5.430E+03	1.718E+06
45	-7.782E+03	1.710E+06
46	-1.494E+04	1.696E+06
47	-5.242E+04	1.643E+06
48	-5.936E+04	1.584E+06

coordinates in the third dimension tend to be extremal. For example, many road distances to Penzance are greater than the direct distance because the Bristol Channel intervenes. This exaggeration is the cause of the high third-dimensional component and anomalous position of these towns. However the magnitude of the second eigenvalue is about four times larger than that of all negative eigenvalues, and its inclusion allows closer agreement under the trace criterion. Although it is small in comparison with the first eigenvalue, on balance we would include it, if in ignorance of the true dimensionality of our land (presumably three!)

The procrustes statistic,  $\gamma_S$ , is 0.02965 and the time taken by the Bath I.C.L. 2980 computer to run the whole job, including procrustes analysis and input/output, was 112 units, a figure we provide for comparison purposes with the other methods. A summary of these two figures for all the methods used is given at the end of the section.

#### (b) Ordinal Scaling

The dissimilarity matrix was scaled twice by the ordinal method, once with a starting configuration generated by random sampling from a uniform distribution over the unit square, and once using the configuration generated by classical scaling as the starting configuration. Fifty iterations were performed in each case, the solution being sought in two dimensions. The global ordering of dissimilarities, and primary treatment of ties were used. For a random start the first step is taken as 0.2, for a 'rational' start 0.05 is the value selected.

#### (i) Random Start

The progress report on the iterations is given in Table 1.6.2. The stress function is approaching convergence, but the stress value is quite high for what should be a splendid fit. As might be suspected, the configuration is about to attain a local minimum. Holyhead, abandoned in the North Sea, is trying to reach its proper

TABLE 1.6.2

<u>IT. NO.</u>	<u>STEP</u>	<u>SLOPE</u>	<u>STRESS</u>
0		0.00485	0.43303
1	0.20000	0.00154	0.40708
2	0.16115	0.00119	0.40346
3	0.09964	0.00097	0.40199
4	0.05700	0.00056	0.40080
5	0.04372	0.00060	0.40003
6	0.03495	0.00060	0.39923
7	0.03162	0.00068	0.39831
8	0.05749	0.00099	0.39611
9	0.10530	0.00170	0.39003
10	0.14242	0.00241	0.37673
11	0.17494	0.00240	0.36011
12	0.15154	0.00253	0.34866
13	0.12853	0.00241	0.33936
14	0.11167	0.00259	0.33165
15	0.09890	0.00219	0.32333
16	0.08917	0.00216	0.31628
17	0.08789	0.00205	0.30910
18	0.08900	0.00195	0.30252
19	0.08681	0.00221	0.29670
20	0.07393	0.00229	0.29037
21	0.07556	0.00242	0.28287
22	0.09838	0.00234	0.27222
23	0.14988	0.00269	0.26222
24	0.12602	0.00358	0.25931
25	0.09420	0.00246	0.25198
26	0.07919	0.00264	0.24572
27	0.06980	0.00285	0.23846
28	0.07845	0.00340	0.22769
29	0.13467	0.00427	0.20580
30	0.22543	0.00678	0.16606
31	0.22418	0.01123	0.16261
32	0.20051	0.00945	0.13880
33	0.11077	0.00477	0.11617
34	0.10860	0.00753	0.11075
35	0.11883	0.00918	0.10936
36	0.06906	0.00293	0.09364
37	0.07209	0.00461	0.08804
38	0.07992	0.00838	0.09237
39	0.04251	0.00228	0.08175
40	0.06825	0.00689	0.08332
41	0.06066	0.00503	0.07995
42	0.01877	0.00180	0.07692
43	0.03998	0.00505	0.07830
44	0.02447	0.00090	0.07528
45	0.01843	0.00105	0.07469
46	0.01665	0.00189	0.07465
47	0.01058	0.00069	0.07420
48	0.00734	0.00076	0.07407
49	0.00543	0.00059	0.07396
50	0.00368	0.00047	0.07389

position but, as it moves that way, is meeting resistance from towns near Humberside that realise that it should not be near them. Hence we have a high procrustes statistic, 0.08574. It has been known for a random start to completely flip Scotland about the North/South Axis. The time taken for the entire run was 106 units.

(ii) Classical Start

The progress report on the iterations is given in Table 1.6.3. It may be noted that the stress value before any iterations have taken place is less than that obtained after 50 iterations from the random start. The slope and stress settle down very quickly and convergence to what we may suppose is the global optimum is rapid, so that about one half of the iterations are really unnecessary. The final procrustes statistic is 0.03194 which is slightly inferior to that obtained from classical scaling itself. The final configuration is more elongated, for example Penzance is displaced southwards to compensate for the Bristol Channel effect, and this accounts for the minor difference. The total time taken was 157 units.

(c) Least Squares Scaling

Again the dissimilarity matrix was scaled twice, once with all weights set equal to unity, and once with weights the reciprocals of the dissimilarity values (non-linear mapping). For both runs the configuration obtained from classical scaling was used as the starting configuration, and the solution obtained in two dimensions.

(i) Constant Weights

The Fletcher-Reeves algorithm satisfied its convergence criteria after 85 evaluations of the function, during which time the residual sum of squares had dropped from 469,498 to 219,062, and the gradient norm from 3981.1 to 0.98. The final procrustes statistic was

TABLE 1.6.3

<u>IT. NO.</u>	<u>STEP</u>	<u>SLOPE</u>	<u>STRESS</u>
0		0.00843	0.07134
1	0.05000	0.00484	0.05589
2	0.06720	0.00614	0.05385
3	0.04182	0.00352	0.05020
4	0.02482	0.00316	0.04951
5	0.01439	0.00138	0.04869
6	0.01217	0.00214	0.04869
7	0.00718	0.00056	0.04832
8	0.00586	0.00056	0.04821
9	0.00467	0.00076	0.04818
10	0.00298	0.00034	0.04813
11	0.00197	0.00030	0.04811
12	0.00137	0.00023	0.04809
13	0.00106	0.00020	0.04808
14	0.00136	0.00017	0.04807
15	0.00188	0.00021	0.04806
16	0.00125	0.00020	0.04806
17	0.00075	0.00011	0.04806
18	0.00049	0.00009	0.04805
19	0.00060	0.00008	0.04805
20	0.00079	0.00009	0.04805
21	0.00054	0.00009	0.04805
22	0.00035	0.00005	0.04805
23	0.00023	0.00005	0.04805
24	0.00028	0.00004	0.04805
25	0.00046	0.00004	0.04805
26	0.00035	0.00006	0.04805
27	0.00021	0.00003	0.04805
28	0.00014	0.00002	0.04805
29	0.00013	0.00002	0.04805
30	0.00014	0.00002	0.04804
31	0.00011	0.00002	0.04804
32	0.00008	0.00001	0.04804
33	0.00007	0.00001	0.04804
34	0.00009	0.00001	0.04804
35	0.00012	0.00001	0.04804
36	0.00008	0.00001	0.04804
37	0.00005	0.00001	0.04804
38	0.00003	0.00001	0.04804
39	0.00004	0.00001	0.04804
40	0.00008	0.00001	0.04804
41	0.00006	0.00001	0.04804
42	0.00003	0.00000	0.04804
43	0.00002	0.00000	0.04804
44	0.00004	0.00000	0.04804
45	0.00003	0.00001	0.04804
46	0.00002	0.00000	0.04804
47	0.00001	0.00000	0.04804
48	0.00001	0.00000	0.04804
49	0.00001	0.00000	0.04804
50	0.00001	0.00000	0.04804

0.02843, and the time taken 147 units. That this is a slight improvement upon the classical scaling result is probably a result of having to use all of the information available in two dimensions. Certainly the near linearity of true and road distances is helping the method to behave well in this example.

(ii) Non-Linear Mapping

This time 88 evaluations of the function were required. The residual sum of squares dropped from 2,523 to 990 and the gradient norm from 23.4 to 0.2. The final procrustes statistic was 0.02824, and the time taken 147 units. Again there is a slight improvement. This is caused by the sensible weighting of residuals, since it is certain that larger discrepancies will occur over greater road distances.

To summarise the results:

	<u>Procrustes</u>	<u>Time</u>
	<u>Statistic</u>	<u>Units</u>
		<u>Taken</u>
Classical .. .. .	0.02965	112
Ordinal (Random Start) .. ..	0.08574	106
Ordinal (Classical Start) .. ..	0.03194	157
Least Squares (Equal Weights) ..	0.02843	147
Least Squares (Non-Linear Mapping)	0.02824	147



## 1.7 Preprocessing the Dissimilarity Matrix

When it is unlikely that the specific values taken in the dissimilarity matrix are reliable, but their rank ordering can be used for multidimensional scaling, we naturally try the Shepard-Kruskal ordinal method. However an alternative is available. We may preprocess the dissimilarity matrix according to some distributional assumptions and use one of the methods that attach significance to the actual numerical values obtained, such as classical scaling or least squares scaling. Prior references are Benzécri (1964), Shepard (1966), Young (1970) and McGinley (1977).

Faced with the ordinal data, we may obtain a provisional numerical structure by replacing the ranking numbers by suitably chosen quantiles from the distribution we would expect the distances to follow if the configuration to be obtained were a sample of independent observations from, say, a multivariate normal distribution. The system of distances between independent points is not itself a system of independent random variables, but it is dissociated, and thus many parallel limit theorems apply. In particular, the empirical distribution function of the distances converges to the distribution of a single distance, and this provides the method with some kind of justification. The theory of dissociated random variables is developed in McGinley and Sibson (1975) and Silverman (1976). In Chapter 3 we explore the effects of assigning numerical values to ordinal data under the assumption that the underlying configuration is spherical normal in two dimensions with unit variances, in which case the distribution of squared interpoint distances is approximately a  $2 \chi^2_2$  distribution. We also investigate the effect of assuming an underlying configuration which is uniform on the unit disc, in which case the interpoint distance density is

$$\frac{4r}{\pi} \{ \cos^{-1} \frac{1}{2}r - \frac{1}{2}r\sqrt{1 - \frac{1}{4}r^2} \} \quad (0 \leq r \leq 2) \quad (1.7.1)$$

as given for example in Bartlett (1964). Where there happen to be ties in the ordinal data, the transformed values may be averaged. Finally we note that this can be an effective way of generating an initial configuration for use in ordinal scaling, although there is no point in using the transformed values themselves for they will have the same rank ordering as the original dissimilarities.

## 1.8 Single-Link Clustering

The single-link method of cluster analysis is the most famous and most straightforward of all the so-called sequential, agglomerative, hierarchical, non-overlapping group of techniques. Sibson (1973) describes an efficient algorithm that enables single-link clustering to be applied to dissimilarity matrices derived from over 1,000 objects at quite reasonable cost in computer time. We provide a brief formulation of this standard technique.

Let  $\Delta = (\delta_{ij})$  ( $i, j=1, \dots, N$ ) be a symmetric, zero-diagonal, dissimilarity matrix formed from  $N$  objects, some entries of which might well be missing. For all  $d \geq 0$  we may group the objects into disjoint sets formed by joining all objects with dissimilarity less than or equal to  $d$ . Thus a chain of links corresponding to dissimilarities less than or equal to  $d$  joins all members of a set, and we have a natural equivalence relation. This clustering may be defined for all values of  $d$ , and so we may derive a dendrogram, and the minimum spanning tree (see Gower and Ross, 1969) from the resultant nested equivalence classes.

This method alone possesses a combination of desirable mathematical properties outlined by Jardine and Sibson (1971). However such hierarchical clustering methods are limited, and single-link clustering is often criticised for its tendency to produce long, thin clusters as a result of its chaining structure. In the next section we describe a non-hierarchical method.

## 1.9 Partition Likelihood Clustering

The partition likelihood method of cluster analysis was developed by Gazard to cluster ceramic objects according to their distribution of trace elements following a neutron activation analysis. See Hammond, Harbottle and Gazard (1976) for details. We use the method to cluster languages according to the distribution of phoenetic value of the first syllable of their words. In this section we provide a brief description of the mathematical background and computer algorithm. Many similarities may be detected between this method and others that minimise within cluster sums of squares, such as that described by Ward (1963) and Ward and Hook (1963). A comprehensive list of other references may be found in the review paper of Cormack (1971).

The basic requirement is an  $N \times P$  data matrix  $X = (x_{ij})$ , whose rows represent  $N$  objects, whose columns represent  $P$  variables, and whose elements  $x_{ij}$  represent the number of 'atoms' of variable  $j$  associated with object  $i$ , where 'atom' can be interpreted as a particle or as a word in the quoted examples. Let us now consider a subset of  $R$  of the  $N$  objects, which we shall label  $1, \dots, R$  for convenience, and which we shall assume derive from a common source. We treat the elements of each row of the submatrix corresponding to these  $R$  objects as observations from a multinomial distribution with unknown parameter

$$\underline{q} = (q_1, \dots, q_p)$$

representing the proportion of each variable in the assumed common source.

Let  $S_{ij} = \frac{x_{ij}}{\sum_{k=1}^P x_{ik}}$  be the observed proportion of atoms of

variable  $j$  associated with object  $i$  ( $i=1, \dots, R$ )

Let  $N_i = \sum_{k=1}^P x_{ik}$  be the total number of atoms corresponding to

object  $i$ .

Then the maximum likelihood estimate of  $q$  is  $\hat{q}$  where

$$\hat{q}_j = \frac{\sum_{i=1}^R N_i S_{ij}}{\sum_{i=1}^R N_i} \quad (j=1, \dots, P) \quad (1.9.1)$$

Thus these maximum likelihood estimates are the centroids of the individual proportion vectors weighted according to the total number of atoms from each object. The associated loglikelihood is approximately proportional to

$$- \sum_{i=1}^R N_i \sum_{j=1}^P S_{ij} \log \left( \frac{S_{ij}}{q_j} \right) \quad (1.9.2)$$

where Stirling's formula has been used for simplifications.

If we now consider the partition of the  $N$  objects into  $T$  clusters it can be shown that the total loglikelihood is just the sum of  $T$  terms of the form (1.9.2). In practice we are unlikely to know  $N_i$ . We can either choose to regard it as constant, or make an assumption that it can be approximated by the total mass corresponding to object  $i$ . This can be interpreted as the mass of elements or

size of dictionary in our two examples. We also need to estimate the unknown  $q$  values by their maximum likelihood estimates  $\hat{q}$ . However we have all the necessary ingredients for an iterative clustering technique. At each stage the loglikelihood of a partition of the objects provides a criterion for assessing its satisfactoriness. We successively amalgamate clusters, choosing to join those which produce the least drop in loglikelihood. At each stage we check to ensure that the partition is admissible, that is that no object's vector of proportions lies nearer to that of a cluster centre in that it would produce an increase in loglikelihood if it were relocated to that group. Gazard produced three computer programs that seek the globally best partition into  $T$  clusters. These programs have varying degrees of sophistication and, correspondingly, make varying demands on computer storage and time. None of them are able to guarantee obtaining the globally best solution. The possibility of relocation at each stage means that this method is not hierarchical.

The method has an intuitive appeal for the non-mathematical user. He can easily think in terms of  $N$  species, each original specimen corresponding to a unique species, and the subsequent amalgamations are then just a matter of determining which division into  $T$  species would most satisfactorily, that is with most likelihood, represent all the specimen data. The concepts of likelihood, relocation and weighting by confidence are then seen to be quite natural.

### 1.10 Principal Component Analysis

Several times we have recourse to principal component analysis. Thus, given a  $N \times P$  data matrix  $X = (x_{ij})$ , where the  $N$  rows correspond to objects and the  $P$  columns correspond to variables, we find the eigenvectors and eigenvalues of  $S$ , defined to be the sample variance/covariance matrix, and project the  $P$ -dimensional point configuration onto the leading principal axes. We note that if we use Euclidean distance to derive a measure of dissimilarity between the objects, then principal component analysis is equivalent to classical scaling. Assuming  $P < N$ , it is computationally more efficient to use principal component analysis, which can then be used to generate a starting configuration for ordinal scaling.

### 1.11 The Generation of Dissimilarities

Sometimes data will arise naturally in the form of a  $N \times N$  matrix of dissimilarities. Such is the case for example when subjects in a psychological test are required to produce estimates of the similarity between objects presented to them. However it is more common to have to derive the measures of dissimilarity, and so in this section we briefly review the particular methods that we use.

The first case we examine arises when the data occur in the form of a  $N \times P$  matrix  $X = (x_{ij})$  of  $N$  rows corresponding to objects, and  $P$  columns corresponding to variables or attributes. The most common measure for us to use is the Euclidean distance given by

$$\delta_{rs} = \left\{ \sum_{j=1}^P (x_{rj} - x_{sj})^2 \right\}^{\frac{1}{2}} \quad (r,s=1,\dots,N)$$

or equivalently

$$\delta_{rs}^2 = (\underline{x}_r - \underline{x}_s)^T (\underline{x}_r - \underline{x}_s) \quad (1.11.1)$$

We need to be careful about the units associated with the variables, which might have vastly differing variances, and we need to beware of correlations among the variables. With both of these considerations in mind we sometimes use the estimated Mahalanobis  $D^2$  statistic given by

$$\delta_{rs}^2 = (\underline{x}_r - \underline{x}_s)^T S^{-1} (\underline{x}_r - \underline{x}_s) \quad (r,s=1,\dots,N) \quad (1.11.2)$$

where  $S$  is the  $P \times P$  sample variance/covariance matrix. We also use the  $P \times P$  sample correlation matrix  $R$  to remove correlations, but not



scale effects via

$$\delta_{rs}^2 = (\underline{x}_r - \underline{x}_s)^T R^{-1} (\underline{x}_r - \underline{x}_s) \quad (r,s=1,\dots,N) \quad (1.11.3)$$

And we may define the diagonal matrix T which has variances on the diagonal to standardise the variance of the variables, but retain correlations. Hence

$$\delta_{rs}^2 = (\underline{x}_r - \underline{x}_s)^T T^{-1} (\underline{x}_r - \underline{x}_s) \quad (r,s=1,\dots,N) \quad (1.11.4)$$

The whole range of Minkowski metrics may be used, where these are defined by

$$\delta_{rs} = \left\{ \sum_{j=1}^P (x_{rj} - x_{sj})^R \right\}^{1/R} \quad (R \geq 1; r,s=1,\dots,N) \quad (1.11.5)$$

These include

- R = 1      The city-block metric
- R = 2      Euclidean distance
- R =  $\infty$     The dominance metric

When interested in the distribution of data values for a particular object taken across the whole range of variables, we sometimes use a measure of dissimilarity called the information radius which has been derived by Sibson (1969) as follows:

Let  $\underline{p}$  and  $\underline{q}$  be the proportion vectors for objects i and j respectively.

Thus  $p_r = \frac{x_{ir}}{\sum_{t=1}^P x_{it}}$  and  $q_s = \frac{x_{js}}{\sum_{t=1}^P x_{jt}}$   $1 \leq r, s \leq P$

Then  $I(\underline{q}/\underline{p}) = - \sum_{k=1}^P q_k \log \left( \frac{q_k}{p_k} \right)$  (1.11.6)

is the information gain of  $\underline{q}$  given  $\underline{p}$ .

We define the dissimilarity between objects  $i$  and  $j$  as

$\delta_{ij} = I \left( \underline{q} / \frac{\underline{p} + \underline{q}}{2} \right) + I \left( \underline{p} / \frac{\underline{p} + \underline{q}}{2} \right)$   $1 \leq i, j \leq N$  (1.11.7)

Thus  $\underline{p} = \underline{q} \Rightarrow \delta_{ij} = 0$ , so that  $\delta_{ii} = 0$

and  $\underline{p} \neq \underline{q} \Rightarrow \delta_{ij} = \delta_{ji} > 0$ .

We are effectively assuming that the vectors  $\underline{p}$  and  $\underline{q}$  represent a probability distribution, and it is therefore necessary that the data values  $x_{ij}$  should all be positive if a natural interpretation is required. We shall call the matrix of  $\delta_{ij}$  values so generated the information radius dissimilarity matrix.

The second case we examine arises when the data are binary and we record for each object the presence or absence of a set of particular characteristics. The first step is to form the 2 x 2 association table for each pair of objects.

Object i

Present    Absent

Present

a	b
c	d

Absent

Object j

Thus  $a + b + c + d = P$ , the total number of characteristics.

We may then define various dissimilarity measures, for example

$$\delta_{ij} = \frac{b + c}{a + b + c + d} \quad (1.11.8)$$

is Hamming distance, familiar from communication theory

$$\delta_{ij} = \frac{b + c}{a + b + c} \quad (1.11.9)$$

is Jaccard distance, familiar from plant ecology. These both satisfy the metric inequality, although the proof that Jaccard distance is metric requires some subtlety, and are constrained to lie in the range 0 to 1. Choosing between them will depend upon how significant we feel the entries in the absent/absent cell to be.

Finally we consider the case of abuttal data. Data for scaling sometimes arise in the form of a three valued dissimilarity coefficient whose values are

- identical (precisely between each object and itself)
- neighbouring
- not-neighbouring.

We may think of the objects as really being regions rather than points, and it is the abuttals between regions which are recorded. Such data have been studied by Kendall (1971, 1974) who, for example, has successfully reconstructed a map of France from the information about the *Départements* that are neighbouring. The method that we use is to represent the regions by points, and then to assign conventionalised regions with the implied neighbour or contiguity

relationship by way of the Dirichlet tessellation (Green and Sibson, 1978). This construct assigns to each point the part of the space nearer to it than to any other point. The problem of reconstructing configurations from abuttal data is not easy and has been tackled by McGinley (1977). A significant step towards a solution is made by replacing the original three-valued dissimilarity coefficient by an integer valued one, the graph-theoretic distance or Wilkinson metric, which is the minimum number of abuttals traversed along a path from one point to another via abuttals. It is this Wilkinson metric that we use as the measure of dissimilarity between such objects.

The methods we have described are but few of many available. Comprehensive treatments of the subject and lists of references may be found in Jardine and Sibson (1971), Gower (1971a) and Cormack (1971).

### 1.12 Choice of Dimensionality in Classical Scaling

The results of this and the next section may be found in Sibson (1979), in which the author undertakes a perturbational analysis of classical scaling based upon small changes in the dissimilarity matrix.

In this section we investigate criteria that Sibson has proposed for determining the appropriate number of dimensions for a solution configuration from classical scaling. The correct choice of dimensionality is a problem in common with all multidimensional scaling methods, although it does not always arise in the same way. Here, for example, we effectively solve the scaling problem for all dimensions simultaneously. This is also true for principal component analysis. The problem is then to decide how many of the possible solution dimensions to accept. In contrast, ordinal scaling and least squares scaling are only solved for one number of dimensions at a time. Often the researcher might produce solutions in several spaces of differing dimensionality and try to determine which combines interpretability, ease of display and accurate representation of the original dissimilarity matrix. Then the situation is comparable to that which is faced with classical scaling. Kruskal (1964) suggested looking at the stress values and deciding at what level there ceased to be 'significant lowering of the stress with increasing dimensionality'. Others have taken the problem further by including an allowance for the number of objects and the suspected error in the dissimilarity values. None of these methods seem particularly satisfactory, especially when it is remembered that solutions in high numbers of dimensions are difficult to inspect visually, which is surely the primary objective of any scaling method.

However, returning to classical scaling, if we have what is believed to be a set of nearly-Euclidean dissimilarities it is possible to make some sensible estimate of the correct choice of dimensionality from the eigenvalue spectrum. Of course, if we have transformed the dissimilarities under the distributional assumptions of Section 1.7, then we have effectively imposed a dimensionality upon the configuration, and that is the only appropriate value to consider. We now follow the argument and notation of Sibson (1978).

Let  $E$  be an exact, squared distance matrix

"  $F$  be a symmetric, zero-diagonal matrix

"  $E(\epsilon) = E + \epsilon F + O(\epsilon^2)$  be a perturbation of  $E$

"  $B = q(E)$  be the exact inner product matrix

"  $\lambda$  be a positive, simple eigenvalue of  $B$

"  $\underline{e}$  be the corresponding unit eigenvector, orthogonal to  $\frac{1}{N}\underline{1}$  in order to satisfy centring conventions.

Then

$$\lambda(\epsilon) = \lambda - \frac{\epsilon}{2} \underline{e}^T F \underline{e} + O(\epsilon^2) \quad (1.12.1)$$

so that

$$E(\lambda(\epsilon) - \lambda) = - \frac{\epsilon}{2} \underline{e}^T E(F) \underline{e} + O(\epsilon^2) \quad (1.12.2)$$

$$\text{Also } \text{tr}(q(E(\epsilon))) = \text{tr } q(E) + \frac{\epsilon}{2N} \frac{1}{N}^T F \frac{1}{N} + O(\epsilon^2) \quad (1.12.3)$$

$$\text{Thus } \Sigma \text{ (eigenvalue perturbations)} = \frac{\epsilon}{2N} \frac{1}{N}^T F \frac{1}{N} + O(\epsilon^2) \quad (1.12.4)$$

But also

$$\begin{aligned} & \Sigma(\text{eigenvalue perturbations}) \\ &= \Sigma(\text{positive eigenvalue perturbations}) \\ &+ \Sigma(\text{zero eigenvalue perturbations}) \end{aligned}$$

Thus,  $\Sigma$  (zero eigenvalue perturbations) =

$$\frac{\epsilon}{2N} \frac{1}{N} F \frac{1}{N} + \frac{\epsilon}{2} \sum_{\text{positive eigenvalues}} \underline{e}^T F \underline{e} + O(\epsilon^2) \tag{1.12.5}$$

We may look at these results in either of two ways. Firstly, suppose that  $K$ , the number of genuine positive eigenvalues, and  $E(F)$  are known, then we can estimate  $\epsilon$  from the observed bias in the zero eigenvalues. Alternatively if we can assume that  $\epsilon$  is small, we can estimate the value of  $K$ , because the sum of the genuine positive eigenvalues ought then to be close to the trace of the perturbed inner product matrix  $q(E(\epsilon))$ . This gives rise to the trace criterion for determining the dimensionality of a solution from classical scaling: the sum of the genuine positive eigenvalues ought to be approximately equal to the sum of all the eigenvalues. This procedure has much greater appeal than the previous rule of thumb concerning looking for a large downwards drop, between the last supposed genuine eigenvalue and the first supposed spurious one, which incorporated the danger of disregarding useful information if the solution configuration should be narrow in some sense.

An allied technique is to reject as spurious those eigenvalues whose absolute magnitude is less than or not much greater than the absolute magnitude of the most negative eigenvalue. Underlying this criterion is the assumption that the perturbation of the multiple zero eigenvalue will be roughly symmetric if there is little error in the dissimilarity matrix. To justify this theoretically would involve the use of multiple eigenvalue perturbation theory and would seem a difficult task.

Thus the magnitude and trace criteria of Sibson both depend

upon the same assumption, and their performances tend to be comparable, as will be demonstrated in Chapter 3. It is clear that the results will be more satisfactory for nearly-Euclidean dissimilarities.



### 1.13 Perturbational Analysis of Procrustes Statistics

Sibson (1979) derived approximate expressions for the procrustes statistics  $G_E$  and  $G_S$  under perturbations to a configuration and subsequently under perturbations to the squared interpoint distance matrix used to construct a configuration by classical scaling. We have two reasons for being particularly interested in these results. Firstly they are fundamental to our approach to the error analysis of classical scaling reported in Section 3.2. Secondly we examine their range of validity in Section 3.5 and Section 3.6. We briefly summarise the results, starting with those relating to perturbations of a configuration.

Let  $X$  be a centred, full rank  $K \times N$  configuration matrix.

$B = X^T X$ , its inner product matrix, has eigenvalues

$\lambda_1 > \lambda_2 > \dots > \lambda_K > 0 \quad \lambda_{K+1} = \dots = \lambda_N$  which have corresponding eigenvectors  $\underline{e}_1, \dots, \underline{e}_N$  where  $\underline{e}_N = \underline{1}_N / \sqrt{N}$ .

Let  $Y = X + \epsilon Z + O(\epsilon^2)$  where  $Z$  is a  $K \times N$  matrix.

Then

$$G_E(X, Y) = \frac{\epsilon^2}{2} \left\{ \sum_{j, k=1}^K \frac{(\underline{e}_j^T (X^T Z + Z^T X) \underline{e}_k)^2}{\lambda_j + \lambda_k} + 2 \sum_{k=K+1}^{N-1} \underline{e}_k^T Z^T Z \underline{e}_k \right\} + O(\epsilon^3) \quad (1.13.1)$$

and

$$G_S(X, Y) = G_E(X, Y) - \frac{\epsilon^2 (\text{tr } (X^T Z))^2}{\text{tr } X^T X} + O(\epsilon^3) \quad (1.13.2)$$

In particular if the entries in  $Z$  are independent  $N(0,1)$  random variables then

$$G_E(X, Y) \sim \epsilon^2 \chi_f^2 + O(\epsilon^3) \quad \text{where } f = NK - \frac{1}{2}K(K+1) \quad (1.13.3)$$

and

$$G_S(X,Y) \sim \epsilon^2 \chi_g^2 + O(\epsilon^3) \quad \text{where } g = f - 1 \quad (1.13.4)$$

It is these forms that we examine in Section 3.5.

Next we summarise Sibson's results concerning perturbations to the squared interpoint distance matrix used to construct a configuration by classical scaling.

$$\begin{aligned} \text{Define } M_1 &= \{ (i,j): 1 \leq i,j \leq N-1 \} \\ \text{" } M_2 &= \{ (i,j): K+1 \leq i,j \leq N-1 \} \\ \text{" } M &= M_1 \setminus M_2 . \end{aligned}$$

Let E be the squared distance matrix of a centred, full rank, configuration X of N points in K dimensions. Let B be the corresponding inner product matrix with the usual eigenvalue structure

$$\lambda_1 > \lambda_2 > \dots > \lambda_K > 0 = \lambda_{K+1} = \dots = \lambda_N$$

and eigenvectors

$$\underline{e}_1, \dots, \underline{e}_N \quad \text{where } \underline{e}_N = \underline{1}_N / \sqrt{N}.$$

Let Y be the K-dimensional configuration recovered from

$$E(\epsilon) = E + \epsilon F + O(\epsilon^2) \quad \text{where } F \text{ is symmetric.}$$

Then,

$$G_E(X,Y) = \frac{\epsilon^2}{8} \sum_{(j,k)} \sum_{\epsilon M} \frac{(\underline{e}_j^T F \underline{e}_k)^2}{\lambda_j + \lambda_k} + O(\epsilon^3) \quad (1.13.5)$$

and

$$G_S(X,Y) = G_E(X,Y) - \frac{\epsilon^2}{16} \left\{ \frac{\sum_{k=1}^K \underline{e}_k^T F \underline{e}_k}{\sum_{k=1}^K \lambda_k} \right\}^2 + O(\epsilon^3) \quad (1.13.6)$$

Thus the expressions for  $G_E$  and  $G_S$  are quadratics in the elements of

F in both cases, albeit very complicated ones. This allows some distributional theory to be developed for the case of random errors in the distances, much in the same way that the  $\chi^2$  results were obtained for the procrustes statistic between two slightly different configurations, see (1.13.3) and (1.13.4), where the procrustes statistic was also a complicated quadratic in the elements of Z. For example if F is symmetric with zero diagonal and its off diagonal entries are independent with zero mean and unit variance, then Sibson shows that if we write

$$G_E(X,Y) = \frac{\epsilon}{2} A + O(\epsilon^3) \quad (1.13.7)$$

then

$$E(A) = \frac{1}{4} \sum_{(j,k)} \sum_{\epsilon M} \left\{ \frac{1 + \delta_{jk} - 2 \sum_m \frac{(e_j)_m^2 (e_k)_m^2}{\lambda_j + \lambda_k}}{\lambda_j + \lambda_k} \right\} \quad (1.13.8)$$

In Section 3.6 we consider this particular case and obtain values for the actual procrustes statistic, the approximation up to  $\epsilon^2$ , and the expected value of this approximation.

#### 1.14 References

- AHMAVAARA, Y. (1957). On the unified factor theory of mind.  
*Ann. Acad. Sci. Fenn. (B)*, 106, pp. 1-176.
- ANDERSON, A. J. B. (1971). Ordination methods in ecology.  
*Journal of Ecology*, 59, pp. 713-726.
- BARTLETT, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, 51, pp. 299-311.
- BENZÉCRI, J. P. (1964). Analyse factorielle des proximités.  
*Publications de l'Institut de Statistique de l'Université de Paris I*, 13, pp. 235-282.
- BLOXOM, B. (1978). Constrained multidimensional scaling in N spaces.  
*Psychometrika*, 43, pp. 397-408.
- CHANG, C. L. and LEE, R. C. T. (1973). A heuristic relaxation method for non-linear mapping in cluster analysis.  
*I. E. E. E. Trans. on Systems, Man and Cybernetics*, 3, pp.197-200.
- CLIFF, N. (1966). Orthogonal rotation to congruence.  
*Psychometrika*, 31, pp. 33-42.
- CORMACK, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A (General)*, 134, pp. 321-353.
- CRITCHLEY, F. (1978). Multidimensional scaling: a short critique and a new method. *Compstat 1978, Proceedings in Computational Statistics (Physica-Verlag: Vienna)*, pp. 297-303.
- CRITCHLEY, F. (1980). Optimal norm characterisations of multidimensional scaling methods and some related data analysis problems.  
*Proceedings of the Second International Symposium on Data Analysis and Informatics (17-19 October, 1979, Versailles)*, North-Holland: Amsterdam.
- GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis.  
*Biometrika*, 53, pp. 325-338.
- GOWER, J. C. and ROSS, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.*, 18, pp. 54-64.
- GOWER, J. C. (1971a). A general coefficient of similarity and some of its properties. *Biometrics*, 27, pp. 857-871.
- GOWER, J. C. (1971b). Statistical methods of comparing different multivariate analyses of the same data.  
*Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson, D. G. Kendall and P. Tautu, eds), pp. 138-149. Edinburgh: University Press.
- GOWER, J. C. (1975). Generalized procrustes analysis.  
*Psychometrika*, 40, pp. 33-51.

- GREEN, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17, pp. 429-440.
- GREEN, P. J. and SIBSON, R. (1978). Computing Dirichlet tessellations in the plane. *Computer Journal*, 21, pp. 168-173.
- GRUVAEUS, G. T. (1970). A general approach to procrustes pattern rotation. *Psychometrika*, 35, pp. 493-505.
- HAMMOND, N., HARBOTTLE, G. and GAZARD, T. (1976). Neutron activation and statistical analysis of Maya ceramics and clays from Lubaantun, Belize. *Archaeometry*, 18, pp. 147-168.
- HURLEY, J. R. and CATTELL, R. B. (1962). The procrustes program: Producing direct rotation to test a hypothesised factor structure. *Behav. Sci.*, 7, pp. 258-262.
- JARDINE, N. and SIBSON, R. (1971). *Mathematical Taxonomy*. New York: Wiley.
- KENDALL, D. G. (1971). Construction of maps from "odd bits of information". *Nature*, 231, pp. 158-159.
- KENDALL, D. G. (1974). The recovery of structure from fragmentary information. *Philosophical Transactions of The Royal Society A. Mathematical and Physical Sciences*, 279, pp. 547-582.
- KRISTOF, W. and WINGERSKY, B. (1971). Generalization of the orthogonal procrustes rotation procedure to more than two matrices. *Proceedings of the 79th Annual Convention of the American Psychological Association*, pp. 81-90.
- KRUSKAL, J. B. (1964a). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, 29, pp. 1-27.
- KRUSKAL, J. B. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika*, 29, pp. 115-129.
- KRZANOWSKI, W. J. (1971). A comparison of some distance measures applicable to multinomial data, using a rotational fit technique. *Biometrics*, 27, pp. 1062-1068.
- McGINLEY, W. G. and SIBSON, R. (1975). Dissociated random variables. *Math. Proc. Camb. Phil. Soc.*, 77, pp. 185-188.
- McGINLEY, W. G. (1977). Some optimisation problems in data analysis. Ph.D. thesis, University of Cambridge.
- MOSIER, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 4, pp. 149-162.
- SAMMON, J. W. (1969). A nonlinear mapping for data structure analysis. *I. E. E. E. Trans. on Computers*, 18, pp. 401-409.

- SCHONEMANN, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31, pp. 1-10.
- SCHONEMANN, P. H. (1968). On two-sided orthogonal procrustes problems. *Psychometrika*, 33, pp. 19-33.
- SCHONEMANN, P. H. and CARROLL, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35, pp. 245-255.
- SHEPARD, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, pp. 125-140.
- SHEPARD, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27, pp. 219-246.
- SHEPARD, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, pp. 287-315.
- SHEPARD, R. N. (1974). Representation of structure in similarity data: problems and prospects. *Psychometrika*, 39, pp. 373-421.
- SIBSON, R. (1969). Information radius. *Z. Wahrscheinlichkeitstheorie verw. G.*, 14, pp. 149-160.
- SIBSON, R. (1972). Order invariant methods for data analysis. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34, pp. 311-349.
- SIBSON, R. (1973). SLINK: an optimally efficient algorithm for single-link cluster analysis. *Computer Journal*, 16, pp. 30-34.
- SIBSON, R. (1978). Studies in the robustness of multidimensional scaling: procrustes statistics. *Journal of the Royal Statistical Society, Series B (Methodological)*, 40, pp. 234-238.
- SIBSON, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41, pp. 217-229.
- SILVERMAN, B. W. (1976). Limit theorems for dissociated random variables. *Advances in Applied Probability*, 8, pp. 806-819.
- SPAETH, H. J. and GUTHERY, S. B. (1969). The use and utility of the monotone criterion in multidimensional scaling. *Multivariate Behavioral Research*, 4, pp. 501-515.
- TORGERSON, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17, pp. 401-419.
- TORGERSON, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.

- WARD, J. H. (1963). Hierarchical grouping to optimise an objective function. *J. Am. Statist. Ass.*, 58, pp. 236-244.
- WARD, J. H. and HOOK, M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educ. Psychol. Measur.*, 23, pp. 69-82.
- YOUNG, F. W. (1970). Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika*, 35, pp. 455-473.
- YOUNG, G. and HOUSEHOLDER, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3, pp. 19-22.

C H A P T E R   T W O

PREVIOUS SIMULATION STUDIES OF MULTIDIMENSIONAL SCALING

	<u>PAGE</u>
2.1    Introduction .. .. .	48
2.2    The Founding Fathers: Shepard and Kruskal .. ..	51
2.3    Robustness of Ordinal Scaling .. .. .	53
2.4    Random Rankings .. .. .	62
2.5    The Choice of Starting Configuration .. .. .	64
2.6    The Relative Importance of the Small, Medium and Large Dissimilarities .. .. .	65
2.7    The Contribution of Lingoes and Roskam .. ..	67
2.8    The Recent Contribution of Shepard .. .. .	69
2.9    A Summary and our Approach .. .. .	74
2.10   References .. .. .	84



## 2.1 Introduction

In this chapter we present a review of papers that have appeared in the multidimensional scaling literature and that have contained at least an element of simulation study. Twenty-four such papers are reported. Most of them have appeared in the psychometric literature. All of them relate to ordinal scaling; none relate directly to either classical scaling or least squares scaling.

Firstly we record the early work of the founding fathers of ordinal scaling, Roger Shepard and Joseph Kruskal, whose original insights are so important that subsequent developments have been successful only to the extent that they have kept close to these original ideas. There then follow three sections, each of which is devoted to a problem that has captivated the imagination of the many who have followed in Shepard and Kruskals' footsteps without their imagination, but with access to better computing facilities. The first of these is the problem of how well the method can reconstruct a given configuration. There are many parameters that have a bearing on this and all of them have been investigated. Thus we find studies on the number of objects to be scaled, the quantity of error introduced to form dissimilarities from the original configuration, the dimensionality of the original configuration, the dimensionality of the reconstructed configuration, the Minkowski metric constant used to form the dissimilarities and the Minkowski metric constant used in forming the reconstruction. Typically either the final value of stress or the correlation between true and derived interpoint distances will be used to measure the success of the reconstruction. The second line of pursuit for scaling programmers has been to try to approximate the cumulative probability distribution for stress by Monte Carlo methods given

random rankings in the dissimilarity matrix. The idea is to be able to gauge when an empirically obtained stress value is sufficiently small to suggest that it is not just the result of random phenomena and that there is some structure in the dissimilarity matrix. Since stress is dependent upon the number of objects and the number of solution configuration dimensions, it is clear that the effect of at least these two parameters has had to be studied. The final popular problem has been the effect of the starting configuration upon the possibility of obtaining a merely local optimum value of the stress function. In the literature, a sharp contrast has been drawn between the approach of using a randomly generated starting configuration, and that of using a 'rational' starting configuration. Often the latter will be the configuration derived from classical scaling or possibly that derived from a procedure based on repeatedly scaling in a high number of dimensions and using all but the least of the solution principal axes dimensions in the starting configuration for the next iteration. Problems with local minima are alleged to have cast serious doubts on the validity of the other studies we have mentioned.

Three extra papers are reviewed for their particular interest. The first by Graef and Spence (1979) tackles the question of which dissimilarities play the most important part in enabling us to reconstruct a configuration adequately: the small, the medium or the large. This is highly relevant for computational considerations when the dissimilarity matrix is large. We address ourselves to this question in Section 3.4. We take special interest in the 1974 review paper by Roger Shepard, entitled 'Representation of Structure in Similarity Data: Problems and Prospects'. The author refers to six major problems and the prospects for overcoming each of them.

We accept his challenge, subsequent chapters providing partial answers to some of these difficulties. The third paper that we refer to at greater length is that of Lingoes and Roskam (1973), in which the authors attempt to compare the mathematical and computational aspects of two algorithms for ordinal scaling, the standard Shepard-Kruskal method and the Guttman-Lingoes SSA-I method.

The chapter is concluded with a statement of our view of the approaches that have been described, and with motivation for our simulation study design as reported in Sections 3.1, 3.2, 3.3 and 3.4.

## 2.2 The Founding Fathers: Shepard and Kruskal

The first published algorithm for ordinal scaling was presented by Shepard (1962a). The author sought that configuration for the objects in his study which would simultaneously have perfect correspondence between the ordering of interpoint distances and dissimilarities and yet lie in a space of minimum dimensionality. This was achieved by providing points too far apart with a force of attraction, providing points too close with a force of repulsion and then taking the vector sum of forces for each point. This process was performed in discrete steps, the configuration being adjusted at each step according to the magnitude and direction of this resultant force, and occasionally being projected into a space of one fewer dimensions. For  $N$  points the starting configuration was the regular simplex in  $N - 1$  dimensions. Shepard (1962b) then proved the power behind his idea by showing that he could recover a known configuration of points from the interpoint distances, even after these had been subjected to monotonic transformations of various kinds. This must have been the first simulation study of ordinal multidimensional scaling. Shepard mentioned five individual simulations amongst which he used three monotonic transformations and configurations lying in spaces of 1, 2 and 3 dimensions. For these simulations he demonstrated how the method was also able to recover both the point configuration and the underlying monotonic transformation to a remarkable degree of accuracy. However no method of comparison between configurations was introduced.

We have seen in Section 1.4 that a more satisfactory algorithm was devised by Kruskal (1964a, b), who employed least squares monotone regression techniques and established an explicit criterion for the measurement of the success of a configuration. The effectiveness of

his algorithm was shown by its ability to solve precisely the same reconstruction problem as that tackled by Shepard's simulations, with the additional distortion of a random error term added to each dissimilarity value. The first test of robustness had been performed, even if subsequent understanding does show that it was a rather mild one. To display his success Kruskal rotated configurations by eye, and measured any disagreement by the percentage difference in corresponding configuration distances.

That such accuracy could be obtained about the metric structure of a configuration from purely ordinal data was a point of fascination to Shepard, and is the basis of the success of ordinal scaling. Shepard (1966) sought to examine how many points were needed for the constraints provided by ranking interpoint distances to effectively determine the configuration. To do this he performed a large Monte Carlo study, using numbers of points lying between 3 and 45, and measuring the correlation between distances in the true configuration and the configuration regarded as optimal after ordinal scaling. The results show that fifteen points can be reproduced quite accurately, and any further improvements 'are of theoretical interest only'.

Thus these original simulation studies can be seen to have justified ordinal scaling, having shown it to be a method that can be relied upon to reproduce configurations very accurately, even after transformation of the underlying dissimilarity matrix and the addition of random error. Furthermore, empirical examples reported in these papers demonstrated the diverse practical applications that were possible.

### 2.3 Robustness of Ordinal Scaling

We now turn to ten other studies that concern the reconstruction of configurations by using ordinal scaling. We may conveniently summarise their approaches in terms of the following ten features:-

1.  $K$ , the dimensionality of a 'true' configuration,  $X$ .
2.  $N$ , the number of points in a 'true' configuration,  $X$ .
3.  $\Gamma(X)$ , the function used to derive the dissimilarity matrix,  $\Delta$ .
4.  $E$ , the number of levels of error introduced into  $\Gamma(X)$ .
5. The particular algorithms used to reconstruct the configuration.
6.  $\gamma(Y)$ , the function used to measure distance in the reconstructed configuration,  $Y$ .
7.  $k$ , the dimensionality of a reconstructed configuration,  $Y$ .
8. The functions used to measure the success of the reconstruction.
9.  $R$ , the number of replications for each combination of parameters.
10. Other properties emphasised in the study, and conclusions.

We consider the papers in chronological order of appearance.

#### (a) Sherman and Young (1968)

1.  $K = 2$  only.
2.  $N = 6, 8, 10, 15, 30$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance

calculation. Distances were then squared and added to 10.

4.  $E = 4$ .
5. The Young-Torgerson implementations of ordinal scaling, TORSCA.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 2$  only.
8. Kruskal's stress; correlation between true and reconstructed distances.
9.  $R = 5$ .
10. Recovery was shown to improve with more points but to worsen with more error. Stress increased with the number of points and so was not regarded as a sensible measure of reliability.

(b) Spaeth and Guthery (1969)

1.  $K = 1, 2, 3$ .
2.  $4 \leq N \leq 36$ .
3.  $\Gamma(X)$  : Unperturbed Euclidean distances were used.
4.  $E = 0$ , no error was added at all.
5. Kruskal's MDSCAL and Guttman-Lingoes' SSA-I.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3$ .
8. No measure apart from eye-appeal.
9.  $R = 1$ .
10. The true configurations formed known geometrical shapes, for example the vertices of a cube. MDSCAL was faster, SSA-I was equipped with a superior initial configuration; neither could guarantee successful reconstruction. We note that classical scaling would reconstruct these configurations exactly, and this would seem the obvious method for generating a starting configuration. Spaeth and Guthery introduce the least squares

criterion (Section 1.5), but do not seem to have implemented any method based upon it.

(c) Spence (1970b)

1.  $K = 1, 2, 3, 4$ .
2.  $6 \leq N \leq 36$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance calculation.
4.  $E = 4$ .
5. Young and Torgerson's TORSCA, Kruskal's MDSCAL, Guttman-Lingoes' SSA-I.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3, 4, 5$ .
8. Kruskal's stress, correlation between true and reconstructed distances.
9.  $R = 2$ .
10. The possibility of becoming entrapped in a local minimum of the objective function was the motivation for this approach. Successful convergence was shown to depend upon the accuracy of the initial configuration, and local minima were shown to be much more prevalent in one-dimensional solutions. The three algorithms produced solutions of comparable quality.

(d) Young (1970)

1.  $K = 1, 2, 3$ .
2.  $N = 6, 8, 10, 15, 30$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point



coordinate immediately prior to each Euclidean distance calculation. Distances were then squared and added to 10.

4.  $E = 5$ .
5. The Young-Torgerson implementation of ordinal scaling, TORSCA.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = K$  precisely.
8. Kruskal's stress, and the squared correlation between true and reconstructed distances.
9.  $R = 5$ .
10. Young laid particular emphasis on the ratio of degrees of freedom in the dissimilarity matrix to the degrees of freedom in the configuration. He showed that when this ratio is large the reconstruction will be good, even for large error. Once again it was shown that more points produce more precision, yet higher stress values.

(e) Wagenaar and Padmos (1971)

1.  $K = 1, 2, 3$ .
2.  $N = 8, 10, 12$ .
3.  $\Gamma(X)$  : Actual Euclidean interpoint distances were multiplied by independent normal random variables with mean 1 and variance  $\sigma^2$  to derive the dissimilarity values.
4.  $E = 5$ .
5. Not specified, presumably not a standard implementation.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3, 4, 5$ .
8. Kruskal's stress.
9.  $R = 11$ .
10. The authors used their results concerning stress values for

different levels of  $k$  to show that given one of either the true dimensionality,  $K$ , or the level of error, it is possible to estimate the other. Kruskal's elbow effect for determining dimensionality was shown to be inadequate for larger amounts of error.

(f) Spence (1972)

1.  $K = 1, 2, 3, 4.$
2.  $6 \leq N \leq 36.$
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance calculation.
4.  $E = 4.$
5. Young-Torgerson's TORSCA, Kruskal's MDSCAL, Guttman-Lingoes' SSA-I.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3, 4, 5.$
8. Kruskal's stress, correlation between true and reconstructed distances.
9.  $R = 2.$
10. This paper presents the results of Spence (1970b) in much greater detail. In particular it is a useful source for its description of the rationale behind the algorithm used and the initial configuration generated in each of the three publicly available computer packages. Guttman's SSA-I technique of rank images, and Young and Torgerson's iterative initial configuration routine are clearly presented. The conclusions are unchanged.

(g) Sherman (1972)

1.  $K = 1, 2, 3$ .
2.  $N = 6, 8, 10, 15, 30$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Minkowski metric calculation. The resultant distances were then squared and added to 10. The Minkowski metric parameter took values 1 (city-block), 2 and 3.
4.  $E = 3$ .
5. The Young-Torgerson implementation of ordinal scaling, TORSCA.
6.  $\gamma(Y)$  : Euclidean distance ( but alternatives are discussed ).
7.  $k = 1, 2, 3$ .
8. Kruskal's stress and the squared correlation between true and recovered distances.
9.  $R = 5$ .
10. This is an extension of the work of Young (1970), with more emphasis placed upon the Minkowski metric parameter. Young's results are substantially reaffirmed. In addition it is shown that it is only helpful to determine the correct Minkowski parameter if the true dimensionality is known.

(h) Spence and Graef (1974)

1.  $K = 1, 2, 3, 4$ .
2.  $N = 12, 18, 26, 36$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance calculation.
4.  $E = 5$ .

5. The Young-Torgerson implementation of ordinal scaling, TORSCA.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3, 4, 5$ .
8. Kruskal's stress.
9.  $R = 5$ .
10. A computer program is described which minimises a quadratic loss function that has been designed to have the correct number of solution dimensions and a reliable estimate of the error at its optimum. The program uses interpolation based on the results of this study. Empirical examples are given. The authors provide warnings; the method requires successful location of minimum stress values and is probably sensitive to the error model used.

(i) Isaac and Poor (1974)

1.  $K = 1, 2, 3$ .
2.  $N = 6, 8, 12, 16, 30$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance calculation.
4.  $E = 6$ .
5. Kruskal's MDSCAL.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 1, 2, 3, 4, 5, 6$ .
8. Kruskal's stress, squared correlation between true and reconstructed distances.
9.  $R = 5$ .
10. The authors propose a measure, termed 'Constraint', designed to enable the user to infer correct solution dimensionality.

It is assumed that the difference between the mean stress value from random ranking (Section 2.4) and a typical stress value obtained from the above procedures, or in practice, will be maximal in the correct dimensionality. The criterion is tested and shown to be moderately effective in general, and quite good when there is little error.

(j) Cohen and Jones (1974)

1.  $K = 3$ .
2.  $N = 9, 12, 15, 18$ .
3.  $\Gamma(X)$  : Independent normal deviates were added to each point coordinate immediately prior to each Euclidean distance calculation. However one of the three coordinate contributions was eliminated; this one being chosen from a probability distribution over the three dimensions. Four such distributions were used. The resulting two-dimensional distance was squared and added to 15 to form the dissimilarity.
4.  $E = 4$ .
5. The Young-Torgerson implementation of ordinal scaling, TORSCA.
6.  $\gamma(Y)$  : Euclidean distance.
7.  $k = 3$ .
8. Kruskal's stress, squared correlation between true and recovered distances, percentage intersection variance.
9.  $R = 4$ .
10. This study was based upon a psychological model that envisaged subjects having to estimate dissimilarities as using only two of three available dimensions in a 'stimulus space'. Those dimensions that were often used were well recovered, others poorly. Cohen and Jones also make some remarks about the

inadequacy of measures of recovery that have been used in the simulation studies.

## 2.4 Random Rankings

Several Monte Carlo studies have been carried out in an attempt to allow interpretation to be given to the optimal value of Kruskal's stress. Kruskal's (1964a) own early attempts at providing an interpretation were soon shown to be inadequate because they were independent of N, the number of objects, and K, the number of solution dimensions. In particular, investigators have sought a value of stress beneath which the dissimilarity matrix could be regarded as having 'significant structure'. Thus attempts have been made to obtain the cumulative probability distribution for stress for varieties of values of N and K, where the dissimilarities are random permutations of the first  $\frac{1}{2}N(N - 1)$  integers. We have encountered six such studies, due to Klahr (1969), Stenson and Knoll (1969), Wagenaar and Padmos (1971), Spence and Ogilvie (1973), Spence (1970a) and Isaac and Poor (1972). The latter two are less accessible being in thesis and unpublished manuscript form only. However we summarise the range of parameters used in the other four studies and the typical number of replications for each combination in Table 2.4.1.

---

TABLE 2.4.1

	<u>Min (N)</u>	<u>Max (N)</u>	<u>Min (K)</u>	<u>Max (K)</u>	<u>Replications</u>
Klahr	6	16	1	5	100
Stenson and Knoll	10	60	1	10	3
Wagenaar and Padmos	7	12	1	5	100
Spence and Ogilvie	12	48	1	5	15

We make a few other observations.

There does seem to be reasonable correspondence between results where separate studies overlap.

The results are also interpreted as an aid to determining correct dimensionality.

Stenson and Knoll consider the effect of tied values, which they demonstrate is very small under all treatments.

Spence and Ogilvie suggest methods of interpolation for intervening parameter values. They also recommend regarding a dissimilarity matrix as 'more than random' if the stress values obtained from scaling in all numbers of dimensions from 1 to 5 are all more than three standard deviations below their respective means. The peaked distribution of stress is the reasoning behind this apparently conservative approach.

None of the studies mentions particular care taken in avoiding the local minimum problem.



## 2.5 The Choice of Starting Configuration

The choice of starting configuration is an important element in the success and efficiency of ordinal scaling. That it can make a lot of difference is clear from the straightforward examples of Section 1.6. Recently attacks have been made upon the whole basis of published simulation studies, because they are alleged to have neglected this problem and been vitiated by the occurrence of local minima of stress. Arabie (1973, 1978a, 1978b) points to discrepancies that exist between papers reporting equivalent Monte Carlo analyses, and suspects that care has not been taken in dealing with merely local optima. He advocates the use of many random initial configurations before choosing that which leads to the lowest overall value of stress. Defences have been attempted by Spence (1974), Clark (1976) and Spence and Young (1978). The controversy has added little to our understanding. Four more helpful contributions are those of Spence (1970b, 1972), discussed in Section 2.3, Lingoes and Roskam (1973), Section 2.7, and Shepard (1974), discussed in Section 2.8.

## 2.6 The Relative Importance of the Small, Medium and Large Dissimilarities

If it is known that, say, the medium sized dissimilarity values contribute most to the success of a reconstruction by ordinal scaling, then other values might possibly be able to be treated as unknown. This would certainly make the scaling more economical and might take into account any redundancy that exists in the dissimilarity matrix. One simulation study of this problem has been carried out by Graef and Spence (1979). We report their findings.

Graef and Spence considered true configurations that had 31 points in 2 dimensions, randomly generated within the unit disc. Error was added to each interpoint distance according to one of the two popular error models reported in Section 2.3. Examples of the uses of the models come in Young (1970) and Wagenaar and Padmos (1971). Five combinations of model and error level were employed. Ten replications were used. The dissimilarity matrices were scaled by TORSCA in five different ways.

- (i) The entire matrix was used.
- (ii) One third of elements were deleted, according to a cyclic design.
- (iii) The smallest one third of elements were deleted.
- (iv) The middle third of elements were deleted.
- (v) The largest third of elements were deleted.

For each of these five treatments the root mean square correlation between true and recovered distances as well as their mean absolute differences were presented for the small distances, the medium distances, the large distances and all distances together. These measures were used because Graef and Spence claim that the

recovery of distances, and the recovery of configurations are very highly correlated.

The main conclusion to come out of the study is that the large distances are by far the most significant. When they are deleted the reconstruction suffers most drastically. It would seem that they determine the coarse structure of the configuration. When they are missing it is difficult to obtain an adequate starting configuration and the subsequent iterations lose their stability. We demonstrate in Section 3.4 that the small dissimilarities provide the fine, local structure.

Graef and Spence proceed to make recommendations about the practical consequences of their findings in terms of data collection methods. They also relate the results to a minimum adequate fraction of the  $\frac{1}{2}N(N - 1)$  dissimilarity values, which yields  $3NK$  as the minimum number of entries required.

These and our findings are of great importance in the applications reported in Chapter 4.

## 2.7 The Contribution of Lingoes and Roskam

The most elaborate simulation study of multidimensional scaling algorithms was presented in a 1973 "Psychometrika" monograph supplement by Lingoes and Roskam. Their objective was to examine the effectiveness of Kruskal's MDSCAL and Guttman-Lingoes' SSA-I in reaching the true optimum quickly. To discuss robustness and speed they were primarily concerned with the avoidance of local optima and the behaviour of the convergence process, two features which they demonstrated were dependent upon:-

- (a) The choice of initial configuration (including even its dimensionality).
- (b) The definition and construction of monotonicity and correspondingly the choice of loss function for minimisation (it is at this point that MDSCAL and SSA-I are most divergent).
- (c) The treatment of tied values.
- (d) The strategies used for guiding the algorithm to a desirable solution, including the calculation of step size.

Lingoes and Roskam performed over three thousand scalings to demonstrate their conclusions, and these were based upon:-

- (a) Some empirical matrices.
- (b) All distinct untied matrices of order 4.
- (c) Randomly generated matrices derived from interpoint distances of configurations lying in spaces of from one to five dimensions with from four to twenty points. The proportion of tied values was forced to vary from none to one half.

No matrix corresponded to more than twenty objects. Matrices were scaled in solution spaces of from one to five dimensions.

The main conclusions to emerge from the work were as follows:-

- (a) The choice of a good initial configuration is most important. Randomly generated configurations and the original Kruskal 'L' configuration have poor properties. Apart from these it is difficult to differentiate between other published suggestions.
- (b) Local minima are more likely when working far below the true dimensionality, and especially in one dimension.
- (c) The use of both MDSCAL and SSA-I loss functions in the same algorithm can reduce the likelihood of local minima, but necessitates more iterations.
- (d) Starting solutions in large numbers of dimensions followed by subsequent projections will also reduce the possibility of local minima. If this strategy is employed then it is worth spending fewer iterations in high dimensional spaces and occasionally projecting through more than one dimension in order to conserve time.
- (e) The primary treatment of ties is to be preferred to the secondary on the grounds of parsimony, as this will produce a less rigid structure. Indeed ties may then even be introduced in a discretisation of the data in order to reduce the solution dimensionality. This may also have the effect of reducing noise if the data is particularly suspect.

These considerations have been incorporated into a synthesis of the two previous algorithms now entitled MINISSA-I.

## 2.8 The Recent Contribution of Shepard

Many facets of multidimensional scaling that are relevant to our simulation studies (Chapter 3) and applications (Chapters 4 - 7) have been discussed in a recent review paper by Shepard (1974), in which the author identified six problem areas in the subject and proceeded to give his opinion as to the prospect for their solution.

These problems concerned:-

- (a) Local minimum solutions.
- (b) Finding meaningful interpretations.
- (c) Choice of dimensionality.
- (d) Loss or imposition of structure caused by degeneracy.
- (e) Choice of the underlying metric.
- (f) Representation of categorical structure.

The first five of these are all highly relevant in this thesis, and accordingly we review Shepard's comments.

### (a) Local Minimum Solutions

Shepard was dissatisfied with the solutions obtained from both random and classical scaling starting configurations. For both, local minimum solutions were likely, especially in spaces of one dimension and with non-Euclidean metrics. Random starts had the additional disadvantage of very slow convergence. Altogether, Shepard advocated the use of at least twenty different random starts in order to ensure attaining the global minimum. Classical scaling produced a poor configuration when the relationship between dissimilarity and distance was highly non-linear. The whole problem cast doubts on the usefulness of published Monte Carlo studies. The problem was put down to the mutual repulsion between dissimilar points which prevented them from 'crossing-over' or swapping positions.

The initial configuration is vital in this problem, and methods were suggested that would generate a successful one. Thus Shepard advocated an iterative approach using classical scaling and transforming the dissimilarity values to linearise the resultant dissimilarity versus distance plot, proceeding until this process converged. A similar approach is adopted in the algorithm TORSCA, in which successive dissimilarity values are replaced by the least squares monotone regression fit to the distances of the configuration produced by classical scaling. Alternatives suggested included permuting the objects into best fit with respect to a previously established configuration, and also building up the initial configuration by adding points one by one in the best available location. Neither of these has been explored. However, the method that Shepard recommended most highly involved scaling in higher numbers of dimensions and successively projecting the solution configuration onto a space of lower dimensionality and using this configuration as the starting point for the new iteration. This method he described as "uniformly successful".

(b) Finding Meaningful Interpretations

The success of an exercise in scaling was to be measured by its interpretability, and conversely a useful interpretation added confirmation to the number of dimensions used, whilst making a local minimum, degenerate or random solution seem unlikely. Stress was to be treated with caution because a low stress could mean a large number of unreliable, uninterpretable dimensions. Equally the axes that defined the solution could be inappropriate for describing its structure. Shepard emphasised the need for awareness of five points.

- (i) Axis rotation may provide extra insight.
- (ii) There is a trade-off between stress and dimensionality.
- (iii) Solutions in spaces of one, two and three dimensions are more easily appreciated.
- (iv) Clusters and circular orderings often arise.
- (v) Objective methods can be used to measure variables on particular sets of rotated axes as an aid to interpretation.

(c) Choice of Dimensionality

In the same spirit Shepard appealed against the common practice of extracting too many dimensions, and the over-importance attached to the stress value. This had been encouraged by the artificial Monte Carlo experiments on the distribution of stress, which often applied to spaces of high dimensionality. Furthermore what were essentially one-dimensional solutions often arose in two- and three- dimensional spaces as 'C', 'S', or helical shapes, a fact that could be demonstrated by reordering the rows (and columns) of the dissimilarity matrix. There was also the hope that powerful mathematics could be developed which would infer the true dimensionality from the constraints within the dissimilarity matrix.

(d) Loss or Imposition of Structure Caused by Degeneracy

Unjustified structure could be imposed if stress took the value zero, in which case several starting configurations would determine the range of possible solutions, and after which more objects could be introduced in the analysis, and less dimensions used. Shepard recommended a minimum of ten objects for two-dimensional solutions.

An alternative cause of low stress was the degenerate case in



which the objects split into clusters such that all within-cluster similarities were greater than all between-cluster similarities, in which case zero stress may also have occurred. Thus true structure within the clusters was obscured. The tendency for many of the least squares monotone regression values to be equal was shown to be an indicator of this type of degeneracy. This behaviour was undesirable from both the statistical and substantive points of view, and if it was possible Shepard recommended that the objects should be chosen so as not to be obviously grouped into clusters. If degeneracy did occur he recommended reanalysing each sufficiently large cluster independently. Finally he showed examples in which the monotone fit was to some parametrised functional form which would force the distinction between points. This can be thought of as a metric method. Indeed this idea has been exploited by Critchley (1978, 1980) as referred to in Section 1.3.

(e) Choice of the Underlying Metric

Shepard dealt with the range of Minkowski  $r$ -metrics available, and pointed out that for values of  $r$  apart from two, there were even greater problems with slow convergence and local minimum solutions. This effect was so pronounced that Shepard advised avoiding random starts altogether. The best chance of reaching the global minimum was obtained by working outwards from the global minimum solution with  $r$  taken as two. Stress values could not be compared for different metrics, and it was shown that tied distances, degeneracies and lower stress values were easier to obtain with non-Euclidean, particularly dominance, metrics. In addition there was the problem of the misleading representation of the structural information for the non-Euclidean metrics which were axis dependent. It would also have been difficult to infer the

correct choice of  $r$  from the dissimilarity matrix, since the rank ordering of optimum configuration distances among the different 'r' values was likely to be similar. In conclusion it seemed that little was to be gained from using anything other than Euclidean distance.

## 2.9 A Summary and our Approach

In order to summarise our reactions to the methodology and rationale behind the studies that have been reported in this chapter we first of all explain the reasoning for our approach reported in Chapter 3. This can be done by developing the points in precisely the same format as that used in Section 2.3, whilst criticising and commenting upon the other studies. This will not cover everything that has arisen, and so we must append the remaining observations to the end of this section. Firstly, however, it seems important to develop the reasoning lying behind the use of a 'true', 'parent' configuration in multidimensional scaling simulation studies.

The crucial point is that the final product of a successful exercise in scaling is a point configuration. It is not a set of distances, nor a least squares monotone regression fit to those distances. Thus to be concerned with the effectiveness of a particular algorithm, or to measure its response to error is to be concerned with configurations. So it is important to be able to measure relative differences between configurations, which the procrustes statistic does naturally, and also to measure the absolute difference of a reconstruction from a specified yardstick, which is provided by a 'true' configuration.

Whilst a reconstruction which corresponds identically with the true configuration minimises the procrustes statistic, it need not necessarily correspond to the global minimum of the scaling method once error has been added. Here the procrustes statistic provides a natural measure of the effect of the introduction of error. Another benefit of using a 'true' configuration is that it enables an easy check to be made to determine whether a scaling method has

converged to a local minimum solution or not. This becomes clear if one or more points has become badly placed. Any methods based upon stress itself or, say, rank correlation between dissimilarities and distances, would not require 'true' configurations, but would correspondingly fail to be sensitive, configuration-based or useful in fixing standards. All reported studies have agreed and used true configurations.

1. Having decided that it is sensible to use 'parent', 'true' configurations we must decide how large to make their dimensionality. It is only possible to appreciate solutions from scaling methods adequately when these lie in one, two or three dimensions. If we realise that one-dimensional solutions show up in two-dimensional spaces as horseshoes and 'S' shapes, then there is no danger in seeking a reconstruction in two dimensions, particularly as this will give less local minimum problems. Equally since three-dimensional solutions are that bit more difficult to appreciate, a two dimensional solution is usually sought first. So the natural first choice when one is scaling is a two-dimensional reconstruction. In Sections 3.3 and 3.4 we are not concerned with the problem of estimating the dimensionality, rather comparing scaling methods and determining the relative contributions made by different sized similarity values. Thus the natural choice for the dimensionality of the 'true' configuration is two. However in Section 3.2 we are concerned with the robustness of one particular method, classical scaling, and the effectiveness of criteria for determining the true dimensionality from the eigenvalue spectrum. So in that section we use spaces of two to six dimensions, although once again we place most emphasis on two-dimensional true configurations.

2. The next step is to consider the number of points that should be used. Here we differ considerably from most previously published studies. It seems that a typical problem in psychology involves about twenty objects. That most of the development of the subject has taken place in psychology is reflected in the small numbers of objects that these studies have used. We seek to demonstrate that scaling has a much broader field of application. It was felt that fifty objects would correspond to a medium-sized problem, and accordingly this is the number that has been used in all our simulations. Subsequent experience suggests that this might even be unrealistically small. For example, the three applications reported in this thesis all involve scaling at least one hundred objects. However, in the interests of conserving computing resources and bearing in mind that different behaviour is unlikely to arise beyond such a level, fifty points would seem to be quite adequate for our simulation studies. Our configurations are generated by realising these fifty points uniformly and independently in the unit disc of appropriate dimensionality, but the only feature of real significance is, in our view, that the configurations should be roughly spherical with no special structure. We do not generate an unlimited supply of new configurations, but we do use enough to provide a check against being misled by the behaviour of any particular one.

3. The models that have so far been described to derive the dissimilarity matrix from the true configuration are open to some criticism. Firstly it is our view that in the majority of applications of scaling methods it is not appropriate to assume that the observed dissimilarities differ from the true interpoint distances by errors that are independent. This has been the case in all the error models we have described so far, for they produce independent

errors. Secondly, all the models we have seen have set up an error distribution by decree, usually adding to interpoint distance a normal deviate. No justification has been attempted for this approach. Thirdly some models have used a non-Euclidean Minkowski metric. To do so would normally presuppose the search for a best representing solution considering all Minkowski parameters, and we accept Shepard's (Section 2.8) warnings about the lack of wisdom of such an approach. In Section 3.1 we describe the four Euclidean models that we use for the generation of dissimilarities. These are designed to be simple, but practically relevant, and are based upon simple ideas from geometry and probability and use measures of similarity commonly employed in taxonomy. At once we are able to introduce dependent errors, test the effects of dependence and use naturally occurring error distributions. It is always possible to produce plots of the actual interpoint distance against the derived dissimilarity, and this we do to show up the spread and nearness to linearity of the relationship.

4. In common with nearly all the other simulation studies we are able to vary the error level and so determine the sensitivity of the scaling methods to different amounts of error. Not only is this a vital element in the study of robustness, but also the use of different error levels enables us to draw conclusions about the types of dissimilarity matrix for which the varying scaling methods will be appropriate. In nearly all of our studies we use six levels of error and these are admitted quite naturally by the models described in Section 3.1.

5. The other studies have demonstrated that a large variety of versions of Kruskal's original algorithm are available in computer packages. All of these are aiming to solve the same problem, even

Guttman-Lingoes' SSA-I in which the actual algorithm is slightly different. The version of MDSCAL that we use was developed by Robin Sibson at Cambridge and offers all of the options outlined in Kruskal's original papers with a few extras, and is written in an exceptionally compact and efficient manner. The classical scaling program was produced by the same author in the same style and relies upon the NAG subroutine FO2ABF for eigenvalue extraction. The other scaling method we describe, least squares scaling, has been implemented by Adrian Bowyer and Robin Sibson at Bath and depends upon Fletcher-Reeves function minimisation by conjugate gradients as programmed by NAG in their subroutine EO4DBF. The procrustes analysis program was produced by Adrian Bowyer, and again depends upon the subroutine FO2ABF. The theory behind these techniques is developed in Sections 1.4, 1.3, 1.5 and 1.2 respectively.

6. In accordance with all of the other studies we have reviewed we use Euclidean distance in relating the reconstructed configuration to the dissimilarity matrix. This is appropriate, in order to correspond to the measurements used in forming the dissimilarities as described in Section 3.1 and defended earlier. Any more elaborate method would require ample justification in the light of Shepard's (1974) warnings.

7. Two situations arise when we consider the number of dimensions to choose in reconstructing our configurations. Firstly when we are comparing methods (Section 3.3) we are not directly concerned with the choice of dimensionality, rather the relative success of the algorithms, and so we do no more than reconstruct the configuration in the correct number of dimensions, assuming that this can be determined in practice. Secondly, however, when we are looking at classical scaling (Section 3.2) we are effectively solving the scaling problem for all levels of dimensionality

simultaneously. The problem with classical scaling comes in choosing how many of the available dimensions to accept. We can examine our criteria without having to compare our solution with the original configuration. After we have done this we can look at the precision of the reconstruction by first choosing exactly the correct number of dimensions and then using our measure of goodness-of-fit, namely the procrustes statistic, to enable us to do this.

8. Nearly all of the studies we have reviewed have used Kruskal's optimal stress value and the correlation between true and recovered distances in order to assess the success of the recovery of the true configuration. One exception was the paper of Cohen and Jones (1974) who used Percentage Intersection Variance as well, but this is also a form of correlation coefficient. Specifically, if  $r_{dd}$  is defined as the correlation between true and reconstructed dimension  $d$  after procrustes fitting, then

$$P.I.V. = \left( \sum_{d=1}^3 r_{dd}^2 \right) / 3 .$$

We feel that these methods leave a lot to be desired. Cohen and Jones have themselves pointed out a number of objections to using correlation to measure recovery. Firstly, it is insensitive for similar configurations, and these will often be our concern. Also it can be drastically misleading following the displacement of one single point. It is unaffected by the addition of a constant to either or both sets of distances. Neither does it deal satisfactorily with the comparison of configurations of differing dimensionality. We would summarise all these objections by saying that the correlation coefficient simply fails to relate properly to either the geometrical or the probabilistic aspects of the problem.



The other common measure of the success of the reconstruction is the Shepard-Kruskal stress achieved at optimality in the ordinal method. This seems to us to reflect a misunderstanding of the role of the optimised stress function, which measures the euclideanness of a set of ordinal data, and not the extent to which the reconstruction matches the original configuration. It is quite easy to derive measures of resemblance other than the correlation function between the distances in the original and reconstructed configurations; some of these look rather like the stress. However, when used as measures of the success of reconstruction, most of the same criticisms may be applied to these as to the correlation coefficient: all such measures miss the point of the problem, which is that it is configurations, not distances, which must be compared. The approach we adopt, using procrustes statistics, avoids these criticisms.

We always compare a recovered configuration  $Y$  with its parent configuration  $X$  by using a procrustes statistic: see Section 1.2 and Sibson (1978). The particular form of statistic employed allows  $Y$  to be fitted to  $X$  under the action of the group of similarity transformations, that is, the group generated by translation, rotation, reflection and uniform scale change. This leads to the statistic  $G_S(X,Y)$  as in (1.2.5) and we normalise this to  $\gamma_S(X,Y)$  as in (1.2.6). We use  $\gamma_S(X,Y)$  which lies in the range  $[0,1]$ , for all our comparisons. It is appropriate to do this even with classical scaling, because in practice the approximately linear relation between dissimilarity and distance is usually an unknown one.

9. Estimation of the variance of the distribution of the procrustes statistic for a particular combination of parameters has been made possible by repeated realisations from the random processes involved. In particular in the classical scaling studies of Section 3.2

we use ten replications. In the comparative studies of Section 3.3 the idea is to use the same dissimilarity matrix with different scaling methods, and so the emphasis is not so much on the distribution of the procrustes statistic and replications are not so essential. However it is possible to repeat the whole procedure for different matrices and this we have done. Similarly the Section 3.4 simulations on the relative importance of different sized dissimilarities for good reconstruction rely on using one matrix in different ways, but again this is repeatable, and we do repeat to a limited extent.

10. It has emerged that there are several features of unusual interest in our scaling studies. Our aim is not so much to compare adaptations of the same algorithm, but rather to treat different methods of scaling. Here the emphasis is on the use of classical scaling and its robustness (Section 3.2), but least squares scaling, ordinal scaling and preprocessing techniques are also considered (Section 3.3). To do this we develop four euclidean models for the generation of dissimilarities (Section 3.1) which enable us to investigate dependence among errors, these models being quite unlike anything that has appeared previously. We are concerned with problems of local minima, and correct choice of dimensionality in classical scaling, but these topics are dealt with as they arise rather than being our focus of attention. The final point of novelty is the development of the procrustes statistic as the tool for comparing configurations.

We now turn to mention other relevant points arising from these studies.

One recurring theme has been the inadequacy of stress for measuring the success of a reconstruction. It has been emphasised that stress is far too dependent upon the number of objects and number

of dimensions to be comparable for different values of these parameters, and it has been demonstrated that larger stresses may correspond to more precisely constrained configurations. In view of this it seems to have been rather futile to expend so much energy in computing random ranking stress distributions. These efforts have been dogged by the need to perform the simulations for every possible combination of parameters, the need to avoid local minima and the need to justify the use of a set of 'structureless similarity matrices'. Unfortunately failure has occurred in all three areas; the ranges of parameters are limited, accusations are made concerning the studies' reliability and little practical use has been made of the results.

The contributions made by different sized similarities are discussed in Section 3.4.

It is our experience that the configuration obtained by classical scaling has been most satisfactory in avoiding merely local minima. Clearly the extent of the euclideanness of the dissimilarities will determine how reliable this method will be. However if the dissimilarities are clearly not very euclidean it is always possible to devise a transformation which will cause an improvement. An additive constant and simple power transformation are useful first steps. More complicated, but in the same spirit and very useful, is the TORSCA technique of replacing the dissimilarities by the least squares monotone regression fit values obtained with the classical scaling configuration. This can establish an iterative approach. With these extra facilities the use of classical scaling seems uniformly powerful.

Finally we note that the applications reported later in this thesis provide examples of several of Shepard's remarks on

interpretability. Circular orderings abound in Chapter 4; Chapter 6 shows the formation of clusters, and the objective choice of axis in aiding interpretation is illustrated in Chapter 4 and Chapter 7.

## 2.10 References

- ARABIE, P. (1973). Concerning Monte Carlo evaluations of nonmetric multidimensional scaling algorithms. *Psychometrika*, 38, pp. 607-608.
- ARABIE, P. (1978a). Random versus rational strategies for initial configurations in nonmetric multidimensional scaling. *Psychometrika*, 43, pp. 111-113.
- ARABIE, P. (1978b). The difference between "several" and "single": a reply to Spence and Young. *Psychometrika*, 43, p. 119.
- CLARK, A. K. (1976). Re-evaluation of Monte Carlo studies in nonmetric multidimensional scaling. *Psychometrika*, 41, pp. 401-403.
- COHEN, H. S. and JONES, L. E. (1974). The effects of random error and sub-sampling of dimensions on recovery of configurations by non-metric multidimensional scaling. *Psychometrika*, 39, pp. 69-90.
- CRITCHLEY, F. (1978). Multidimensional scaling: a short critique and a new method. *Compstat 1978, Proceedings in Computational Statistics (Physica-Verlag:Vienna)*, pp. 297-303.
- CRITCHLEY, F. (1980). Optimal norm characterisations of multidimensional scaling methods and some related data analysis problems. *Proceedings of the Second International Symposium on Data Analysis and Informatics (17-19 October, 1979, Versailles)*, North-Holland: Amsterdam.
- GRAEF, J. and SPENCE, I. (1979). Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, 86, pp. 60-66.
- ISAAC, P. D. and POOR, D. D. S. (1972). On the estimation of appropriate dimensionality in data with error. *Unpublished manuscript, Ohio State University, 1972.* Cited in Spence (1972).
- ISAAC, P. D. and POOR, D. D. S. (1974). On the determination of appropriate dimensionality in data with error. *Psychometrika*, 39, pp. 91-109.
- KLAHR, D. (1969). A Monte Carlo investigation of the statistical significance of Kruskal's scaling procedure. *Psychometrika*, 34, pp. 319-330.
- KRUSKAL, J. B. (1964a). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, 29, pp. 1-27.
- KRUSKAL, J. B. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika*, 29, pp. 115-129.
- LINGOES, J. C. and ROSKAM, E. E. (1973). A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika*, 38, Monograph Supplement, pp. 1-93.

- SHEPARD, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, pp. 125-140.
- SHEPARD, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27, pp. 219-246.
- SHEPARD, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, pp. 287-315.
- SHEPARD, R. N. (1974). Representation of structure in similarity data: problems and prospects. *Psychometrika*, 39, pp. 373-421.
- SHERMAN, C. R. (1972). Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters. *Psychometrika*, 37, pp. 323-355.
- SHERMAN, C. R. and YOUNG, F. W. (1968). Nonmetric multidimensional scaling: a Monte Carlo study. *Proceedings, 76th Annual Convention, American Psychological Association*, 3, pp. 207-208.
- SIBSON, R. (1978). Studies in the robustness of multidimensional scaling: procrustes statistics. *Journal of the Royal Statistical Society, Series B (Methodological)*, 40, pp. 234-238.
- SPAETH, H. J. and GUTHERY, S. B. (1969). The use and utility of the monotone criterion in multidimensional scaling. *Multivariate Behavioural Research*, 4, pp. 501-515.
- SPENCE, I. (1970a). Multidimensional scaling: An empirical and theoretical investigation. *Unpublished Ph.D. thesis, University of Toronto*, 1970.
- SPENCE, I. (1970b). Local minimum solutions in nonmetric multidimensional scaling. *Proc. of the Soc. Stats. Section of the American Statistical Association*, 13, pp. 365-367.
- SPENCE, I. (1972). A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika*, 37, pp. 461-486.
- SPENCE, I. (1974). On random rankings in nonmetric scaling. *Psychometrika*, 39, pp. 267-268.
- SPENCE, I. and GRAEF, J. (1974). The determination of the underlying dimensionality of an empirically obtained matrix of proximities. *Multivariate Behavioural Research*, 9, pp. 331-341.
- SPENCE, I. and OGILVIE, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioural Research*, 8, pp. 511-517.
- SPENCE, I. and YOUNG, F. W. (1978). Monte Carlo studies in nonmetric scaling. *Psychometrika*, 43, pp. 115-117.

- STENSON, H. H. and KNOLL, R. L. (1969). Goodness-of-fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin*, 71, pp. 122-126.
- WAGENAAR, W. A. and PADMOS, P. (1971). Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology*, 24, pp. 101-110.
- YOUNG, F. W. (1970). Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika*, 35, pp. 455-473.

C H A P T E R   T H R E E

FOUR SIMULATION STUDIES OF MULTIDIMENSIONAL SCALING

	<u>PAGE</u>
3.1    Euclidean Models for the Generation of Similarities	88
3.2    Simulation Studies of Classical Scaling    .. ..	102
3.3    Comparison of Scaling Methods    .. .. ..	123
3.4    Simulations on Scaling Subsets of Similarities    ..	128
3.5    Procrustes Statistics Arising from Slightly Different Configurations    .. .. ..	136
3.6    Procrustes Statistics Arising from Slightly Different Squared Distance Matrices    .. ..	144
3.7    References    .. .. ..	148



### 3.1 Euclidean Models for the Generation of Similarities

The techniques that we use for testing the operation of scaling methods have been previewed in Section 2.9 as a response to the work of others. In particular we have defended the use of 'parent', 'true' configurations and introduced our choice of such configurations as those of fifty points lying in spaces of dimensionality from two to six, generated independently from a uniform distribution over the unit disc. We now turn to a detailed account of the four euclidean models that are used to generate similarities. These are based upon simple geometrical and probabilistic constructs with a view to allowing dependent errors and different quantities of error. A range of conditions are thus provided for our various scaling methods. In addition the models use measures that are commonly employed in taxonomy and so have an added appeal. The effects of dependence are examined more exactly by arranging that one of the four models is a version of another with errors forced to be independent. See also Sibson, Bowyer and Osmond (1981) for an alternative account of these models. Much of the derivation of the following models is due to the first author of that trio.

#### Binomial Hyperplane Model

The first model is an attempt to represent the simple matching coefficient of numerical taxonomy in which objects are compared by counting the number of attributes (variables) in which they concur and normalising with respect to the total number of attributes. If the attributes are coded into 0 (absent) and 1 (present) we have a straightforward form of binary data. The total number of differences between any pair then provides a metric which is known in communication theory as Hamming distance. For a specified configuration in  $k$  dimensions, random Hamming distances may be

generated by a number of randomly located hyperplanes, each of which divides the space into two half-spaces, one arbitrarily coded as 0, the other as 1. If the hyperplanes are realised from a Poisson hyperplane process, then each Hamming distance is almost surely well defined, finite, and has a Poisson distribution, the parameter of which is the product of the euclidean distance between the two points and the intensity of the process, where this is expressed in appropriate units. The Hamming distance may then be seen to be just the number of hyperplanes that are traversed in passing from one point directly to the other. The mean Hamming distance will then be proportional to the corresponding euclidean distance, so that the relationship between them is roughly a linear one. As the intensity of the Poisson process becomes large, so the relative values of the system of Hamming distances converge to those of the euclidean distances. The Hamming distances are not independent; any two of them together have a bivariate Poisson distribution (see Mardia, 1970) whose parameters may be expressed in geometrical terms.

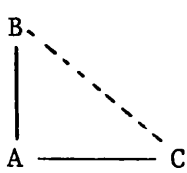
For example, consider the two-dimensional situation where

$P(\mu)$  denotes a Poisson random variable with mean  $\mu$ .

$N_{XY}$  denotes the number of Poisson lines intersecting line segment  $XY$ .

$\lambda$ , the intensity, is expressed in units to make the expected number of lines intersecting a segment of unit length equal to two.

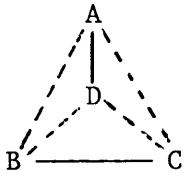
Four non-trivial cases may occur (see e.g. Kendall & Moran (1963) ):-

(1)  
$$\begin{aligned} N_{AB \ \& \ AC} &\sim P(AB+AC-BC) \\ N_{AB \ \& \ BC} &\sim P(AB+BC-AC) \\ N_{AC \ \& \ BC} &\sim P(AC+BC-AB) \end{aligned} \quad \left. \begin{array}{l} ) \\ ) \\ ) \end{array} \right\} \text{independent}$$

$$N_{AB} = N_{AB \ \& \ AC} + N_{AB \ \& \ BC}$$

$$N_{AC} = N_{AB \ \& \ AC} + N_{AC \ \& \ BC}$$

(2)

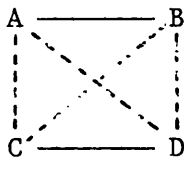


$$\begin{array}{l}
 N_{AD} \ \& \ BC \quad \sim P(2AD+DC+DB-AB-AC) \\
 N_{AD} \ \& \ BC^c \quad \sim P(AB+AC-BD-CD) \\
 N_{BC} \ \& \ AD^c \quad \sim P(2BC-2AD-BD+AB+AC-CD)
 \end{array}
 \left. \begin{array}{l}
 ) \\
 ) \\
 )
 \end{array} \right\} \text{independent}$$

$$N_{AD} = N_{AD} \ \& \ BC + N_{AD} \ \& \ BC^c$$

$$N_{BC} = N_{BC} \ \& \ AD + N_{BC} \ \& \ AD^c$$

(3)

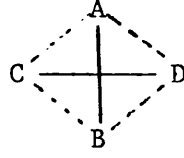


$$\begin{array}{l}
 N_{AB} \ \& \ CD \quad \sim P(AD+BC-AC-BD) \\
 (N_{AB\&AC} + N_{AB\&BD}) \sim P(2AB+AC+BD-BC-AD) \\
 (N_{CD\&AC} + N_{CD\&BD}) \sim P(2CD+AC+BD-AD-BC)
 \end{array}
 \left. \begin{array}{l}
 ) \\
 ) \\
 )
 \end{array} \right\} \text{independent}$$

$$N_{AB} = N_{AB} \ \& \ CD + (N_{AB} \ \& \ AC + N_{AB} \ \& \ BD)$$

$$N_{CD} = N_{CD} \ \& \ AB + (N_{CD} \ \& \ AC + N_{CD} \ \& \ BD)$$

(4)



$$\begin{array}{l}
 N_{AB} \ \& \ CD \quad \sim P(2AB+2CD-AC-CB-BD-DA) \\
 (N_{AC\&AD} + N_{BC\&BD}) \sim P(AC+AD+BC+BD-2CD) \\
 (N_{CA\&CB} + N_{DA\&DB}) \sim P(AC+AD+BC+BD-2AB)
 \end{array}
 \left. \begin{array}{l}
 ) \\
 ) \\
 )
 \end{array} \right\} \text{independent}$$

$$N_{AB} = N_{AB} \ \& \ CD + (N_{AC} \ \& \ AD + N_{BC} \ \& \ BD)$$

$$N_{CD} = N_{CD} \ \& \ AB + (N_{CA} \ \& \ CB + N_{DA} \ \& \ DB)$$

In each case we are interested in the joint distribution of two Poisson variables  $X, Y$  which may be written as:

$$X = Z_1 + Z_3$$

and  $Y = Z_2 + Z_3$

where  $Z_1, Z_2, Z_3$  are independent Poisson variables. But this is precisely the condition specified for the bivariate Poisson

distribution described by Mardia.

The system as a whole will have as its joint distribution a multivariate generalisation of this bivariate Poisson distribution. All realisations of systems of Hamming distances arising from this joint distribution will automatically satisfy the metric inequality. Hamming distance is in fact just one of a large class of dissimilarity functions which do so (Gower 1971) . We call this model for the generation of euclidean-like distances the Poisson hyperplane model. However in practice it is more convenient to condition on the total number of hyperplanes involved, whereupon Poisson distributions become binomial, and it is in this form that we actually realise the model. We give two examples of the dependence of this binomial hyperplane model distance upon euclidean distance in the configuration. Fig. 3.1.1 was produced from 50 hyperplanes; Fig. 3.1.2 derived from 500 hyperplanes. In Fig. 3.1.2 the points lie in a narrow band demonstrating the near-linearity of the two distances, and with less hyperplanes the width of the band is correspondingly greater.

#### Independent Binomial Model

In order to be able to assess the effects of the dependence structure we have described in the binomial hyperplane model we introduce another model in which each individual dissimilarity has the same distribution as in the binomial hyperplane model, but the dependence is removed, producing a model with independent errors. Thus each dissimilarity has a binomial distribution with parameters  $N$ , the number of hyperplanes, and  $p$ , the probability of one hyperplane intersecting the line segment of interest. The dissimilarity-against-distance plots arising from this independent binomial model are visually indistinguishable from those arising in the binomial hyperplane model.

FIG. 3.1.1 Binomial Hyperplane Dissimilarity Plotted Against  
Euclidean Distance; 50 Hyperplanes

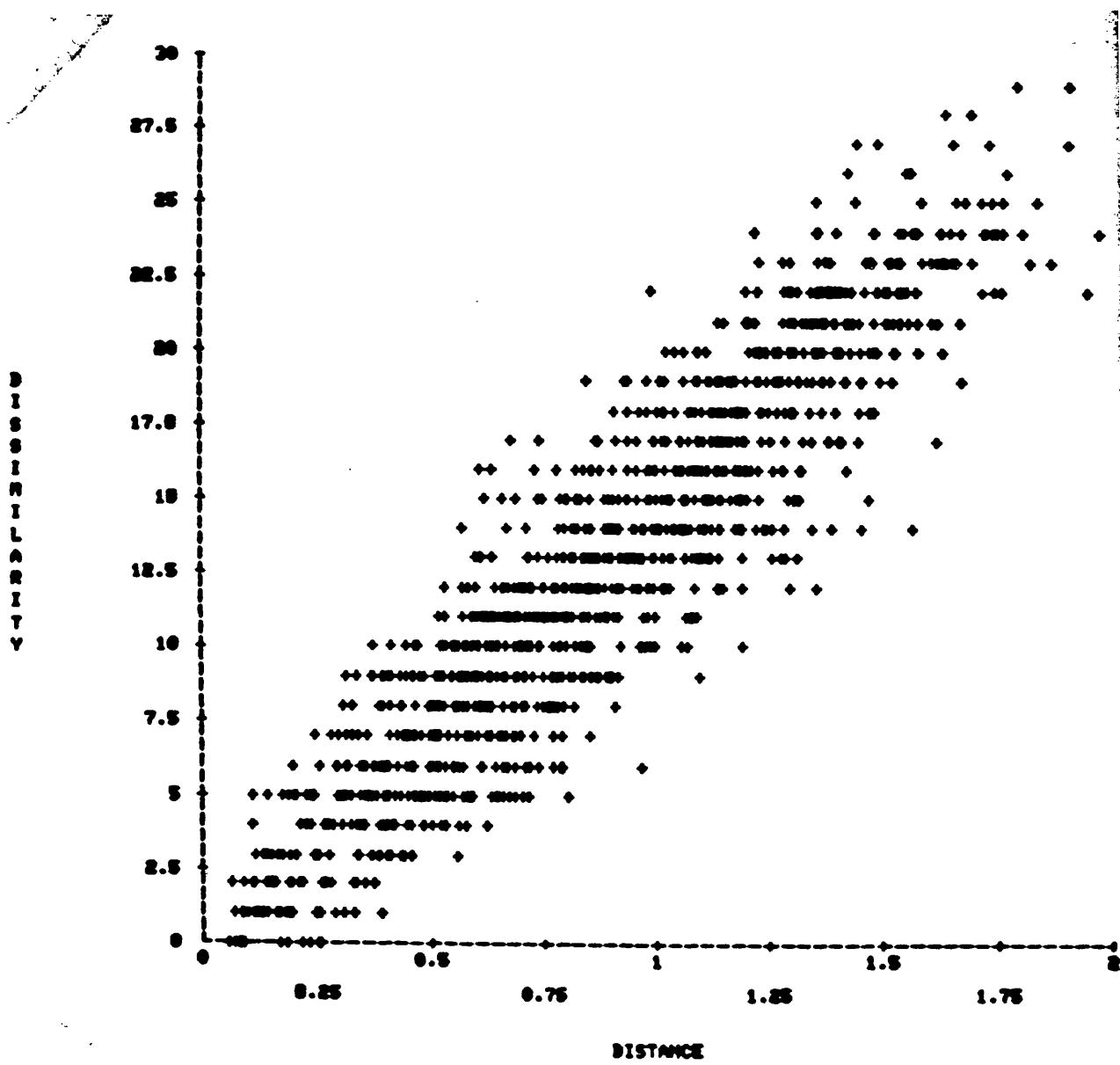
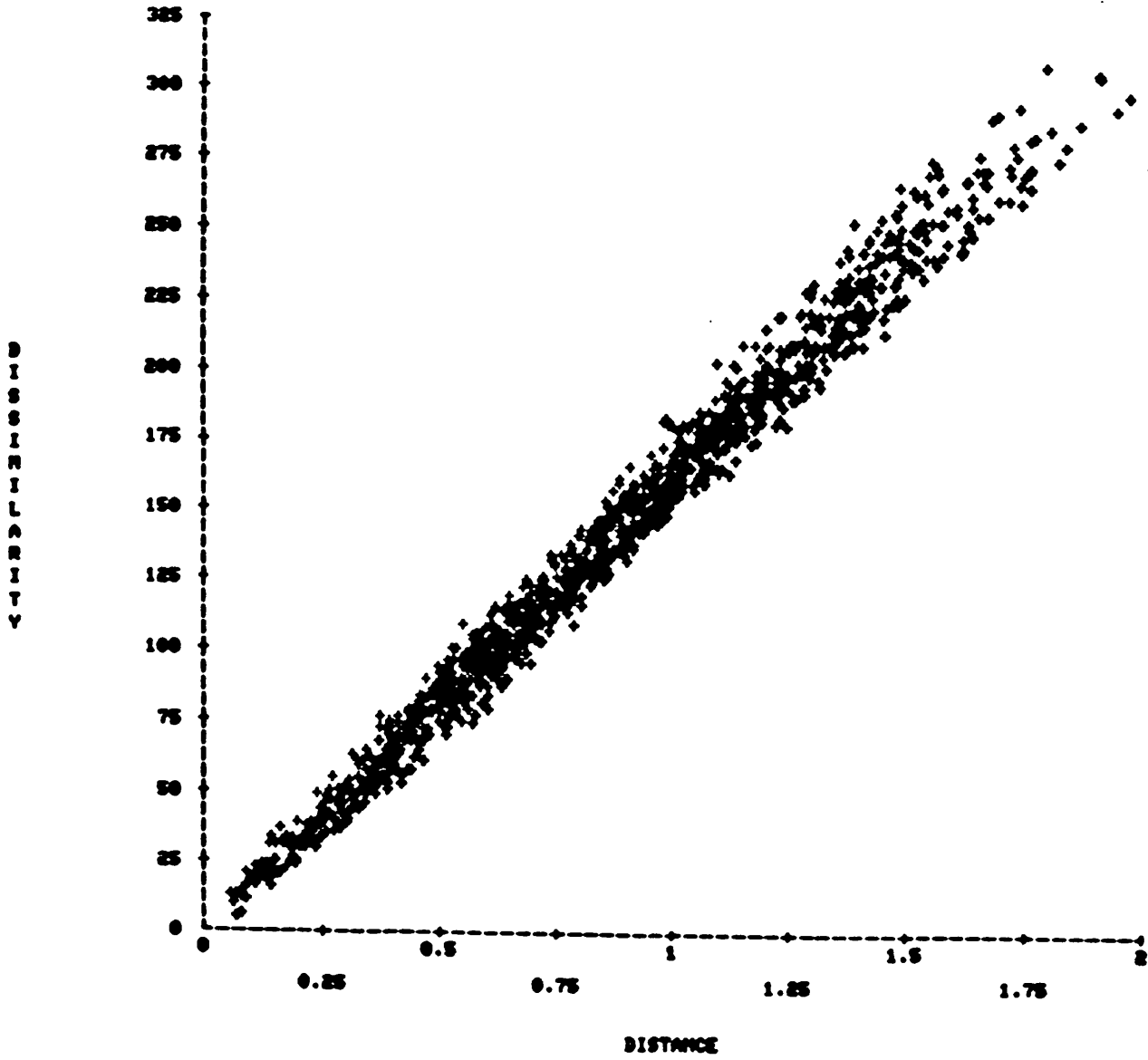


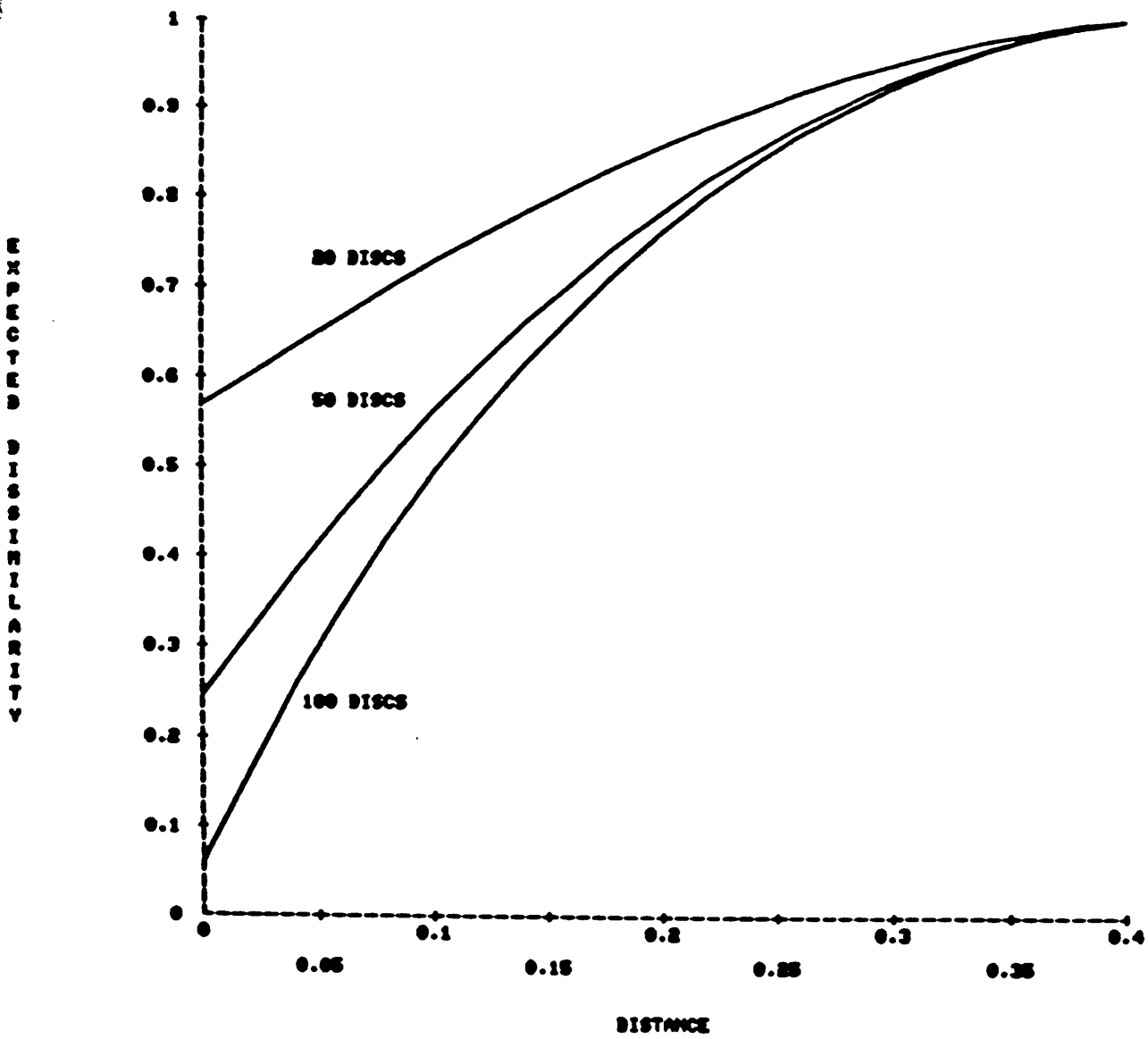
FIG. 3.1.2      Binomial Hyperplane Dissimilarity Plotted Against  
Euclidean Distance; 500 Hyperplanes



### Jaccard Distance Model

The third model is chosen to treat dissimilarities which arise in problems where data are in binary objects-by-attributes form, and where 0 corresponds to absence, 1 to presence. Such data occur in plant ecology, where the objects are sites, and the attributes are plant species which are recorded as either present or absent at each of the sites. A commonly used coefficient in such cases is Jaccard's coefficient, which we use to form our Jaccard distance model. This is obtained by dividing Hamming distance by the number of attributes present in either or both of the two objects under consideration. Jaccard distance also produces a metric, although it takes values in the range zero to one, and so its relationship to euclidean distance certainly cannot be linear. We describe a method of generating random Jaccard distances as follows. Each attribute is "present" over a region of space interior to a disc, whose radius is drawn from some fixed distribution, and whose centre is a point in a realisation of a Poisson point process. Provided that the expected disc area is finite all Jaccard distances will be almost surely well-defined, except when two objects each lie in no discs at all. In this latter case we arbitrarily assign value unity. Any version of this model is characterised by the rate of the Poisson point process and the nature of the radius distribution. For a fixed radius distribution the relationship between euclidean distance and expected Jaccard distance is a monotone one, and has decreasing fluctuation about the mean as the intensity increases. In Fig. 3.1.3 we show the form of the value of the expected Jaccard distance plotted against euclidean distance for a fixed radius of 0.2. The three curves correspond to fixed values of 20, 50 and 100 discs and are based upon exact (up to computer accuracy) calculations.

FIG. 3.1.3 Expected Jaccard Distance Dissimilarity Plotted Against  
Euclidean Distance; 20, 50 and 100 Discs of Fixed  
Radius 0.2



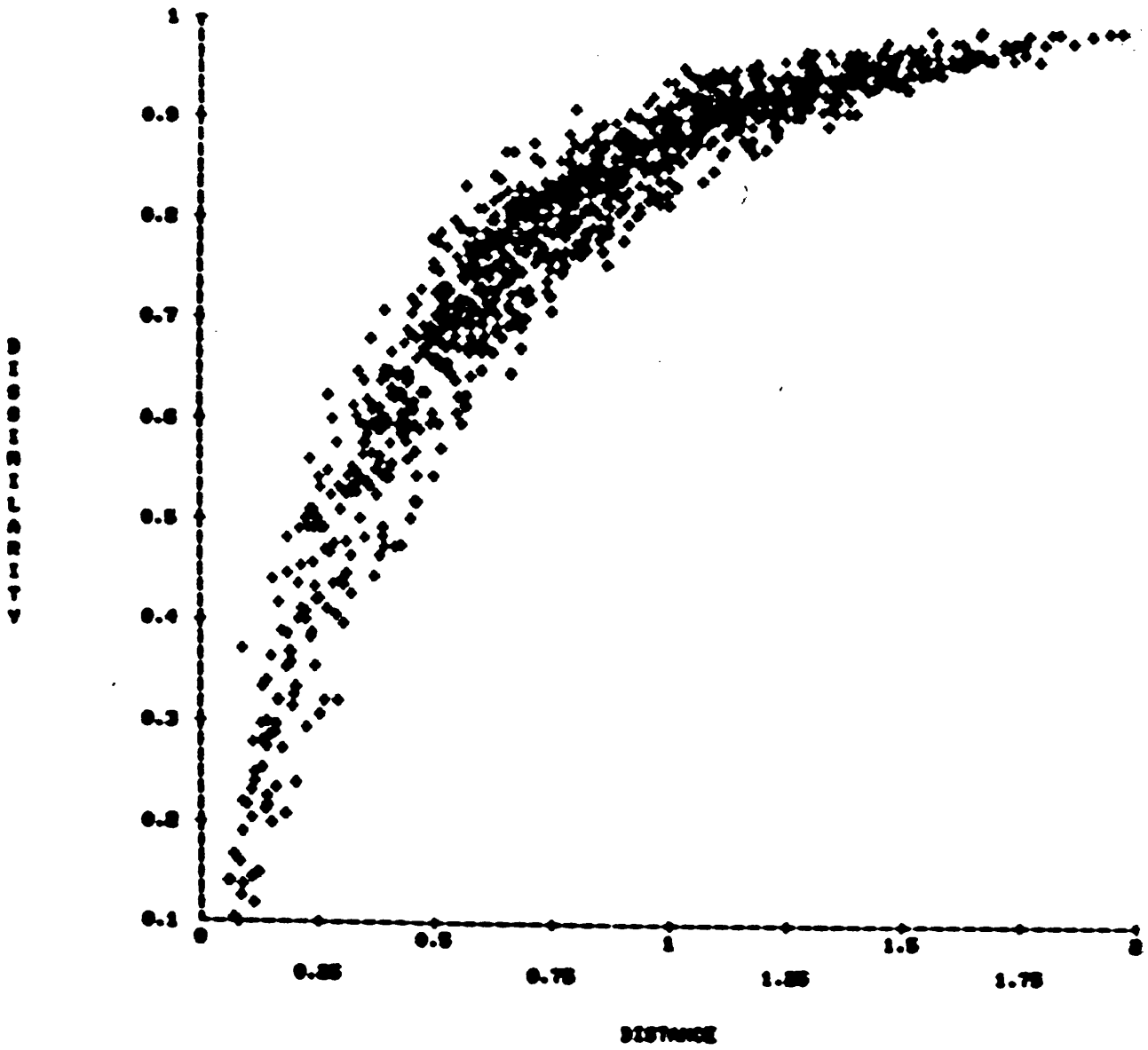


The concave relationship is shown up. The dependence of this monotone relationship on the nature of the radius distribution is calculable in principle, but has no simple form and is better treated as unknown. This model is of particular interest for comparisons between classical and ordinal scaling, for the non-linearity of distance relationship would be expected to be handled more effectively by the ordinal method. In Fig. 3.1.4 we show a typical dissimilarity-versus-distance scatterplot. In practice we use an exponential radius distribution or a constant radius distribution and condition on the total number of discs. In Fig. 3.1.4, 500 discs were used and the exponential radius distribution had mean 0.2. Again the width of the spread about the expected line is inversely related to the number of discs.

#### Wilkinson Metric Model

The final model relates to abuttal data. As derived in Section 1.11, the Wilkinson metric or graph-theoretic distance between two points is the minimum number of contiguities traversed along a path from one point to the other via contiguous points. Contiguity is here defined via the Dirichlet tessellation. The distribution of euclidean distance between contiguous points in a planar Poisson process is known (Miles, 1970; Sibson, 1980) but this knowledge does not extend to points at larger Wilkinson distances. However it appears from simulation that the mean euclidean distance is close to linear with the Wilkinson metric in two dimensions, which for computational reasons is the only currently practicable case. A simulation model may be obtained by taking a fixed configuration of points between which the values are to be calculated, and superimposing on this a realisation of a Poisson process. The Wilkinson metric for the combined configuration may

FIG. 3.1.4     Jaccard Distance Dissimilarity Plotted Against  
Euclidean Distance; 500 Discs with Radius Distribution  
Exponential Mean 0.2



then be calculated from its Dirichlet tessellation. In practice, of course, only finitely many additional points are generated. These are taken over a region larger than that occupied by the original configuration in order that edge effects may be negligible. As the number of additional points increases it appears that, as in the other models, the relative variability of the Wilkinson metric decreases. We generate the model, conditioning on the number of additional points, these being taken from a disc concentric with that containing the original points, but of radius two. In Fig. 3.1.5 we show the effect of 3,200 additional points in this larger disc.

### Preprocessing Techniques

In Section 1.7 we described the reasoning underlying preprocessing techniques, and introduced the two particular transformations corresponding to assumptions of normal and uniform distributions of the parent configuration. It is instructive to see what form of dissimilarity-versus-distance plots such transformations induce. Figs. 3.1.6 and 3.1.7 are based upon two dissimilarity matrices from the Jaccard distance model, using an exponential radius distribution of mean 0.2 with 1000 discs. The two figures correspond to the normal and uniform assumptions respectively. For the normal assumption the plot is much more linear than was the original matrix, however there is a tendency for values at large distance to be too high and this corresponds to the lack of upper bound on the normal distribution. For the correct uniform assumption the linearity of the plot is greater still, and this applies at all levels of distance. This behaviour is not surprising and is not specific to this particular choice of matrices.

FIG. 3.1.5 Wilkinson Metric Dissimilarity Plotted Against  
Euclidean Distance; 3,200 Additional Points in the  
Larger Disc

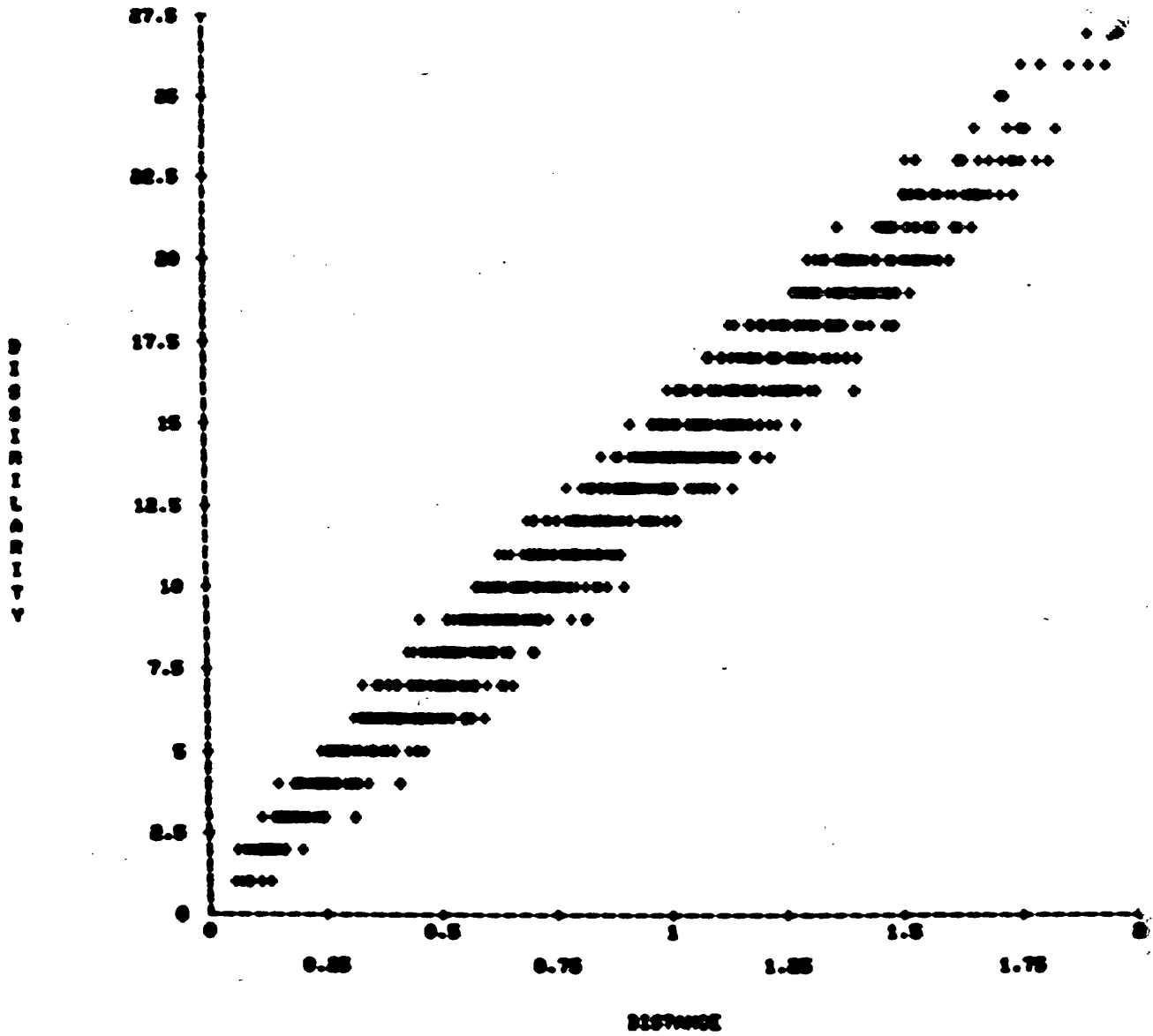


FIG. 3.1.6 Jaccard Distance Dissimilarity Plotted Against Euclidean Distance; 1,000 Discs with Radius  
Distribution Exponential Mean 0.2. Dissimilarities Processed According to the Assumption  
of an Underlying Bivariate Normal Configuration

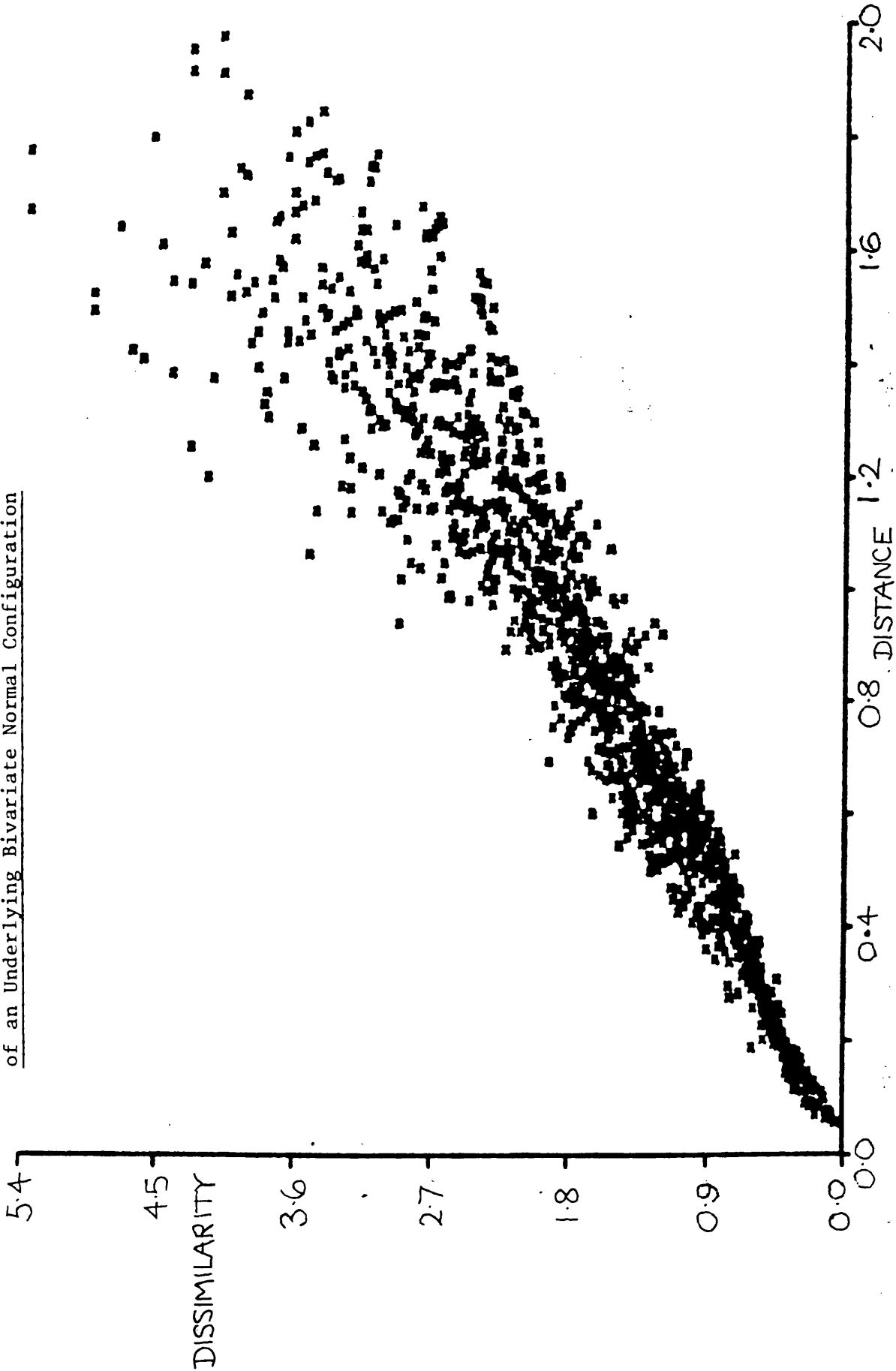
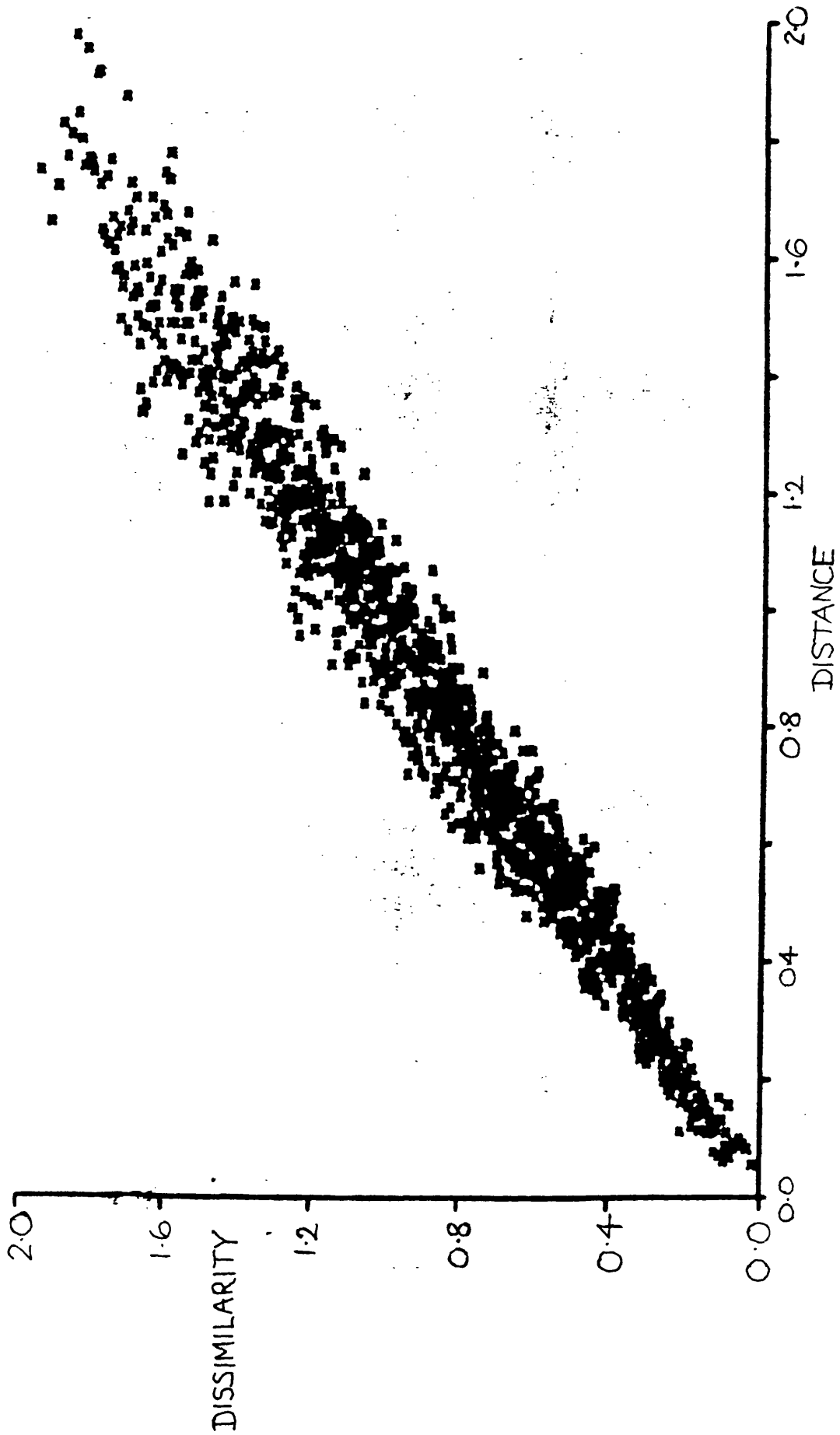


FIG. 3.1.7 Jaccard Distance Dissimilarity Plotted Against Euclidean Distance; 1,000 Discs with Radius  
Distribution Exponential Mean 0.2. Dissimilarities Processed According to the Assumption  
of an Underlying Uniform Configuration



### 3.2 Simulation Studies of Classical Scaling

Each of the four euclidean models derived in the previous section is now used to assess the robustness of classical scaling. We use procrustes statistics to measure the degree of departure from the original configuration. In these studies we also examine the effect of the trace criterion, magnitude criterion and preprocessing transformations.

#### Design

Altogether 976 trials of classical scaling are reported in this section. The design is summarised in Table 3.2.1. Five two-dimensional configurations are used; one each in three, four, five and six dimensions. Six levels of error are allowed for each model by varying the number of hyperplanes, discs or points used to generate values. To compare different numbers of dimensions we make the standardisation that the expected number of hyperplanes intersecting a line segment of given length should be a constant. Thus in six dimensions "1000 hyperplanes" should be interpreted as that number of hyperplanes which will give the same expected number of cuts of a line segment of length  $\ell$ , as would 1000 hyperplanes in two dimensions; generally this number will be higher as the number of dimensions increases, for in  $n$  dimensions the probability that a hyperplane cuts a line segment of length  $\ell$  is given by

$$\frac{\ell \Gamma\{\frac{1}{2}n\}}{(n-1)\sqrt{\pi} \Gamma\{\frac{1}{2}(n-1)\}}$$

For the Jaccard distance model the radius of discs that cut the unit disc is given an exponential distribution with mean either 0.2 or 1.0. It is not necessary that the centre of the disc should lie within the

TABLE 3.2.1 Numbers of Replications for Model, Number of Dimensions, Configuration and Error Level

Model	No. of Dimensions & Configuration	No. of Hyperplanes					
		20	50	100	200	500	1000
Binomial Hyperplane	2 dims., first configuration	10	10	10	10	10	10
	" " second	10	10	10	10	10	10
	" " third	10	10	10	10	10	10
	" " fourth	10	10	10	10	10	10
	" " fifth	10	10	10	10	10	10
	3 " only	10	10	10	10	10	10
4 " "	10	10	10	10	10	10	
5 " "	10	10	10	10	10	10	
6 " "	10	10	10	10	10	10	
Independent Binomial	2 dims., third configuration	10	10	10	10	10	10
	6 " only	10	10	10	10	10	10
Jaccard Distance	2 dims., first configuration (exponential, mean radius 1.0)	10	10	10	10	10	10
	2 dims., first configuration (exponential, mean radius 0.2)	10	10	10	10	10	10
Processed Jaccard Distances (Normal Assumption)	2 dims., first configuration (exponential, mean radius 1.0)	5	5	5	5	5	0
	2 dims., first configuration (exponential, mean radius 0.2)	10	10	10	10	10	10
Wilkinson Metric	$\sqrt{(50 + \frac{1}{2} \text{ extra points})}$	7.1	11.2	15.0	18.0	21.2	26.5
	2 dims., first configuration	1	20	20	20	20	30



unit disc. For the Wilkinson metric model extra points are added to the centre disc of radius two, the combination being tessellated in the square  $-4 \leq x, y \leq 4$ . Both of these are precautionary measures to minimise edge effects. Since the Wilkinson distances increase approximately as the square root of the number of points in the unit disc, we may use  $\sqrt{(50 + \frac{1}{4} \text{ extra points})}$  to measure the level. Preprocessing is also assessed, using the assumption (which is incorrect) of an underlying bivariate normally distributed configuration. This is done on different matrices from the unprocessed version, so that comparisons are unmatched.

### Results

Mean values and sample standard deviations for the procrustes statistic are provided in Table 3.2.2. Corresponding log/log plots of procrustes statistic against error level are shown in Fig. 3.2.3 (binomial hyperplane), Fig. 3.2.4 (independent binomial), Fig. 3.2.5 (Jaccard Distance), Fig. 3.2.6 (processed Jaccard distance) and Fig. 3.2.7 (Wilkinson metric).

Overall Impressions. The distribution of the procrustes statistic is known to be approximately of a general  $\chi^2$  in type (Sibson, 1979), so it is not surprising that it is quite skewed, with occasional values being very high. The range and standard deviations of the replications are thus quite large. A few general rules stand out. The mean value increases with dimensions, decreases as the rate of the underlying Poisson process increases and is always smaller for the larger of the two Jaccard distance model disc radii.

Binomial Hyperplane Model. Each of the doubly logarithmic plots is remarkably linear with slope close to  $-1$ , indicating that the dominant term in the procrustes statistic is constant/number of hyperplanes. The actual value of the constant is dependent upon the

TABLE 3.2.2 Results from Classical Scaling. Mean Procrustes Statistic (and Sample Standard Deviation)

Model	No. of Hyperplanes		No. of Discs			
	20	50	100	200	500	1000
Binomial Hyperplane						
2,1	0.1145 (0.0348)	0.0495 (0.0167)	0.0284 (0.0111)	0.0144 (0.0060)	0.0054 (0.0028)	0.0033 (0.0020)
2,2	0.1786 (0.0602)	0.0677 (0.0261)	0.0382 (0.0109)	0.0182 (0.0051)	0.0063 (0.0019)	0.0028 (0.0012)
2,3	0.1695 (0.1105)	0.0532 (0.0191)	0.0272 (0.0106)	0.0149 (0.0050)	0.0058 (0.0022)	0.0022 (0.0008)
2,4	0.1131 (0.0455)	0.0476 (0.0161)	0.0249 (0.0055)	0.0141 (0.0036)	0.0068 (0.0031)	0.0025 (0.0009)
2,5	0.1336 (0.0360)	0.0628 (0.0305)	0.0247 (0.0072)	0.0122 (0.0084)	0.0056 (0.0018)	0.0031 (0.0011)
3	0.1642 (0.0623)	0.0727 (0.0160)	0.0376 (0.0106)	0.0206 (0.0057)	0.0082 (0.0032)	0.0039 (0.0010)
4	0.2288 (0.0601)	0.0951 (0.0190)	0.0440 (0.0078)	0.0252 (0.0080)	0.0111 (0.0024)	0.0055 (0.0008)
5	0.2806 (0.0620)	0.1166 (0.0177)	0.0578 (0.0120)	0.0303 (0.0043)	0.0136 (0.0018)	0.0060 (0.0008)
6	0.2974 (0.0534)	0.1387 (0.0299)	0.0699 (0.0097)	0.0375 (0.0047)	0.0142 (0.0025)	0.0069 (0.0005)
Independent Binomial						
2,3	0.0494 (0.0081)	0.0184 (0.0033)	0.0093 (0.0013)	0.0049 (0.0009)	0.0018 (0.0004)	0.0009 (0.0001)
6	0.3923 (0.0237)	0.1737 (0.0335)	0.0764 (0.0086)	0.0367 (0.0040)	0.0142 (0.0014)	0.0070 (0.0007)
Jaccard (mean rad. 1.0)						
2,1	0.1498 (0.0358)	0.0995 (0.0359)	0.0532 (0.0098)	0.0318 (0.0091)	0.0166 (0.0022)	0.0142 (0.0021)
Distance (mean rad. 0.2)						
2,1	0.5895 (0.1871)	0.2994 (0.1230)	0.2435 (0.1804)	0.1228 (0.0249)	0.1022 (0.0126)	0.0883 (0.0132)
Processed Jaccard (mean rad. 1.0)						
2,1	0.1795 (0.0471)	0.1121 (0.0608)	0.0576 (0.0198)	0.0407 (0.0073)	0.0211 (0.0021)	---- (----)
Jaccard (mean rad. 0.2)						
2,1	0.5888 (0.1881)	0.2624 (0.0831)	0.1570 (0.0413)	0.0896 (0.0271)	0.0650 (0.0224)	0.0421 (0.0119)
Wilkinson Metric						
2,1	√(50 + 1/2 extra points) 7.1	11.2	15.0	18.0	21.2	26.5
2,1	0.0579 (----)	0.0398 (0.0083)	0.0282 (0.0083)	0.0224 (0.0061)	0.0186 (0.0051)	0.0141 (0.0047)

FIG. 3.2.3      Log Mean Procrustes Statistic Plotted Against Log  
Hyperplanes for Classical Scaling Simulations of  
Binomial Hyperplane Model

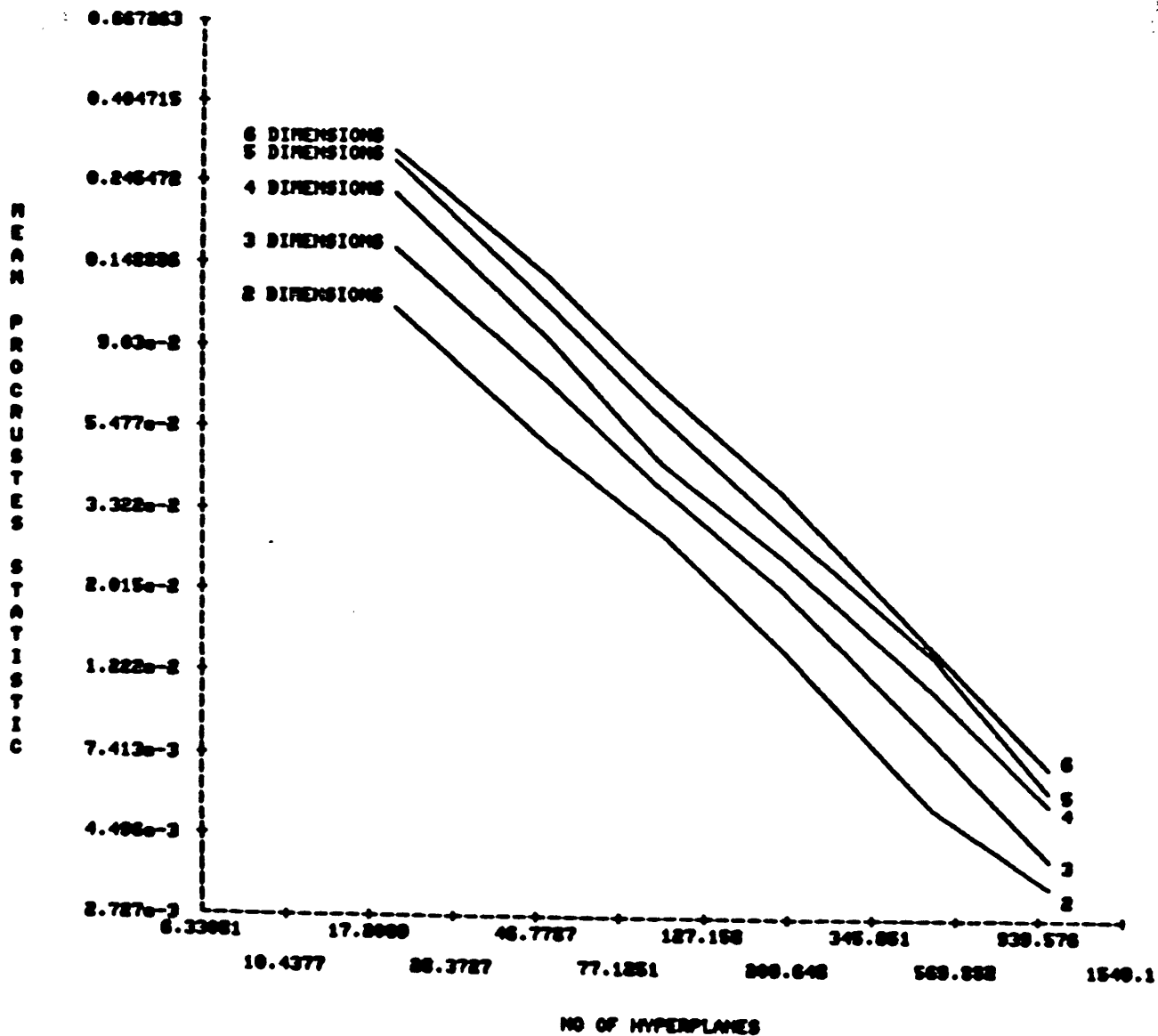


FIG. 3.2.4 Log Mean Procrustes Statistic Plotted Against Log  
'Hyperplanes' for Classical Scaling Simulations of  
Independent Binomial Model

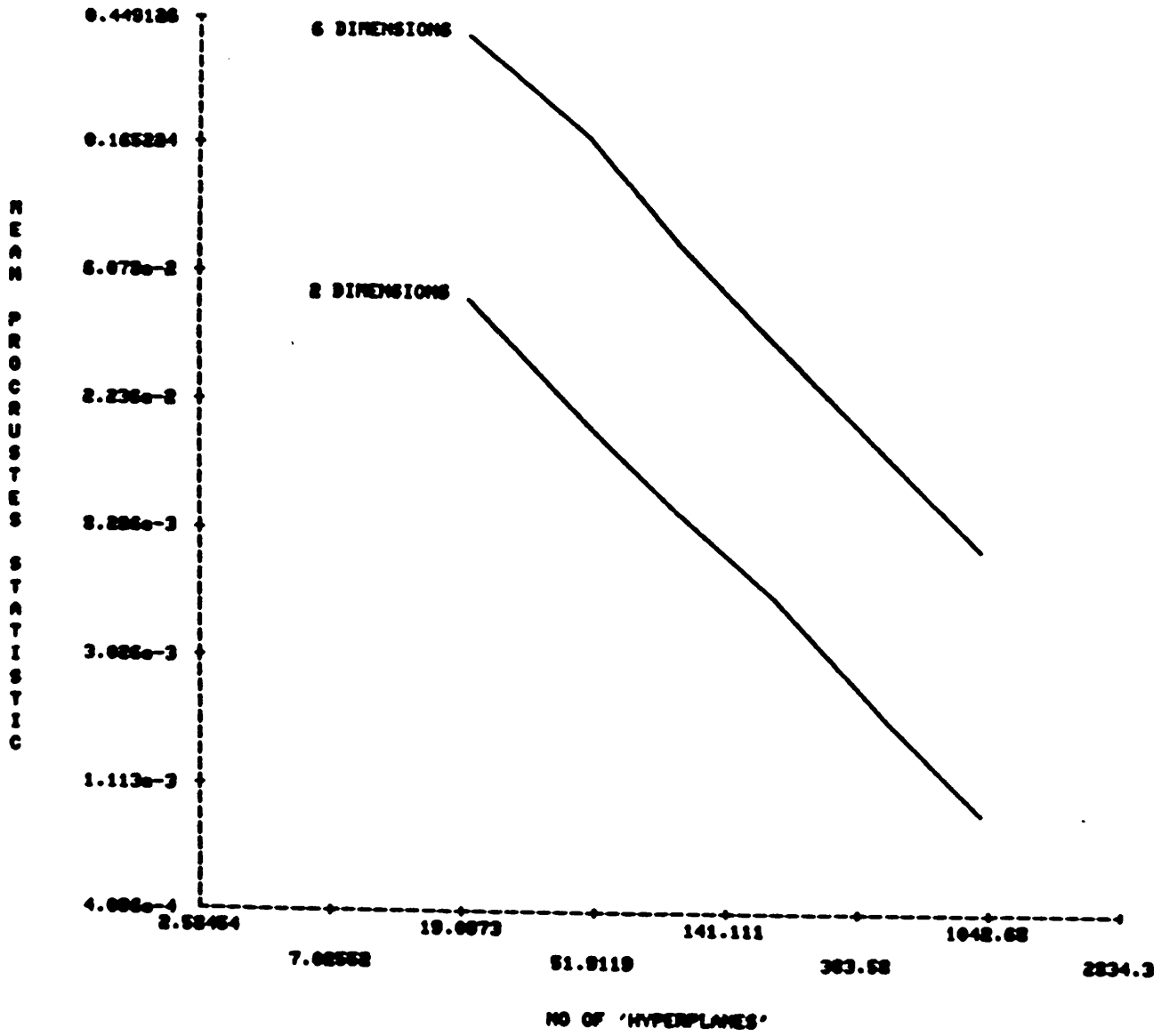


FIG. 3.2.5      Log Mean Procrustes Statistic Plotted Against Log  
Discs for Classical Scaling Simulations of Jaccard  
Distance Model

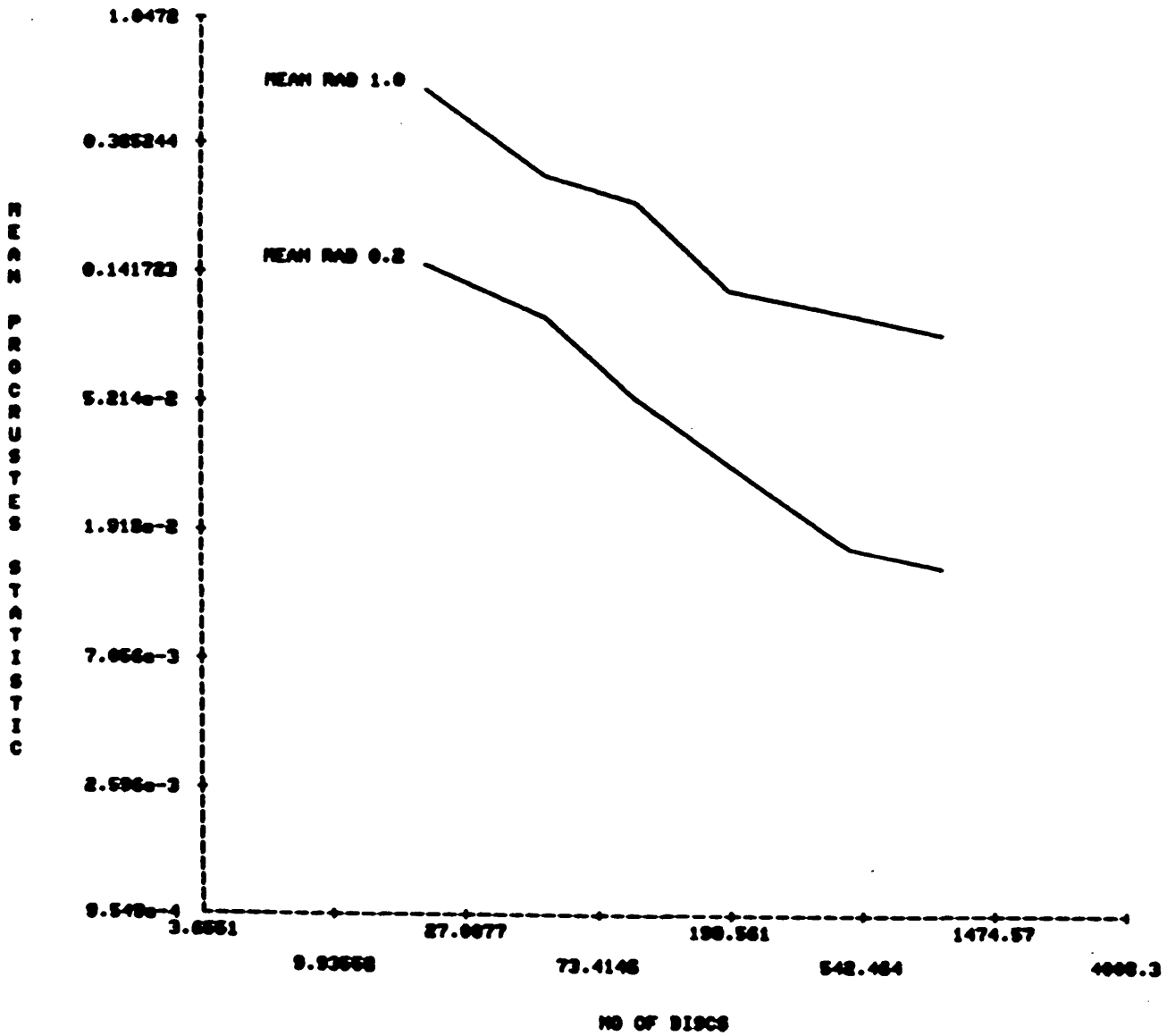


Fig 3.2.6: Processed Jaccard Distances

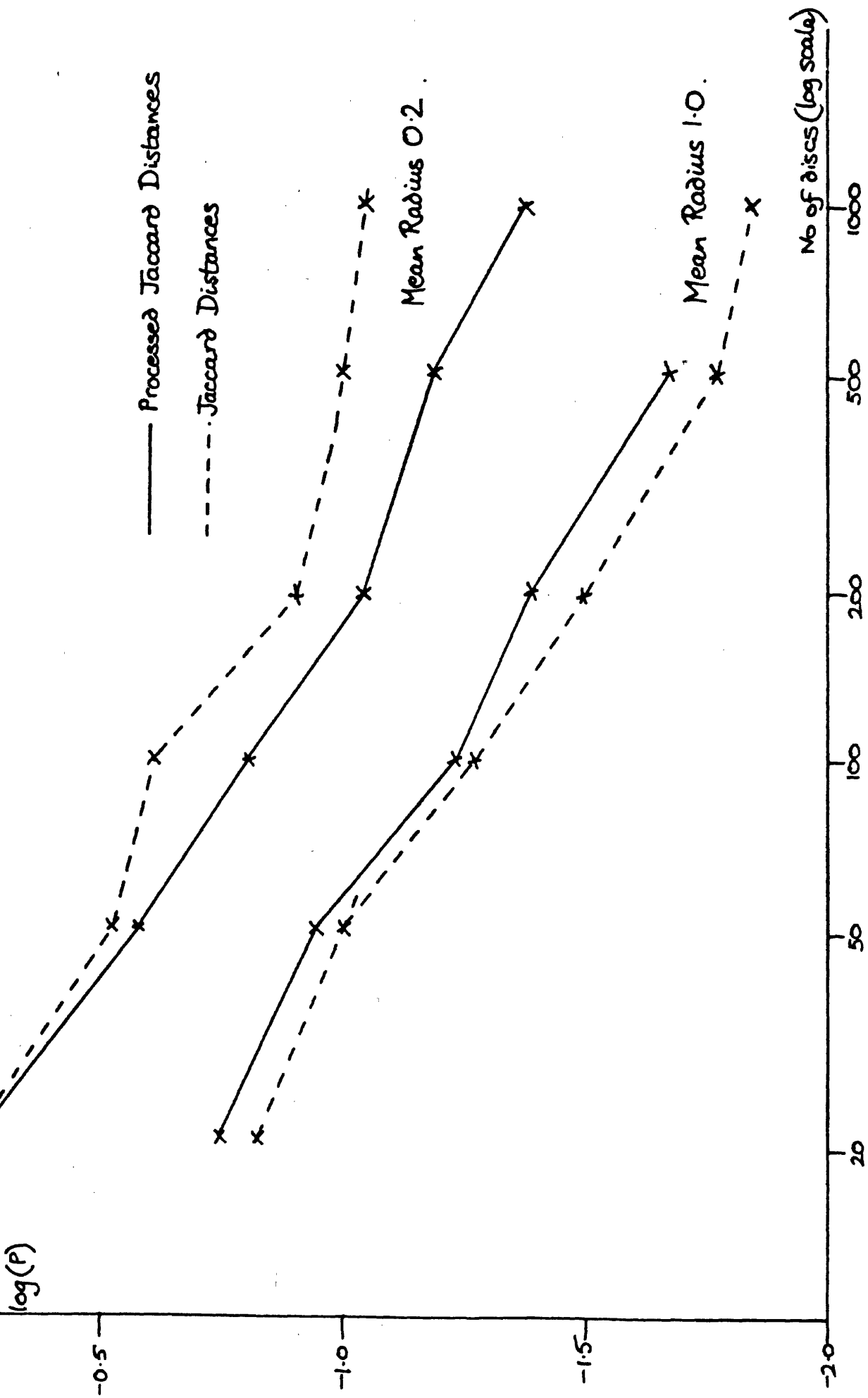
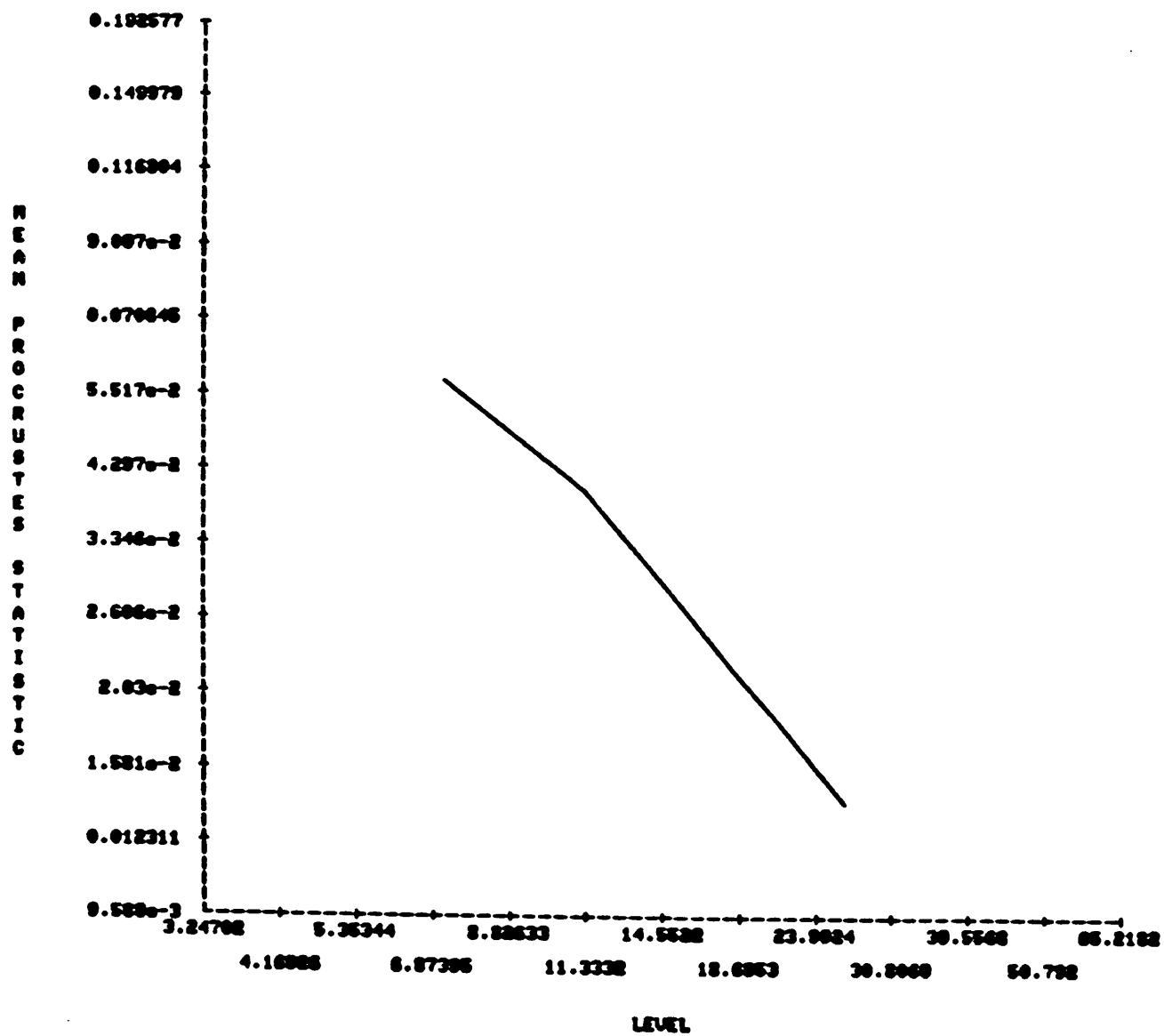


FIG. 3.2.7 Log Mean Procrustes Statistic Plotted Against Level  
for Classical Scaling Simulations of Wilkinson Metric  
Model



dimensionality, the standardisation used to compare dimensions, and the procrustes statistic normalisation used. It would seem that the procrustes statistic could be made arbitrarily small with any specified probability less than one, given sufficient hyperplanes.

Independent Binomial Model. Here again there is emphatic "constant/number of hyperplanes" behaviour for both dimensionalities. However the constant is quite different from that found in the binomial hyperplane model. For the two-dimensional configurations it is smaller by a factor of about two. For the six-dimensional configuration it is slightly larger, although there is a tendency for the difference to diminish with more hyperplanes. The only possible cause for the inferior performance of the binomial hyperplane model in two dimensions is the covariance structure which must act to reduce the available information about true interpoint distance.

Why does this effect not work for six dimensions? As we move from two to just three dimensions we observe a slackening of the constraints upon the line cutting process. For example, in two dimensions three non-collinear line segments may be arranged so that a line may pass through at most two of them. However in three dimensions a plane may pass through any three such segments. As the dimensionality of the space increases such freedoms increase and the covariance structure is much weaker. On the other hand if we reduce the dimensionality and consider what would happen in one dimension, we see that the 1225 entries of the data matrix are entirely determined by the 49 Poisson (or binomial) random variables defining the number of cuts between neighbours of the configuration on the interval  $(-1,1)$ . Thus the 49 random variables contributing to the Poisson hyperplane model may be compared with the 1225 random variables producing the independent



binomial model, which has therefore much more information and is likely to provide a more satisfactory scaling solution.

Jaccard Distance Model. Here the doubly logarithmic plot is much flatter than before. As the number of discs increases the dissimilarities become closer to their expected values, which are not linear with configuration distance, so that it is uncertain that classical scaling will ever be able exactly to reproduce the original configuration. It is not practical for us to allow enough discs to gain a clear impression as to whether the procrustes statistic may become arbitrarily small, or will remain above a fixed level dependent upon the disc radius distribution. The limitations of classical scaling are seen most clearly for this model, and the improvements provided by other models are demonstrated in the following section.

Wilkinson Metric Model. Here again the doubly logarithmic plot is quite linear with slope about -1. This has been achieved by transforming the number of extra points, so that the behaviour is constant/ $\sqrt{\text{no. of extra points}}$ . Again it seems likely that the procrustes statistic could be made arbitrarily small.

Processed Jaccard Distances. For discs of mean radius 1.0 the processing does not improve the mean procrustes statistic. Two factors contributing to this are the incorrect assumption about the underlying configuration, and the lower values that are obtained for the discs of mean radius 1.0 as compared with those of mean radius 0.2. For these latter discs quite considerable improvements are made, for the original values are that much further from euclidean. This may be deduced from the fact that for discs of constant radius 1.0 the intercepts of the expected Jaccard distance curves are pulled down towards the origin relative to those of Fig. 3.1.3. Thus there is

evidence in favour of preprocessing a matrix that is known to be far from euclidean.

Eigenvalue Spectra. When scaling in practice it is normal to inspect the eigenvalue spectrum in order to gauge how many dimensions are required for the solution. We provide an impression of how the spectra appear in the presence of differing amounts of error in the following figures: Fig. 3.2.8. (binomial hyperplane model, 2 dimensions), Fig. 3.2.9 (binomial hyperplane model, 6 dimensions), Fig. 3.2.10 (independent binomial model, 2 dimensions), Fig. 3.2.11 (independent binomial model, 6 dimensions), Fig. 3.2.12 (Jaccard distance model) and Fig. 3.2.13 (Wilkinson metric model). The form of each of these is as follows. The loading on each of the leading nine eigenvalues is plotted for six levels of error where the loading is defined as the percentage of the sum of the first nine eigenvalues, averaged over replications. Direct comparison of Fig. 3.2.8. and Fig. 3.2.10 shows that the binomial hyperplane model produces a more clearly defined configuration dimensionality. The same is seen by comparing Fig. 3.2.9 and Fig. 3.2.11. If hyperplanes are sparse in some direction, dissimilarities measured perpendicular to this direction tend to be small, and the resultant configuration is more one-dimensional than the original. This accounts for the frequent overloading on the first dimension at high error levels of the binomial hyperplane model, which does not occur in the independent equivalent. These two major differences can only be attributed to the correlation structure. For the Jaccard distance model there is a different sort of behaviour. The perturbed zero eigenvalues do not die away so quickly and, as the number of discs increases, more of them become positive. This latter effect would also be observed if an increasing constant term was

Fig 3.2.8

Eigenvalue Spectra.

Binomial Hyperplane Model.

Two dimensions.

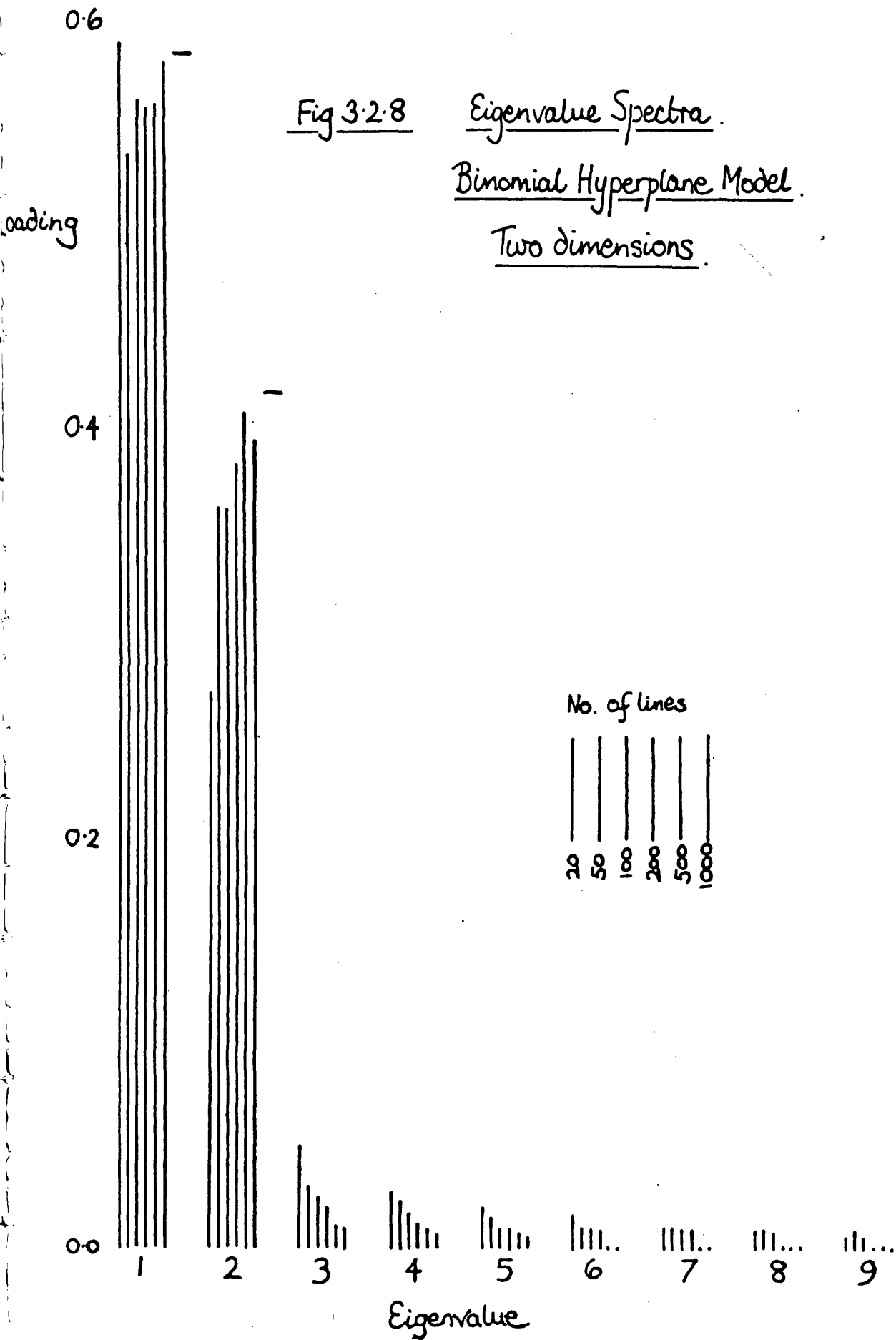


Fig 3.2.9

Eigenvalue Spectra  
Binomial Hyperplane Model  
Six dimensions

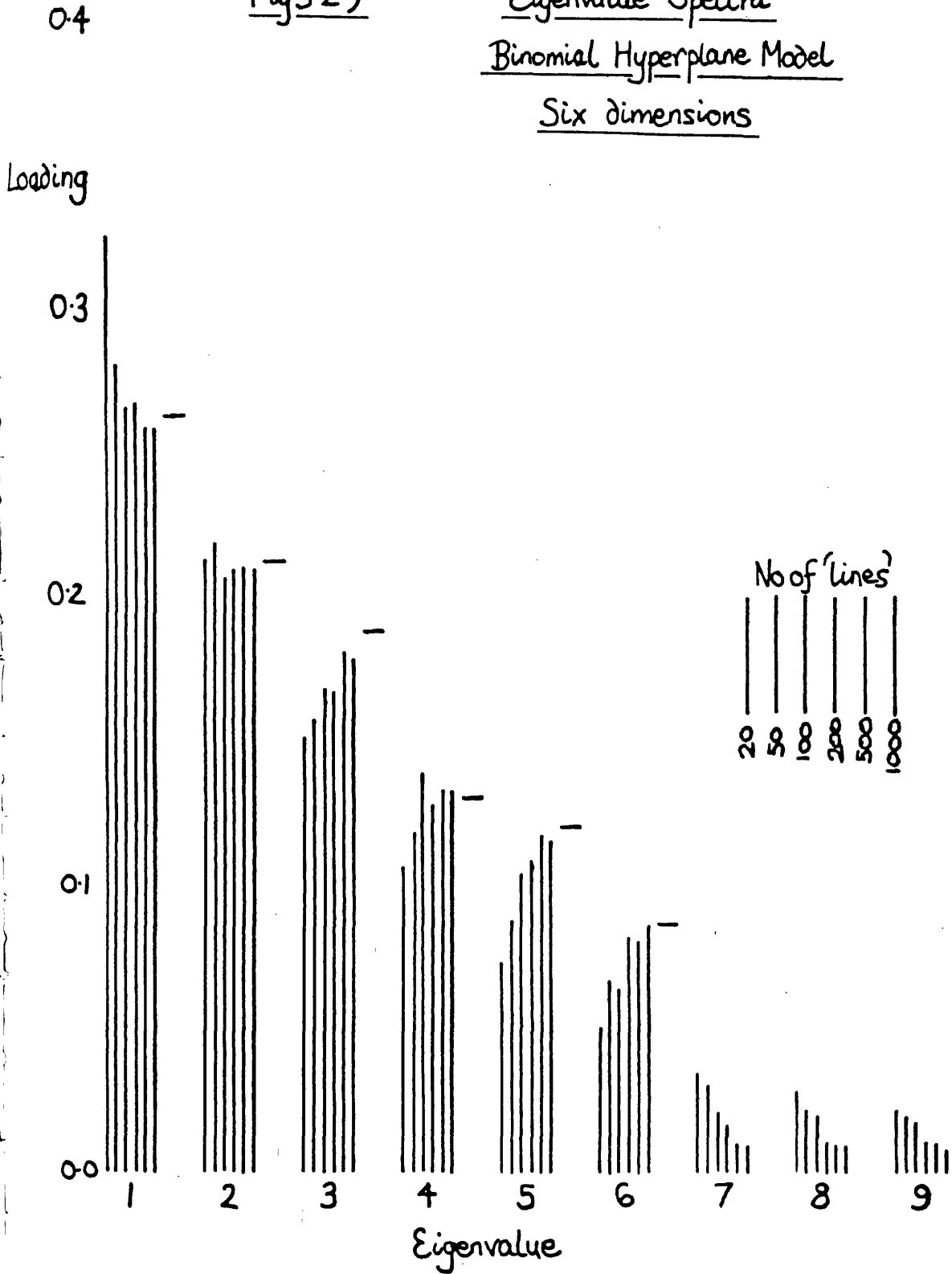
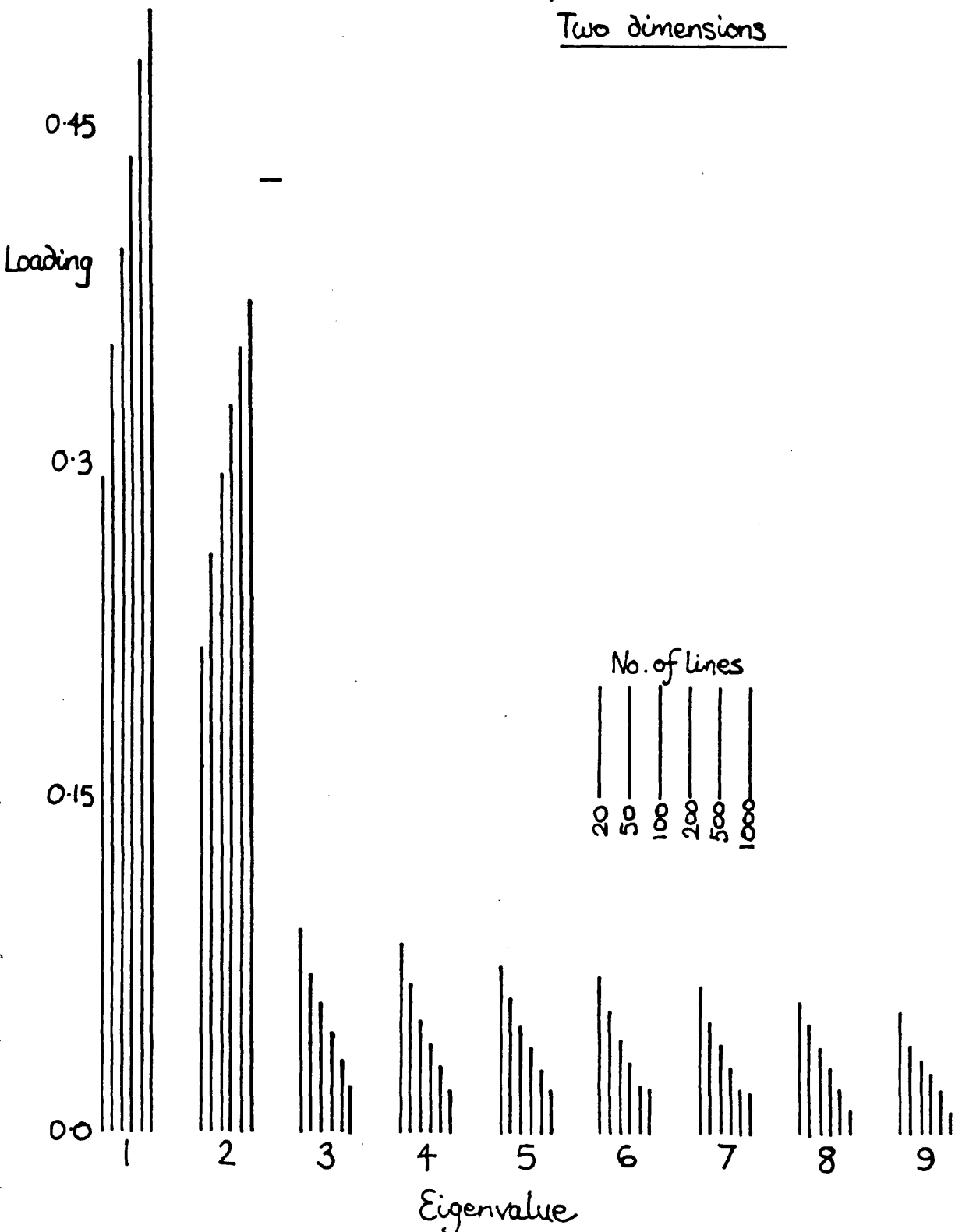


Figure 3.2.10

Eigenvalue Spectra  
Independent Binomial Model  
Two dimensions



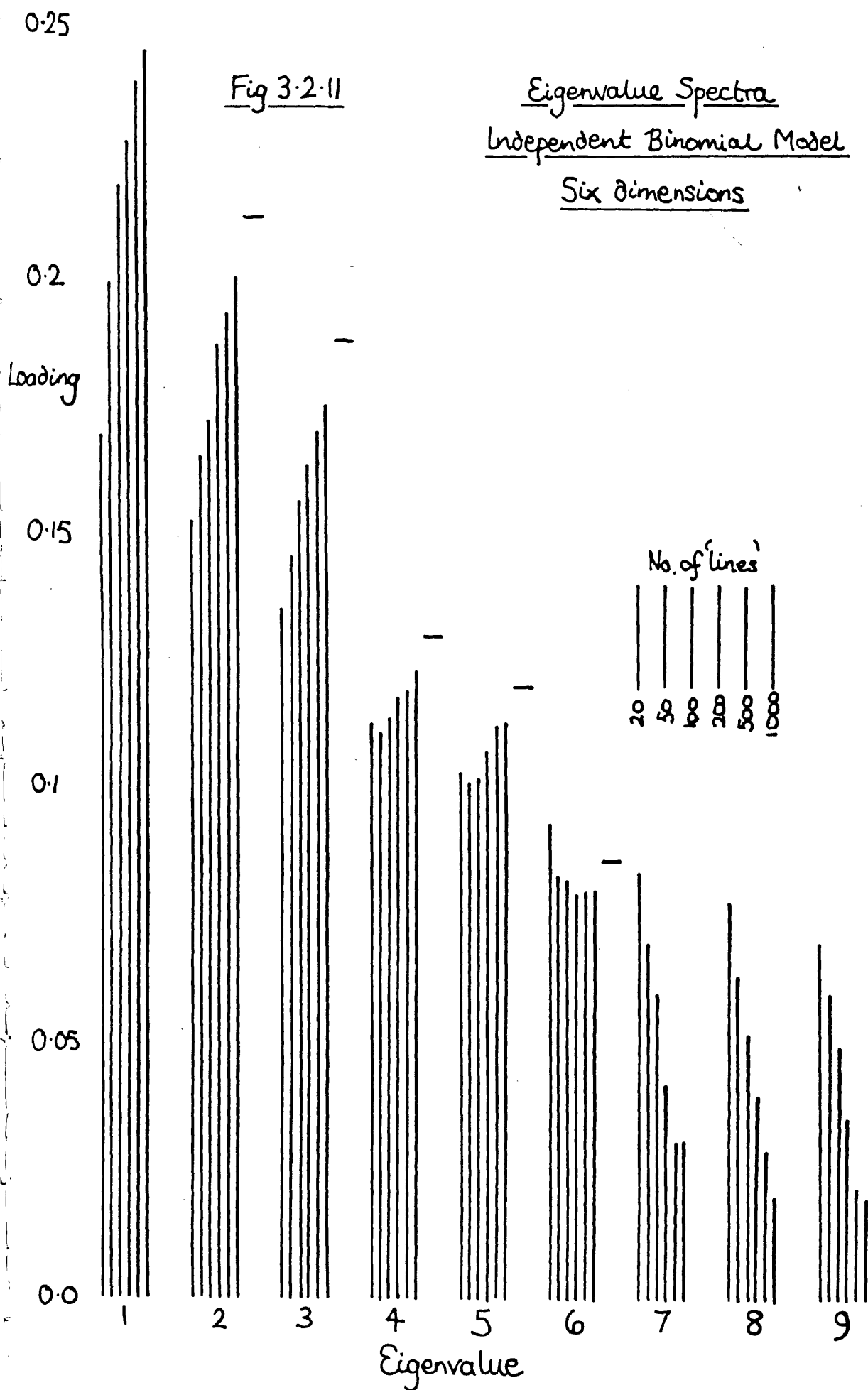


Fig 3.2.12

Eigenvalue Spectra  
Jaccard Distance Model  
Two dimensions

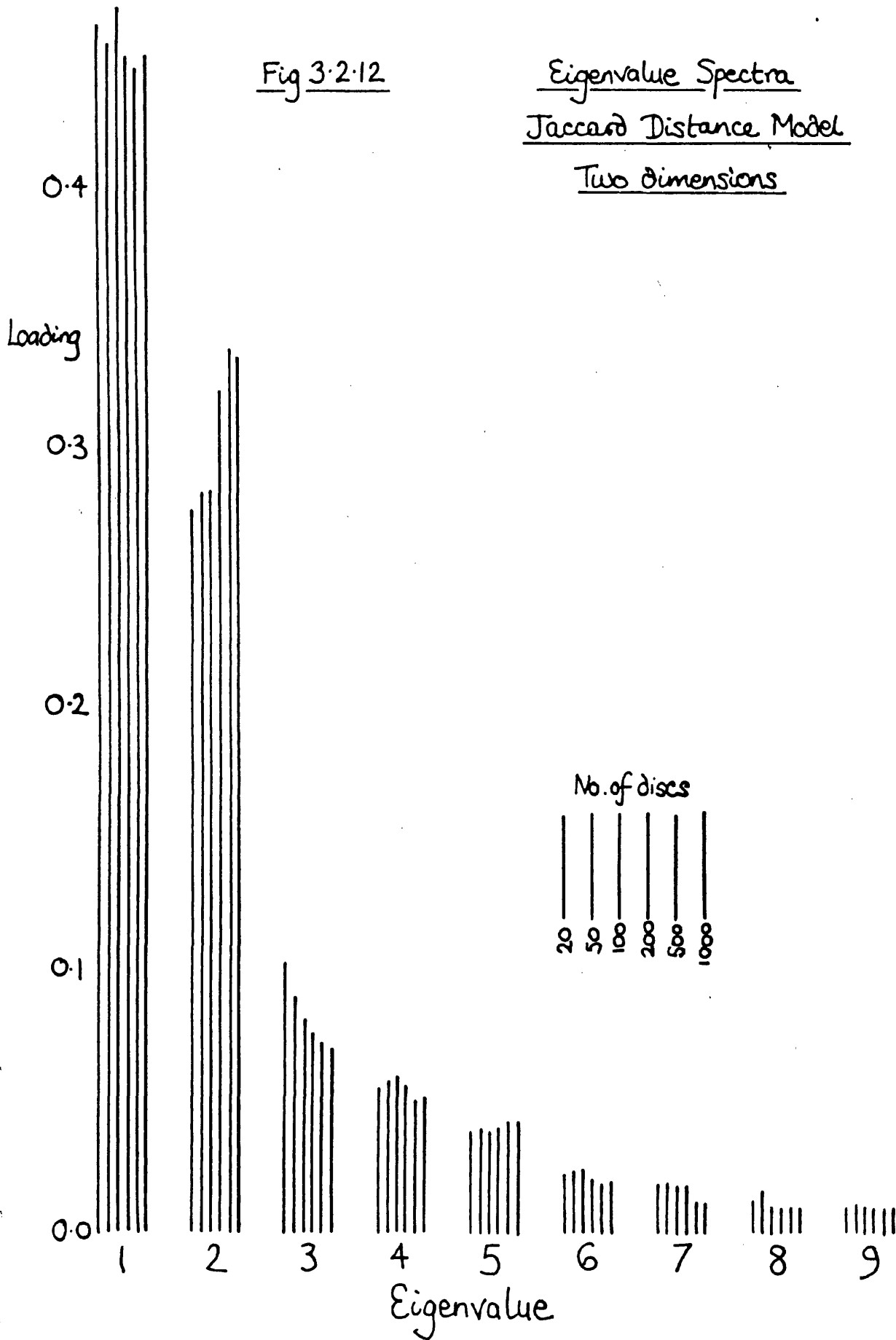
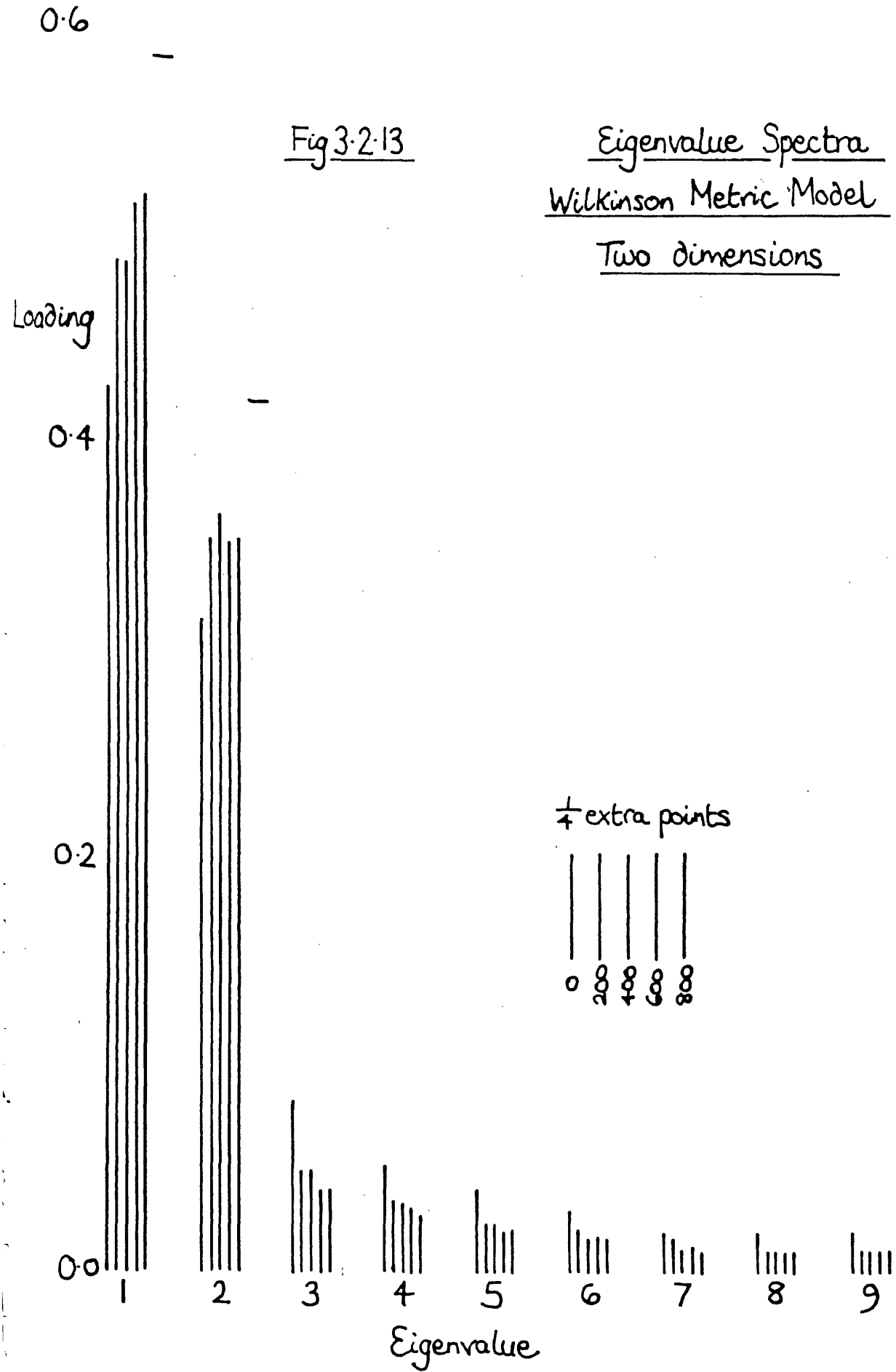


Fig 3.2.13

Eigenvalue Spectra  
Wilkinson Metric Model  
Two dimensions





added to all dissimilarities, although in that case all eigenvalues would become similar in magnitude, even the genuine positive ones. Here intermediate dissimilarities are being forced to be larger than would be anticipated by the concave dissimilarity/true configuration distance relationship. For the Wilkinson model slow convergence is seen. When there is loading left in the higher dimensions this corresponds to additional information, both useful and noisy, that can be used by ordinal scaling, as the comparative results will show.

Trace and Magnitude Criteria. The trace criterion for determining the true dimensionality of a configuration suggests that the sum of genuine positive eigenvalues ought to be approximately equal to the sum of all the eigenvalues. The magnitude criterion suggests that any positive eigenvalue whose magnitude does not substantially exceed that of the largest negative eigenvalue should be rejected as spurious. Here we have an ideal test for these ideas for we know the true configuration dimensionality. In Table 3.2.14 we provide the most common estimate of dimensionality for all of the combinations of model and error. The criteria are applied strictly in that the closest approximation to total trace is used for the first; the eigenvalues larger in magnitude than the most negative are used for the second. It is doubtful whether such an approach would be adopted for the magnitude criterion in practice. A multiplicative factor of at least two would probably be used. The results show the need for the dissimilarity/distance relationship to be roughly linear for this approach to work well. For the binomial hyperplane and independent binomial models, where this is so, there is some success, particularly as the underlying intensities increase. The magnitude criterion is

TABLE 3.2.14 Trace and Magnitude Criteria. Most Commonly Inferred Dimensionality;  
Trace Criterion First, then Magnitude

Model	No. of Dims.	No. of Hyperplanes					
		20	50	100	200	500	1000
Binomial Hyperplane	2	1,2	2,2	2,2	2,2	2,2	2,2
	3	2,2	2,3	2,3	3,3	3,3	3,3
	4	3,4	3,4	3,4	3/4,4	4,4	4,4
	5	3,5	4,5	4,5	4,5	4,5	5,5
	6	3,5	4,6	5,6	5,6	5,6	6,6
	Independent	2	2,2	2,2	2,2	2,2	2,2
Binomial	6	4,4	5,5	5,5	5,5/6	5,6	6,6
No. of Discs							
Jaccard (mean rad. 1.0) Distance (mean rad. 0.2) Processed (mean rad. 1.0) Jaccard (mean rad. 0.2)	2	3/4,4	5,6	7,8	9,10	12,13	15,16
	2	11,13	15,17	20,22	26,26	34,33	39,38
	2	2,2	2,2	2,2	2,2	2,2	---
	2	6,6	4,4	3,3	2,2	2,2	2,2
$\sqrt{(50 + \frac{1}{2} \text{ extra points})}$							
Wilkinson Metric	2	7.1	11.2	15.0	18.0	21.2	26.5
	2	2,2	2,1	3,3	3,2	3,3	3,3

marginally better. For the Wilkinson metric model the methods are less successful, often identifying a third dimension. However for the Jaccard distance model, the results are affected by the increasing numbers of positive perturbed zeros and are increasingly wayward. If the two-dimensional normal assumption is made, then these results are reversed, for the set of dissimilarity values are forced to look two-dimensional, and the criteria reflect this.

### 3.3 Comparison of Scaling Methods

We now turn to consider other scaling methods, and use the four probabilistic models and procrustes statistics to compare the relative accuracy of the configurations recovered by them.

#### Design

Thirty dissimilarity matrices were derived from all combinations of the four probabilistic models, Jaccard distance being included twice with different disc radius distributions, and the six levels of intensity. These matrices were then used as input for the scaling methods which were thus compared on the same data. Only two-dimensional configurations were used. Classical scaling, ordinal scaling (based on classical and random starts) and least squares scaling (with weights all one, and with weights  $1/\delta_{ij}$ ) were all used with these matrices. For the Jaccard distance matrices the two preprocessing transformations we have defined were also applied before using classical scaling. The two disc radius distributions were exponential with mean 0.2, and constant radius 0.2. Several other simulations were attempted in order to test the consistency of the results presented here, but these were not systematic in nature and are not presented. They were in general agreement.

In summary we may express the three main aims of this comparison as follows:-

- (i) To determine the relative accuracy of the methods.
- (ii) To examine their behaviour for matrices with differing degrees of linearity.
- (iii) To examine their behaviour for matrices with differing degrees of error.

Additional factors that emerge are:-

- (iv) The influence of the weights in least squares scaling.
- (v) The influence of the assumed distribution in the preprocessing transformation.
- (vi) The relative merits of random and classical starting configurations.

### Results

The results are summarised in Table 3.3.1 which provides the procrustes statistic from each of the 144 combinations of model, scaling method and error level.

To answer (ii) above we look at the results from the point of view of the probabilistic models.

Binomial Hyperplane. Although the configuration generated by classical scaling is always least accurate, the differences are not great. Ordinal scaling and least squares scaling with unequal weights provide the best solutions, the latter for the lower levels of error. The eigenvalue spectrum for the binomial hyperplane model is clearly two-dimensional, so that it is not surprising that the methods tend to work equally well.

Independent Binomial. Again the classical scaling configuration is always the worst, and this time the differences are quite marked. The eigenvalue spectrum for the independent binomial model reveals more loading on the later eigenvalues, which contain information that may be used by the other methods. One of the least squares scaling formulations is always the best, indicating that these methods work particularly well for such strongly euclidean matrices.

Jaccard Distance (Variable Radius). Classical scaling works badly for this non-linear dissimilarity/distance model and

TABLE 3.3.1 Comparison of Scaling Methods. Procrustes Statistics Arising from Applying Scaling Methods to the Same Dissimilarity Matrix

Model	Method	No. of Hyperplanes					
		20	50	100	200	500	1000
Binomial Hyperplane	Classical	0.1574	0.0501	0.0170	0.0077	0.0039	0.0021
	Ordinal	0.1396	0.0434	0.0148	0.0063	0.0032	0.0019
	Least Squares (1)	0.1446	0.0442	0.0148	0.0062	0.0032	0.0019
	Least Squares (1/6 <sub>ij</sub> )	0.1465	0.0436	0.0145	0.0059	0.0031	0.0019
Independent Binomial	Classical	0.0486	0.0177	0.0088	0.0042	0.0017	0.0008
	Ordinal	0.0191	0.0076	0.0038	0.0018	0.0007	0.0004
	Least Squares (1)	0.0172	0.0065	0.0032	0.0016	0.0006	0.0003
	Least Squares (1/6 <sub>ij</sub> )	0.0304	0.0077	0.0035	0.0017	0.0006	0.0003
Jaccard Distance (expl. rad. distn. mean 0.2)	No. of Discs						
		20	50	100	200	500	1000
	Classical	0.6601	0.3700	0.2166	0.1427	0.1002	0.0877
	Ordinal	0.6682	0.3804	0.1120	0.0479	0.0115	0.0081
	Least Squares (1)	0.8995	0.2405	0.0588	0.0296	0.0180	0.0169
	Least Squares (1/6 <sub>ij</sub> )	0.8970	0.2740	0.0585	0.0279	0.0157	0.0148
Jaccard Distance (constant rad. 0.2)	Preprocessing (Normal)	0.6597	0.3127	0.1727	0.0951	0.0394	0.0292
	Preprocessing (Uniform)	0.6597	0.3124	0.1665	0.0711	0.0216	0.0147
	Classical	0.8998	0.8664	0.7603	0.7775	0.7668	0.7463
	Ordinal	0.8999	0.8298	0.7461	0.7359	0.6921	0.5922
Wilkinson Metric	Least Squares (1)	0.9200	0.8595	0.9033	0.8433	0.9525	0.9850
	Least Squares (1/6 <sub>ij</sub> )	0.9162	0.9007	0.8772	0.8291	0.9753	0.9606
	Preprocessing (Normal)	0.9000	0.7975	0.7355	0.7172	0.7118	0.6717
	Preprocessing (Uniform)	0.9000	0.7972	0.7354	0.7170	0.7117	0.6713
√(50 + 1/4 extra points)							
	7.1	15.8	21.2	25.5	29.2	30.8	
Wilkinson Metric	Classical	0.0579	0.0222	0.0095	0.0101	0.0111	0.0110
	Ordinal	0.0506	0.0175	0.0076	0.0074	0.0095	0.0100
	Least Squares (1)	0.0493	0.0160	0.0076	0.0076	0.0093	0.0097
	Least Squares (1/6 <sub>ij</sub> )	0.0458	0.0155	0.0070	0.0069	0.0089	0.0091

produces the highest procrustes statistics. It can be improved by the preprocessing transformations. Similarly the least squares methods are inferior to ordinal scaling when the error levels are low. For high levels of error none of the methods can cope adequately, and in these circumstances the least squares results are some of the best.

Jaccard Distance (Constant Radius). In this formulation all dissimilarities corresponding to points at distance greater than 0.4 are equal and 1.0. The least squares methods are quite unable to cope with this, and consistently produce bad reconstructions. Classical scaling is slightly better and can be improved by preprocessing. Ordinal scaling never produces a configuration that is markedly inferior to the other models', and becomes the best when less error is present.

Wilkinson Metric. Similar observations may be made as to those from the binomial hyperplane model. Classical scaling reconstructs the configuration slightly less satisfactorily, while, amongst the other methods, weighted least squares scaling is particularly successful.

To answer (i) we look at the results from the point of view of the scaling methods.

Classical Scaling. The success of classical scaling is highly dependent upon the euclideanness of the dissimilarity matrix. It compares favourably with the ordinal method for euclidean matrices, particularly where the eigenvalue spectrum shows clear indications of the true dimensionality.

Least Squares Scaling. Again this method is more suitable for euclidean matrices, for which it is superior to classical scaling and often to ordinal scaling. It is less successful in dealing with non-euclidean matrices.

Ordinal Scaling. This method hardly ever produces a solution that is badly inferior to that of another method, and so it seems uniformly trustworthy.

Preprocessing. The technique of preprocessing nearly always improves the procrustes statistic from classical scaling.

Other Important Considerations. The weighted form of least squares scaling is more accurate in twenty-two of the thirty possible comparisons, and is therefore recommended. This finding conforms to the maximum likelihood theory underlying this choice of weights. In the same vein, the preprocessing transformation based upon the (correct) uniform assumption is superior in ten of the twelve possible comparisons. This suggests that the technique may be quite sensitive to the details of the underlying structure. Finally it is important to record that nearly one half of the random starting configurations that were used failed to reach the same minimum of the stress function as attained by the classical scaling starting configuration, and thus often resulted in inflated values of the procrustes statistic. We have no guarantee that we always reach the global minimum, but repeated random starts and other arrivals at the same minimum suggest that this problem is not too serious, when a classical starting configuration is used.



### 3.4 Simulations on Scaling Subsets of Similarities

In this section we refer to simulation tests on the feasibility of using a small part of a similarity matrix. One complete matrix is used, and this is derived from the Jaccard distance model of Section 3.1 with parameters, the number of discs as 1000 and the radius distribution as exponential with mean 0.2. The underlying true configuration is two-dimensional with fifty points in the unit disc. Jaccard distances provide the sternest test of the ability of ordinal scaling to reconstruct a configuration.

Two different approaches are used. In the first, each point is taken in turn, and  $K$  other points are randomly (independently from a uniform distribution) selected so that the similarity with them is regarded as known. Pairs that are not selected in this way are taken to have unknown similarity value. For any pair, the probability of the value being defined is thus: 
$$\frac{2(K)}{(49)} - \frac{(K)^2}{(49)^2} .$$

In the second approach each point is again taken in turn, but this time the  $K$  smallest dissimilarity values with other points are treated as known. In this case the acceptance of a value depends upon its size, and we may expect that less values will be known for the equivalent  $K$ .

Ordinal scaling is applied to the reduced matrix in the following way. First of all, because we want to measure the fineness of detail that is contained within the matrix, we do not want to become entrapped in any locally minimum solutions that would distort the results. A particular device that we use to reduce the possibility of such an event is to start the iterative method from the true configuration. In comparison studies we were genuinely interested in the ability of the method to avoid local minima; here this is

not the immediate problem. Secondly, as the values are selected radially from each point it seems appropriate to use local order scaling. The other merit that this has in this context is that it reduces the number of comparisons that need to be made to those with a common endpoint, and thus streamlines the computation even further. All other parameters are chosen with the conventional values used in this thesis.

The measure of departure from the true configuration is made by a procrustes statistic, as usual. This is produced for nine different values of  $K$  for each of the similarity selection approaches. In addition results from the complete matrix are available. Summaries of the success of the approaches are presented for the random selections in Table 3.4.1 and Fig. 3.4.2, and for the small values in Table 3.4.3 and Fig. 3.4.4.

As the minimum number of similarities in each row increases (and hence so does the density) the final stress value also increases for both cases. This behaviour is similar to that by which the final stress tends to increase with more points, as widely observed in the papers reported in Chapter 2. It is not surprising that the initial stress is higher for the small value selection procedure because the values here are more tightly packed, and hence more difficult to order correctly. However it is the behaviour of the procrustes statistic that is of particular interest. For the random selection procedure the procrustes statistic decreases with more values, whereas for the small value selection procedure it increases. When  $K$  is 49 both methods are trivial and identical so that the final values of the procrustes statistic must be equal.

How may these results be interpreted? An immediate reaction might be that because the true configuration is used to start the

TABLE 3.4.1

Results from Using Random Entries of the Matrix

<u>K</u>	<u>S<sub>1</sub>(%)</u>	<u>S<sub>2</sub>(%)</u>	<u>R(%)</u>	<u>V</u>	<u>P</u>
5	2.934	0.230	19.5	.991	0.01784
10	3.070	1.212	36.9	.988	0.01172
15	3.394	1.901	52.1	.985	0.00785
20	3.544	2.142	64.2	.984	0.00711
25	3.723	2.329	75.7	.978	0.00653
30	3.813	2.425	84.2	.977	0.00632
35	3.876	2.482	91.3	.976	0.00594
40	3.923	2.500	96.7	.976	0.00598
45	3.917	2.493	99.2	.976	0.00596
49	3.935	2.505	100.0	.976	0.00600

K = Minimum number of entries used in each row.

S<sub>1</sub> = Original stress value.

S<sub>2</sub> = Final stress value.

R = Density, which is the number of cells used expressed as a percentage of those available.

V = Proportion of different values amongst the cells used.

P = Procrustes statistic between the final configuration and the true configuration.

Fig 3.4.2 : Random Neighbours

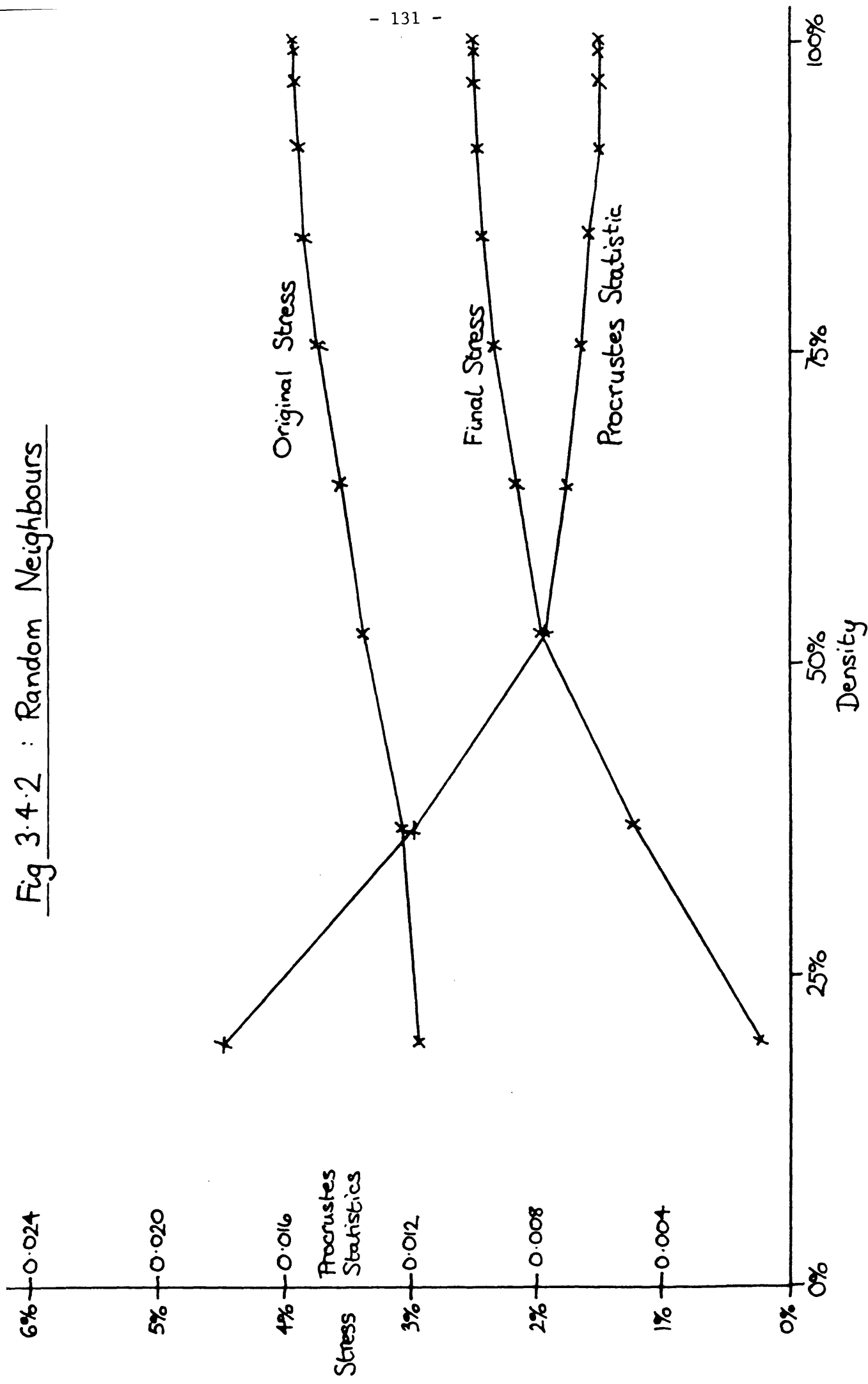


TABLE 3.4.3

Results from Using the Smallest Entries of the Matrix

<u>K</u>	<u>S<sub>1</sub> (%)</u>	<u>S<sub>2</sub> (%)</u>	<u>R (%)</u>	<u>V</u>	<u>P</u>
5	4.283	0.000	12.6	1.000	0.00111
10	3.835	0.371	24.1	.996	0.00259
15	3.656	0.736	35.9	.994	0.00313
20	3.628	1.083	47.3	.990	0.00434
25	3.949	1.425	60.7	.988	0.00478
30	3.934	1.645	72.6	.981	0.00528
35	3.930	1.911	82.6	.979	0.00553
40	3.822	2.165	90.4	.977	0.00514
45	3.854	2.370	98.0	.977	0.00567
49	3.935	2.505	100.0	.976	0.00600

K = Minimum number of entries used in each row.

S<sub>1</sub> = Original stress value.

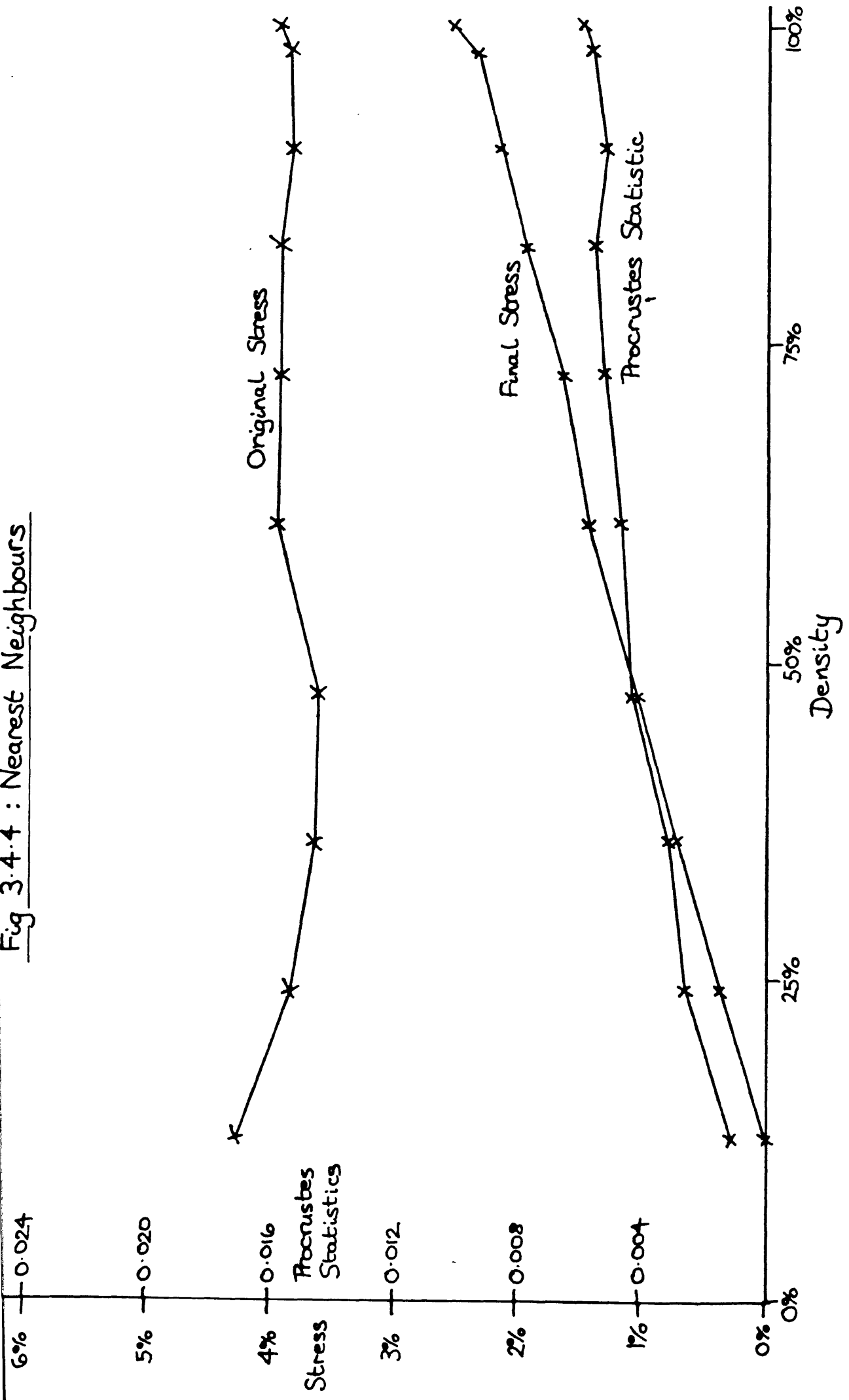
S<sub>2</sub> = Final stress value.

R = Density, which is the number of cells used, expressed as a percentage of those available.

V = Proportion of different values amongst the cells used.

P = Procrustes statistic between the final configuration and the true configuration.

Fig 3.4.4 : Nearest Neighbours



process, the less constraints that are applied, the less the adjustment that will be required. But this will not do for, although the procrustes statistic increases for the small selection procedure, the opposite is true for the random process. Thus there must be something specific to the small values that is causing this behaviour. If we believe the results of Graef and Spence (1979), which seem reasonable, then we are forced to the conclusion that it is the large dissimilarities that determine the coarse structure of the configuration, and that they must be known in order to provide a good starting configuration. However these results would suggest that it is important to know the small values in order to deduce the fine structure.

How may these results be applied? In some circumstances there will be a natural constraint upon the dissimilarity matrix values that may be obtained. For example, in Chapter 4 we argue that it is not meaningful to derive a dissimilarity value for a pair of M.P.s if, as a pair, they fail to vote in a sufficiently large number of divisions. In these cases it is likely that we will be in the random sampling situation, although it may be that only within-cluster type measurements may be made in which case the values will be small. More often the whole range of values will be at our disposal. If the matrix is of reasonable size (less than eighty by eighty, say) then we would have no hesitation in using all available values. Beyond this range, restriction to a subset becomes attractive and, if it is possible to generate a reasonable starting configuration, the use of the small values would seem to be justified.

This set of results demonstrates that there is great potential in the idea of using just a portion of the dissimilarity matrix. It would be a worthwhile extension to this exercise to

undertake a more complete study of the properties of reduced matrices which would involve:-

- (i) Different starting configuration approaches.
- (ii) Different degrees of euclideaness of matrix.
- (iii) Different amounts of error.
- (iv) Global and local order scaling.
- (v) Other subsets also defined by dissimilarity value.  
(e.g. What would happen if we used the largest values?)
- (vi) Different configurations. (Different numbers of points would allow statements concerning possible economies.)

Such a study would add confidence to what has been proposed, but even these results provide an adequate basis for the neglect of certain values if their computation is regarded as dubious.



3.5 Procrustes Statistics Arising from Slightly Different Configurations

In (1.13) we referred to the results of Sibson (1979) concerning procrustes statistics for two slightly different configurations. In particular we noted that if  $X$  is a centred, full rank,  $K \times N$  configuration matrix which is perturbed to  $Y = X + \epsilon Z + O(\epsilon^2)$ , where  $Z$  is another  $K \times N$  matrix, then both  $G_E(X,Y)$  and  $G_S(X,Y)$  can be represented as quadratics in the elements of  $Z$ , the precise forms of which are given in (1.13.1) and (1.13.2).

Sibson then illuminates these results by considering the case in which the entries in  $Z$  are independent  $N(0,1)$  random variables. It then follows ( (1.13.3) and (1.13.4) ) that

$$G_E(X,Y) \sim \epsilon^2 \chi_f^2 + O(\epsilon^3) \text{ where } f = NK - \frac{1}{2}K(K + 1)$$

$$G_S(X,Y) \sim \epsilon^2 \chi_g^2 + O(\epsilon^3) \text{ where } g = f - 1$$

These last two results are independent of the matrix  $X$ . They can be justified loosely by noticing that the original procrustes statistic, before translation, orthogonal transformation or scale change, is the sum of  $NK$  squares of independent  $N(0,1)$  random variables, that is a  $\chi_{NK}^2$  random variable, and subsequently  $K$  degrees of freedom are lost through translation, approximately  $\frac{1}{2}K(K - 1)$  through orthogonal transformation, and finally 1 through scale change.

We now examine the range of validity of the above approximations (1.13.3) and (1.13.4). To do this we use a configura-

tion of 50 points independently and uniformly distributed over the unit disc  $\{(x,y) : x^2 + y^2 \leq 1\}$  in two dimensions, and then centre it at the origin by translation. Only one such configuration was generated, since the results are known to be independent of the base configuration. Initial values of  $\epsilon$  were 0.02, 0.05, 0.1 and 0.2. For each value of  $\epsilon$  ten independent realisations of  $Z$  were produced from a pseudo random number generator, and the values of  $G$  (the original procrustes statistic),  $G_T$  (the procrustes statistic after translation),  $G_E$  and  $G_S$  were obtained. If a random variable  $V$  has distribution  $(\epsilon^2 \chi_h^2)$ , where  $h$  is sufficiently large (greater than 50 in practice) then

$\frac{\sqrt{2V}}{\epsilon^2}$  is approximately  $N(\sqrt{(2h-1)}, 1)$  in distribution.

In Table 3.5.1. we record  $G$ ,  $\sqrt{\frac{2G}{\epsilon^2}}$ ,  $G_T$ ,  $\sqrt{\frac{2G_T}{\epsilon^2}}$ ,  $G_E$ ,  $\sqrt{\frac{2G_E}{\epsilon^2}}$ ,  $G_S$  and  $\sqrt{\frac{2G_S}{\epsilon^2}}$  for each level of  $\epsilon$  and each replication.

In Table 3.5.2 we present the mean values of the transformed statistics taken over the ten replications, and provide 95% confidence limits for these means based upon the normal approximation. Of course the values of  $G$ ,  $G_T$ ,  $G_E$  and  $G_S$  are highly interdependent, and this must be remembered in interpreting the mean values.

The results show that the approximations to  $G$ ,  $G_T$  and  $G_E$  are entirely satisfactory for this range of values of  $\epsilon$ . That this is true for  $G$  and  $G_T$  is clear from the above intuitive arguments. That the result is true for  $G_E$  is of more interest, and is caused by the lack of any systematic rotation needed to match the original and perturbed configurations. We proceed by extending the range of  $\epsilon$  to

TABLE 3.5.1

	G	$\frac{\sqrt{2G}}{\epsilon^2}$	$G_T$	$\frac{\sqrt{2G_T}}{\epsilon^2}$	$G_E$	$\frac{\sqrt{2G_E}}{\epsilon^2}$	$G_S$	$\frac{\sqrt{2G_S}}{\epsilon^2}$
$\epsilon=0.02$	0.0440	14.83	0.0437	14.78	0.0435	14.75	0.0421	14.51
	0.0414	14.39	0.0374	13.67	0.0374	13.67	0.0374	13.67
	0.0408	14.28	0.0407	14.27	0.0400	14.14	0.0390	13.96
	0.0383	13.84	0.0381	13.80	0.0381	13.80	0.0380	13.78
	0.0397	14.09	0.0395	14.05	0.0395	14.05	0.0390	13.96
	0.0411	14.34	0.0409	14.30	0.0403	14.20	0.0403	14.20
	0.0436	14.76	0.0405	14.23	0.0404	14.21	0.0401	14.16
	0.0473	15.37	0.0471	15.34	0.0469	15.31	0.0466	15.26
	0.0432	14.69	0.0431	14.67	0.0429	14.65	0.0429	14.65
	0.0353	13.28	0.0344	13.11	0.0343	13.10	0.0341	13.06
$\epsilon=0.05$	0.1878	12.25	0.1834	12.11	0.1834	12.11	0.1757	11.85
	0.2206	13.28	0.2143	13.09	0.2089	12.93	0.2051	12.80
	0.3016	15.53	0.3014	15.52	0.3006	15.51	0.2912	15.26
	0.1948	12.48	0.1946	12.47	0.1928	12.42	0.1868	12.22
	0.2354	13.72	0.2211	13.29	0.2157	13.14	0.2058	12.83
	0.2584	14.37	0.2508	14.16	0.2498	14.14	0.2498	14.13
	0.2286	13.52	0.2284	13.51	0.2187	13.23	0.2171	13.17
	0.2130	13.05	0.2003	12.65	0.1962	12.52	0.1956	12.50
	0.2768	14.88	0.2761	14.86	0.2715	14.74	0.2647	14.55
	0.2741	14.80	0.2740	14.80	0.2728	14.77	0.2717	14.74
$\epsilon=0.1$	0.8886	13.33	0.8659	13.15	0.8553	13.08	0.8544	13.07
	0.9509	13.79	0.9144	13.52	0.9142	13.52	0.8266	12.85
	1.1382	15.08	1.1337	15.05	1.1337	15.05	1.0119	14.22
	0.7566	12.30	0.7541	12.28	0.7493	12.24	0.7350	12.12
	1.0159	14.25	1.0069	14.19	1.0034	14.17	0.9159	13.53
	1.0468	14.46	1.0465	14.46	1.0393	14.42	1.0129	14.23
	1.0208	14.28	1.0142	14.24	1.0058	14.18	0.8923	13.35
	1.2159	15.59	1.2043	15.51	1.2019	15.50	1.1172	14.94
	0.7719	12.42	0.7717	12.42	0.7433	12.19	0.7038	11.86
	0.7186	11.98	0.7115	11.92	0.6605	11.49	0.6197	11.13
$\epsilon=0.2$	4.0675	14.26	3.9917	14.12	3.9452	14.04	3.7020	13.60
	3.8395	13.85	3.7548	13.70	3.7140	13.63	3.1253	12.50
	4.2028	14.49	4.1182	14.34	4.1136	14.34	2.9417	12.12
	3.6000	13.41	3.5821	13.38	3.5820	13.38	3.2647	12.77
	3.5461	13.31	3.4372	13.10	3.4368	13.10	2.9659	12.17
	4.9116	15.67	4.7585	15.42	4.7581	15.42	3.8728	13.91
	3.8518	13.87	3.7386	13.67	3.7376	13.67	3.2819	12.81
	4.5006	15.00	4.2776	14.62	4.2414	14.56	3.6223	13.45
	3.9690	14.08	3.6863	13.57	3.6591	13.53	2.8684	11.97
	4.1496	14.40	4.0287	14.19	4.0283	14.19	3.3099	12.86

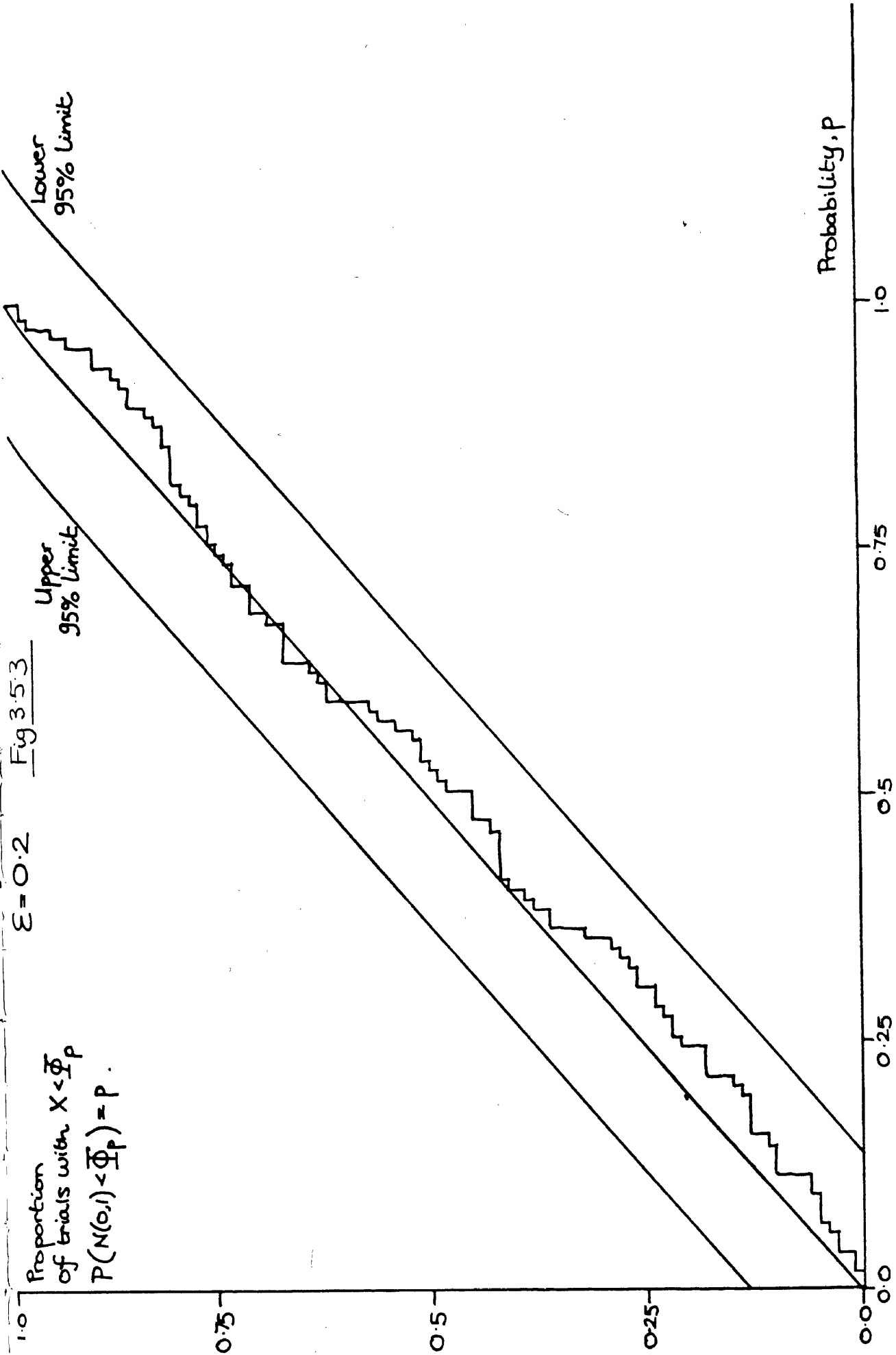
TABLE 3.5.2

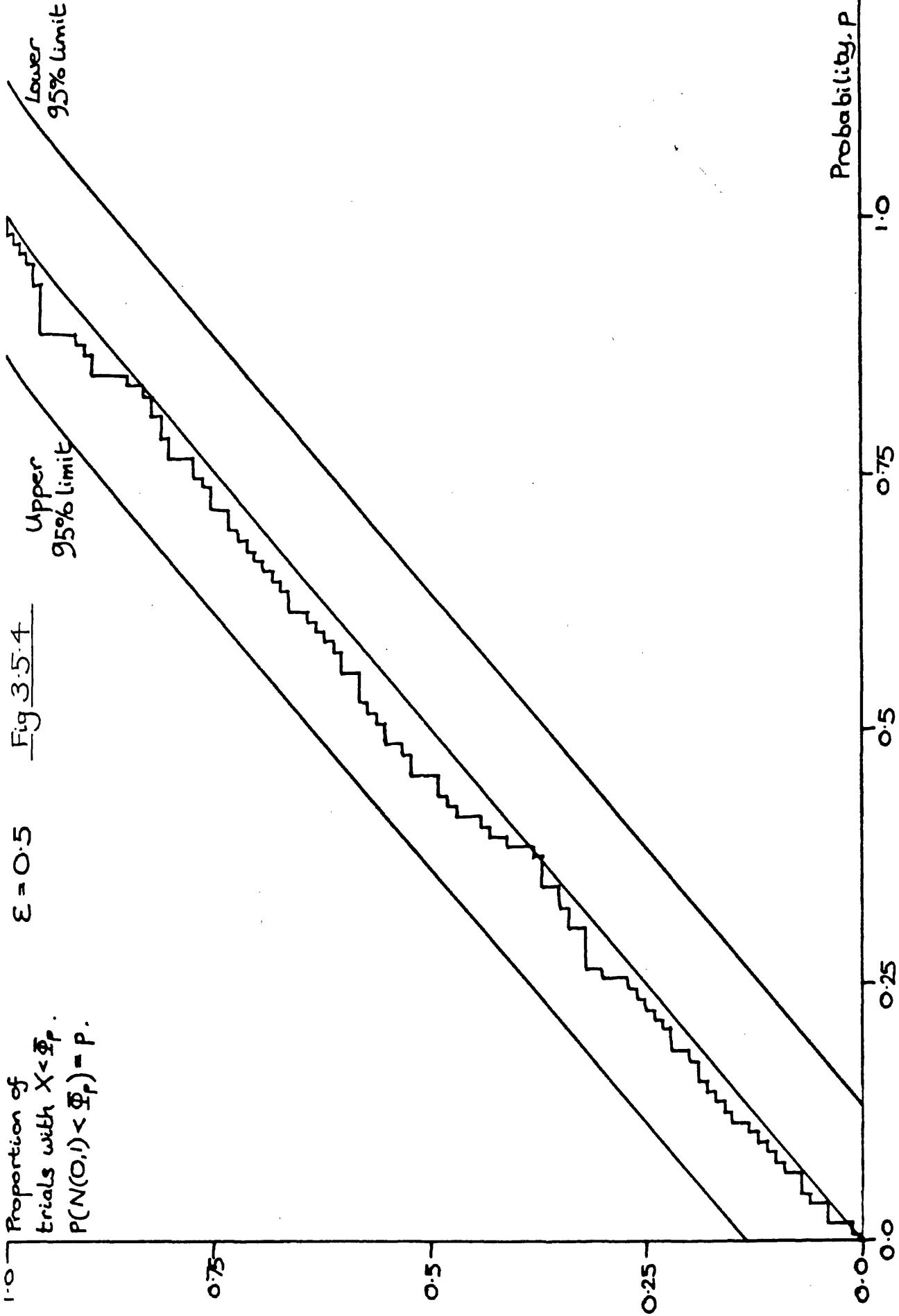
Lower Confidence Limit	13.49	13.34	13.27	13.20
Mean Value of	$\frac{2G}{\sqrt{\frac{\epsilon^2}{2}}}$	$\frac{2G_T}{\sqrt{\frac{\epsilon^2}{2}}}$	$\frac{2G_E}{\sqrt{\frac{\epsilon^2}{2}}}$	$\frac{2G_S}{\sqrt{\frac{\epsilon^2}{2}}}$
$\epsilon = 0.02$	14.39	14.22	14.19	14.12
$\epsilon = 0.05$	13.79	13.65	13.55	13.41
$\epsilon = 0.1$	13.75	13.67	13.58	13.13
$\epsilon = 0.2$	14.23	14.01	13.99	12.82
Upper Confidence Limit	14.73	14.58	14.51	14.44

test the robustness of the distributional approximation to  $G_E$ . But first we observe that the approximation to  $G_S$  is not nearly as satisfactory, especially for higher values of  $\epsilon$ . S. Langron (personal communication) has demonstrated that the coefficient of the  $\epsilon^3$  term in the approximation to  $G_S$  is large and, if ignored, causes  $G_S$  to be over-estimated. Certainly we shall expect the mean square object to origin distance to be greater for the perturbed configuration, and so some systematic scale change will be required.

In order to test the approximation to  $G_E$  even more severely we use the values  $\epsilon = 0.2, 0.5$  and  $1.0$  and proceed as before, this time generating 100 values of the procrustes statistic for each level of  $\epsilon$ . The results are summarised in the three graphs, Figs. 3.5.3, 3.5.4 and 3.5.5, which give the empirical distribution function for the 100 values and the distribution function for the null hypothesis that the distribution is  $\epsilon^2 \chi_f^2$ , where both functions have been transformed so that the latter follows the line  $y = x$  in  $(0,1)$ . Two lines are also marked giving 95% limits for the acceptance of the null hypothesis under the Kolmogorov-Smirnov test.

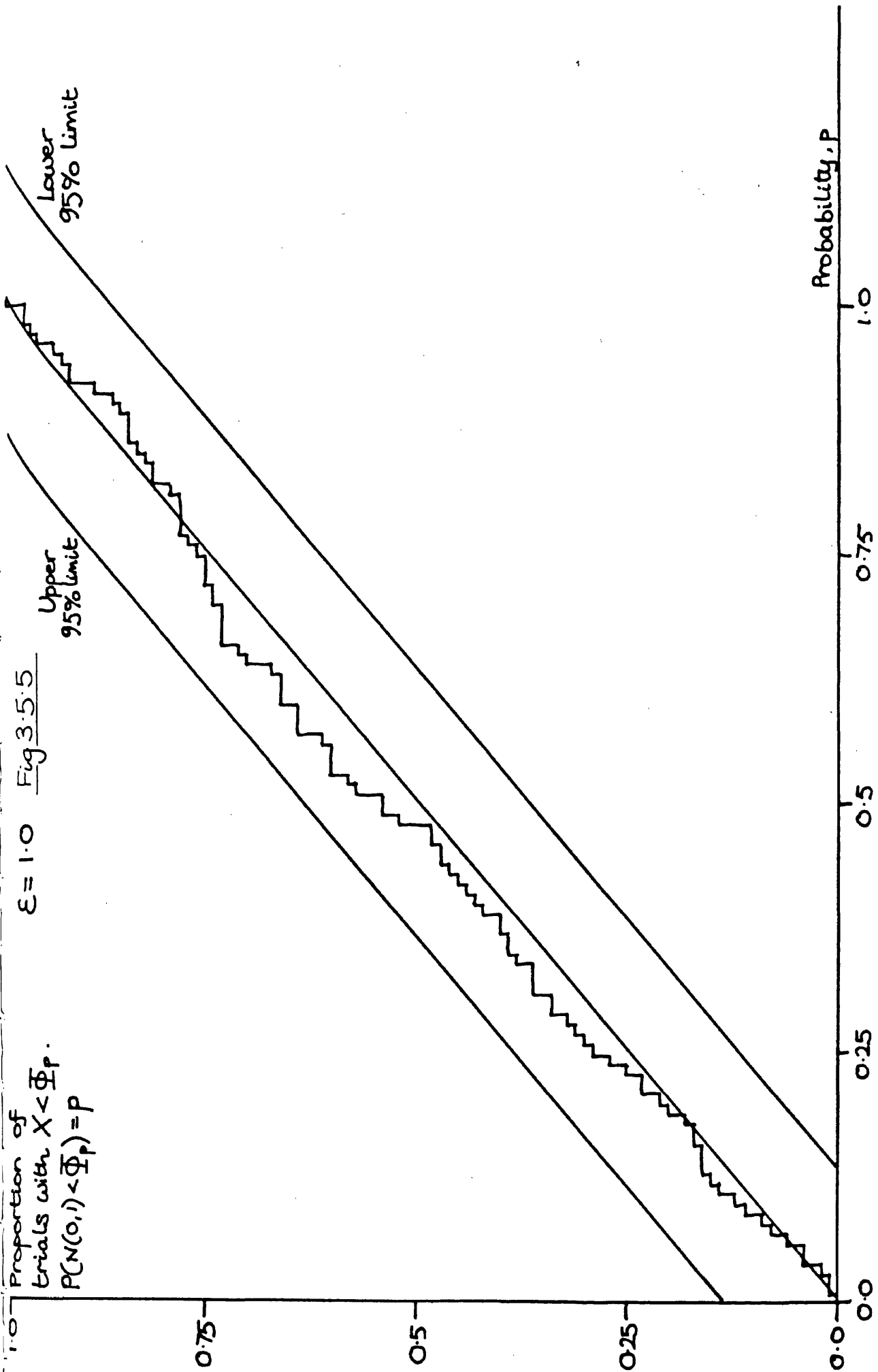
At each of these levels of  $\epsilon$  the empirical distribution function lies well within the limits and we may conclude that the approximate distribution (1.13.3) is satisfactory even with  $\epsilon$  as large as  $1.0$ . At this level of  $\epsilon$  there will be gross changes in the configuration but, as has been emphasised, there will be no systematic rotation effect. We may conclude that the approximation is robust, even with large displacements of the original configuration corresponding to  $\epsilon = 1$ .





Proportion of trials with  $X < \Phi_p$ .  
 $P(N(0,1) < \Phi_p) = p$

$\epsilon = 1.0$  Fig 3.5.5





### 3.6 Procrustes Statistics Arising from Slightly Different Squared Distance Matrices

There are two other results of Sibson (1979) that are examined here. These relate to procrustes statistics arising from two configurations produced from classical scaling when the method has been applied to two slightly different squared distance matrices. The results were mentioned in Section 1.13 as (1.13.5), (1.13.6), (1.13.7) and (1.13.8), which provide explicit expressions for  $G_E$  and  $G_S$  in terms of the elements of a symmetric matrix,  $F$ , used to perturb a parent squared distance matrix. In much the same way as in the previous section, Sibson sheds light on these results by following through the calculations for the specific case in which the entries in  $F$  are symmetric, zero on the diagonal, and independent off the diagonal with mean zero and variance one, and by calculating the expected value of  $G_E$ . We retain the notation of Section 1.13.

We turn to an investigation of the range of validity of the expressions (1.13.5) and (1.13.8). This is done in a manner very similar to that used in the previous section. One configuration,  $X$ , of 50 points lying in the two-dimensional unit disc was generated and centred. Exact squared interpoint distances were computed from this configuration to form the matrix  $E$ . Four perturbation matrices,  $H$ , were obtained using a pseudo-random number generator such that the off diagonal entries were sampled independently from the  $N(0,1)$  distribution, the diagonal was zero, and the matrix was symmetric. Four values of  $\epsilon$  were used. These were 0.1, 0.2, 0.5 and 1.0. For each value of  $\epsilon$  and each perturbation matrix we calculated  $G_{E, \frac{\epsilon^2}{2}A}$  (as in (1.13.5)), and then calculated  $E(\frac{\epsilon^2}{2}A)$

(as in (1.13.8) ). The results are given in Table 3.6.1.

The approximation to  $G_E$  is quite satisfactory, being good for the smaller values of  $\epsilon$  and less accurate for  $\epsilon = 1.0$  . Indeed for  $\epsilon = 1.0$  the errors that are introduced into the squared distance matrix  $E$  are of the same order of magnitude as the squared distances themselves. Indeed it may occasionally be the case that the perturbed matrix  $F$  will have negative entries, and devices introduced to circumvent this problem (setting them as zero in our case) will change the expectation of the approximation, causing a little extra inaccuracy in that row of Table 3.6.1. If this is done, corresponding alterations have to be made to  $H$  , so that the approximation itself is calculated accurately.

In addition we present the results of applying the approximation (1.13.6) for  $G_S$  to some specific cases. These arise from the comparative studies of Section 3.3. We treat the two- and six-dimensional versions of the binomial hyperplane and independent binomial models. Considering the six standard error levels this provides twenty-four matrices. For each matrix we compare the procrustes statistic after classical scaling with the approximation based upon the perturbation induced by the hyperplane process. The results are displayed in Table 3.6.2. We see that the approximation is quite satisfactory, and particularly good for the low levels of error.

Together these results demonstrate further how classical scaling processes the errors in the dissimilarity matrix. Allied with the simulation study findings of Section 3.2 we may agree with the conclusion of Sibson that "classical scaling is a method that is robust against errors which leave observed dissimilarities still approximately linearly related to distance".

TABLE 3.6.1

		<u><math>\epsilon=0.1</math></u>	<u><math>\epsilon=0.2</math></u>	<u><math>\epsilon=0.5</math></u>	<u><math>\epsilon=1.0</math></u>
First H	G	0.0210	0.0848	0.549	2.39
	$\frac{\epsilon^2 A}{2}$	0.0212	0.0849	0.531	2.12
Second H	G	0.0256	0.1050	0.716	3.32
	$\frac{\epsilon^2 A}{2}$	0.0250	0.1001	0.626	2.50
Third H	G	0.0150	0.0603	0.389	1.79
	$\frac{\epsilon^2 A}{2}$	0.0151	0.0603	0.377	1.51
Fourth H	G	0.0187	0.0745	0.468	1.91
	$\frac{\epsilon^2 A}{2}$	0.0204	0.0817	0.510	2.04
	E $\left(\frac{\epsilon^2 A}{2}\right)$	0.0208	0.0833	0.521	2.08

TABLE 3.6.2

Exact and Approximated Procrustes Statistics

Model	No. of Dims.	No. of Hyperplanes					
		20	50	100	200	500	1000
Binomial 2	Exact	0.15744	0.05015	0.01699	0.00768	0.00387	0.00206
	Approx.	0.14554	0.05112	0.01832	0.00864	0.00368	0.00214
Independent Binomial 2	Exact	0.34853	0.11730	0.06814	0.02998	0.01322	0.00776
	Approx.	0.51316	0.25380	0.09550	0.03607	0.01291	0.00930
Hyperplane 6	Exact	0.35751	0.15082	0.07851	0.03943	0.01552	0.00769
	Approx.	0.41518	0.16033	0.07846	0.03821	0.01522	0.00760

### 3.7 References

- GOWER, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, pp. 857-871.
- GRAEF, J. and SPENCE, I. (1979). Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, 86, pp. 60-66.
- KENDALL, M. G. and MORAN, P. A. P. (1963). *Geometrical Probability*. Griffin: London.
- MARDIA, K. V. (1970). *Families of Bivariate Distributions*. Griffin: London.
- MILES, R. E. (1970). On the homogeneous planar Poisson process. *Mathematical Biosciences*, 6, pp. 85-127.
- SIBSON, R. (1979) . Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41, pp. 217-229.
- SIBSON, R. (1980) . The Dirichlet tessellation as an aid in data analysis. *Scandinavian Journal of Statistics*, 7, pp. 14-20.
- SIBSON, R., BOWYER, A. and OSMOND, C. (1981). Studies in the robustness of multidimensional scaling: euclidean models and simulation studies. *Journal of Statistical Computation and Simulation*

C H A P T E R   F O U R

AN APPLICATION IN BRITISH POLITICAL HISTORY

	<u>PAGE</u>
4.1    Introduction: Motivation for the Project    ..    ..	150
4.2    The Extent of Earlier Statistical Analysis of Voting    ..    ..    ..    ..    ..    ..    ..    ..	154
4.3    Acquiring and Assembling the Data    ..    ..    ..	157
4.4    A Preliminary Approach ..    ..    ..    ..    ..    ..	168
4.5    The Measurement of Dissimilarity    ..    ..    ..	176
4.6    Definitions of Groups Used in the Analyses    ..	178
4.7    Results Obtained by Single-Link Clustering    ..	180
4.8    Results Obtained by Ordinal Scaling    ..    ..    ..	183
4.9    Results Obtained by Least Squares Scaling    ..    ..	198
4.10   Economies in Use of Similarities    ..    ..    ..	201
4.11   Conclusions    ..    ..    ..    ..    ..    ..    ..    ..	204
4.12   References    ..    ..    ..    ..    ..    ..    ..    ..	206

#### 4.1 Introduction: Motivation for the Project

Little is known about the consistency of parliamentary voting in the last 150 years. It has been generally assumed that M.P.s have only rarely voted against their party whips' advice since the emergence of the nationally organised political parties in the late nineteenth century. This project was conceived as an application of multidimensional scaling with the aim of analysing all House of Commons' divisions in one parliamentary session (1861) so that as full and as unbiased a picture of voting behaviour as possible could be discerned. If this proved informative, extensions to more than one session were envisaged. The intention was to monitor the voting behaviour of specified sets of M.P.s on specified sets of divisions by producing multidimensional scaling maps. The project arose from a fusion of the research interests of Valerie Cromwell, Reader in History at the University of Sussex, who has a particular interest in nineteenth century British political history, and Professor Robin Sibson, who has been responsible for developments in multidimensional scaling. It was regarded as exploratory in nature, treating just one parliamentary year as it did, and was supported by a research grant from the Social Science Research Council. It has been very much a collaborative effort with Valerie Cromwell, who is acknowledged with gratitude. The form of parts of this chapter has evolved from the end-of-grant report that she has submitted to the Social Science Research Council (Cromwell, 1980).

As far as we know, multidimensional scaling has not been applied to the analysis of voting records. This attempt was designed to explore the usefulness and adaptability of the method when applied to Commons' division lists. If successful the strength or weakness of

party and other group loyalties would be exposed, and assumptions derived from other sources such as political diaries and correspondence, press reports and comment could be tested.

The choice of session was important if assessment of the usefulness of the method was to be possible. A particular attraction of the 1860's as a period for detailed attention of this kind was the apparent fragility of the political structure. Party loyalties in the early 1860's are generally considered to have been weak and fluctuating. Palmerston's government had come into office in the summer of 1859 with a slim majority, which had been calculated as being at most 16 by Mowbray (1900). During the early years of the ministry contemporary commentators repeatedly remarked on the difficulties created for the government by radical dissidents on their side of the Commons. There was some evidence that the conservative opposition leadership more than once indicated to liberal ministers a general reluctance to turn the government out of office as long as it pursued a moderate financial policy (Monypenny and Buckle, 1916 ). Although whig and radical support had assisted Palmerston to take office, it appeared that it was whig and conservative support which enabled the government to pass such legislation and parliamentary business as it dared to introduce in its lifetime. It was to test these assumptions that our analysis of voting in divisions was designed.

Within this period, the choice of which parliamentary session to use was obviously important. 1861 was chosen as a year close to the beginning of a ministry, but which did not see a change of government. The political complexion of the Commons near the beginning of a ministry would have reflected closely the political sympathies of the electorate. A change of government in the middle of a parliamentary session would have presented additional



problems of data management which would have been unnecessary in such a feasibility study. Another advantage of 1861 was that no major divisive domestic or foreign issue polarised political opinion in that year. It was its very 'ordinariness' which made it a suitable choice.

The advantage of using votes by members was that it provided positive evidence, well recorded. It would have been even more convenient if all evidence of pairing and abstention were available. However when compared with the random, fragmentary evidence provided by diary, press, comment and correspondence it afforded a very hard, complete set of data. On a certain date, a member was prepared to walk into one of two Division Lobbies in support of a particular opinion and to have that vote recorded and published. While a vote might only have had procedural significance, as for instance when it ensured full debate of an issue, it was still an indication of an opinion of a very positive sort.

One problem with some divisions was their low participation, and this could have been seen to present a difficulty, but such smallness did not necessarily indicate lack of importance. Time of evening or stage of the session often affected voting in all but the most politically important divisions. The House was always thin at the dinner hour and towards the end of the session attendance at the Commons was poor. On the other hand we had size problems of a different nature to negotiate. The number of members with seats in the Commons was 662; the number of divisions in the session was 187. These numbers were typical historically, but were much larger than commonly used in scaling applications. Additionally the M.P.s had highly variable participation rates ranging from a high of 182 to a low of 0 votes. Effectively this produced a group of M.P.s about whom nothing could

be inferred and forced there to be varying degrees of reliability attached to findings on other M.P.s.

An accurate comparison between any pair of M.P.s required that there should have been a reasonably large number of divisions in which they both voted. Comparisons of rarely voting M.P.s on small sets of divisions were unlikely to be reliable. Thus we were unable to assess some quite famous M.P.s because they voted so infrequently in 1861, and little could be done to examine attitudes to Scotland, Civil Service reform or India, for example, because those issues were rarely debated in that year. Under less extreme conditions we were able to produce maps and look for changes in position of individual M.P.s for different sets of divisions.

## 4.2 The Extent of Earlier Statistical Analysis of Voting

The advantages to be gained from a statistical analysis of voting have been appreciated for some time. The first numerical analyses of division lists were quite straightforward in concept and bore much resemblance to the initial data analyses that we report in Section 4.4. However physical developments in computing power and statistical developments in multidimensional scaling have considerably extended the range of possible analyses, as we demonstrate. In particular it has become possible to obtain an objective assessment of each individual's patterns of voting with respect to the rest of his colleagues.

Returning to traditional analyses, an index of party cohesion for each division was derived as:-

$$\frac{|Y - N|}{Y + N}$$

where Y represented the number of 'Aye' votes in the particular party in that division and N represented the corresponding number of 'No' votes. Introducing A as the number of abstainers, an index of abstention was formed as:-

$$\frac{A}{Y + N + A}$$

To compare two parties the index of party likeness was defined as:-

$$1 - \left| \frac{Y_1}{Y_1 + N_1} - \frac{Y_2}{Y_2 + N_2} \right|$$

where the suffices represented the two parties being compared.

These indices were used alongside measures such as:-

- (i) The percentage of votes for and against for each party in any division.
- (ii) The percentage of divisions unanimous for each party.
- (iii) The percentage of votes against the party majority in any division.

(iv) An individual's percentage of successes in passing legislation., in order to provide useful summaries. No study attempted to consider numbers as large as those found in this present analysis. Principal works in this area have been those of Wahlke and Eulau (1959), Aydelotte (1963, 1966, 1972, 1977), Anderson, Watts and Wilcox (1966), Berrington (1968), Loveday (1975), Loveday, Martin and Parker (1977) and Beringer (1978), who provides a useful summary.

A popular technique with early authors was Guttman scale analysis. A scale comprised two rank orderings, one of a set of divisions and one of a set of voters. The idea was that if the ordering of the divisions was carefully chosen it would represent an 'axis of attitude' such that individual voters would vote 'Aye' up to a certain point and 'No' thereafter. If such an ordering of the divisions could be found the voters could then be ranked according to the point at which their response altered. A successful scale would appear as follows:-

		<u>Division</u>						
		1	2	3	4	5	6	7
	1	Aye	Aye	Aye	Aye	Aye	Aye	Aye
	2	Aye	Aye	Aye	Aye	Aye	Aye	Aye
	3	No	Aye	Aye	Aye	Aye	Aye	Aye
	4	No	No	No	Aye	Aye	Aye	Aye
	5	No	No	No	Aye	Aye	Aye	Aye
	6	No	No	No	Aye	Aye	Aye	Aye
<u>Voter</u>	7	No	No	No	Aye	Aye	Aye	Aye
	8	No	No	No	No	Aye	Aye	Aye
	9	No	No	No	No	Aye	Aye	Aye
	10	No	No	No	No	No	Aye	Aye
	11	No	No	No	No	No	Aye	Aye
	12	No	No	No	No	No	No	Aye
	13	No	No	No	No	No	No	No

Various freedoms were allowed when such perfect fits were not quite obtainable. It was hoped that both rank orderings would then contain useful information. The technique is mentioned in the works of Aydelotte and Beringer given above.

More recently Heyck and Klecka (1973) and Heyck (1974) have used techniques of discriminant analysis to classify radical M.P.s based upon the voting behaviour of known radicals in divisions that were specially chosen because of their known importance to the radical cause. Veitch and Jaensch (1974) have tailored principal component analysis and factor analysis to fulfil the special requirements of voting data. These ideas have also been used by Loveday, Martin and Parker (1977). Hartigan (1972, 1974, 1975) has successfully applied his direct clustering algorithms to the voting of countries in the United Nations. In his work the original data matrix containing objects (voters) by variables (divisions) had rows and columns permuted simultaneously in order to highlight rectangular blocks of consistent behaviour. A much less powerful version of the same type of method was used by Hatzenbuehler (1972) who formed a similarity matrix of percentage agreements between voters and sought squares of entries greater than a specified baseline, along the leading diagonal.

Some of these ideas are incorporated in the preliminary statistical analyses reported in Section 4.4. However we have been able to go much further by using scaling techniques. The range of interpersonal agreements and conflicts that can be studied is much greater, as is the sensitivity of the final results. The final results are appealing in their simplicity and interpretability. Many of the other techniques are limited in that they produce what is effectively a one-dimensional solution. This equally applies to single-link clustering (Section 4.7). Thus multidimensional scaling is intrinsically more powerful.

4.3 Acquiring and Assembling the Data

The main sources on which the project was based were the Commons' Division Lists as printed for the House. A typical list has the following format:-

"Mercurii, 13<sup>o</sup> die Martii, 1861.

Numb. 19.

County Franchise Bill, - Order for Second Reading read; Motion made, and Question proposed, "That the Bill be now read a second time:" - Whereupon Previous Question put, "That that Question be now put:" - (Mr. Augustus Smith:) - The House divided; Ayes 220, Noes 248.

A Y E S.

Acton, Sir John Dalberg

Adair, Hugh Edward

. . .

. . .

Wyld, James

Wyvill, Marmaduke

Tellers for the Ayes, Mr. Locke King and Mr. Hastings Russell.

N O E S.

Adderley, Rt. Hn. Charles Bowyer

Arbuthnott, Hon. General

. . .

. . .

Wynne, Wm. W. E. (Merioneth)

Yorke, Hon. Eliot Thomas

Tellers for the Noes, Mr. Augustus Smith and Mr. Du Cane."

Subsequent pages contained corrections and these we incorporated.

Some limited biographical information about each individual was also obtained. This nearly always came from Dod's "Parliamentary Companion" for 1861 and 1862 and the "Dictionary of National Biography". Nominal party allegiance was obtained from Dod. The advertising slip for the 1879 edition of Dod vouched for these political descriptions by asserting, "In all possible cases the exact words of the member himself has been preferred to any other indication of his political opinions".

From these two sources we derived a matrix of voting behaviour,  $M = (m_{ij})$ , where  $1 \leq i \leq 662$  corresponding to M.P.s  
 $1 \leq j \leq 187$  " " divisions

and  $m_{ij}$  was a variable that took one of the following six values:

A - M.P. i voted with the Ayes in division j

N - " " " " Noes " " "

B - " " was teller for the Ayes in division j

M - " " " " " Noes " " "

X - " " did not vote in division j

Z - " " was not able to vote in division j, as he was not then a member.

There were obvious reservations which had to be recognised in the use of division lists as an indicator of political or other allegiance, as there were with any records of voting behaviour. M.P.s may have had many reasons for not voting in a division, but we were only able to allow for the most straightforward. That is, casual or occasional absence from a division was separately coded from inability to vote for reasons such as resignation (Chiltern Hundreds etc.) or death. Where members were eligible to vote, their abstentions were uniformly coded as 'X'. This failed to allow for two common

practices, pairing and deliberate abstention. No evidence of any systematic kind existed about pairing in particular votes, or about pairing agreements. It would therefore have been misleading to introduce into the coding such random evidence of pairing as existed. Similarly, it seemed dangerous to try to introduce any qualitative criterion into our coding of abstention. Although evidence existed of decisions by certain members to abstain from specific votes, such evidence was of a random nature. We decided that we could only use positive evidence, that of the actual votes by members.

A four-letter acronym was designed for each M.P. so that the alphabetical ordering of acronyms and surnames as used on the division lists should correspond and so that the acronym should give a good clue as to the M.P.'s identity. Thus PALM was derived for Viscount Palmerston and DISR for the Rt. Hon. Benjamin Disraeli etc. This required careful collection of a complete set of M.P.s for the year, seventeen new names having appeared during the year.

The following information was also collated for each M.P.:-

- (a) His party allegiance defined by Dod.
- (b) His full title, as used on the division lists.
- (c) His constituency.
- (d) Whether the constituency was in England and Wales, Scotland or Ireland.
- (e) Whether the constituency was in a borough or a county.
- (f) The M.P.'s age.
- (g) Whether the M.P. had ever had a brother as an M.P., or any other relation.
- (h) Whether the M.P. was a past or present government office holder.
- (i) Whether the M.P. had served in the militia, regular army or navy.



Other information that was available identified East India Co. directors or proprietors, governors and directors of the Bank of England, merchant navy officers, brewers and so on, but applied only to very small numbers of M.P.s and was thus considered unworthy of being encoded.

It was the placing of the information onto a computer that presented the next challenge. Would this best be done interactively at the terminal or was an intermediate step of preparing coding sheets likely to save time and minimise error? It was decided that the interactive approach would be more cumbersome, require more training of staff and be more liable to error, so we devised a procedure based on coding sheets.

Fig. 4.3.1 shows one of the coding sheets (the A-sheet) prepared for M.P.s 26-50 in the alphabetical ordering of M.P.s. Each line eventually corresponded to a punched card of eighty characters. The first four characters on each card were the M.P.'s acronym. The last three on each card were one of 61A, 61B, 61C, 61D, 61E or 61F indicating that the year of study was 1861 and then the particular sheet chosen. This allowed for the extension of the project to other years. If that took place it would probably not still be possible to maintain the same order among acronyms and surnames while keeping the acronyms for M.P.s still serving. Otherwise the lines contained for M.P. no. i:-

- (a) A sheets:  $m_{ij}$  for  $1 \leq j \leq 56$  )
- (b) B sheets:  $m_{ij}$  for  $57 \leq j \leq 112$  ) recorded in groups of 4 in 5 cols.,
- (c) C sheets:  $m_{ij}$  for  $113 \leq j \leq 168$  ) for ease of punching.
- (d) D sheets:  $m_{ij}$  for  $169 \leq j \leq 187$  )



(e) E sheets: Columns 6 - 8: Dod's party label, coded as

LIB = liberal

LIC = liberal-conservative

CON = conservative

REF = reformer

RAD = radical

WHI = whig.

Columns 11-75: The M.P.'s full title as used most often in the division lists, which included an indication of constituency in some ambiguous cases.

(f) F sheets:

Cols. 6-8: The constituency held by the M.P. (The three-digit number corresponds to the position of the constituency in the alphabetical ordering of all consituencies.)

Col. 9: 'S' means the consituency is in Scotland

'I' " " " Ireland

'E' " " " England or Wales.

Col. 11: 'B' " " " a borough constituency

'C' " " " county "

Col. 12: '2' means the M.P. is in his 20's )

'3' " " " 30's ) (Some ages

'4' " " " 40's ) were guessed

'5' " " " 50's ) from evidence

'6' " " " 60's ) about

'7' " " " 70's ) university

'8' " " " 80's ) education etc.)

'9' " " " 90's )

'0' " we have no information. )

Col. 13: 'A' means the M.P. had no brother as an M.P. (in 1861)  
          'B'       "       "       had a brother as an M.P. (in 1861)  
Col. 14: 'F'       "       "       had another member of family as M.P.,  
  e.g. father, father-in-law  
          'U'       "       "       was unrelated to any other M.P.,  
  or we have no information  
Col. 16: 'P'       "       "       is a past government office holder  
          'G'       "       "       "       present       "       "       "  
          'H'       "       "       has never been       "       "       "  
Col. 17: 'M'       "       "       has been in the militia  
          'R'       "       "       "       the regular army  
          'X'       "       "       "       neither.  
Col. 18: 'N'       "       "       "       the navy  
          'X'       "       "       has not been in "

The design of the E sheets, which were the first to be prepared, made it simple to encode the original division lists, division by division, onto the coding sheets. An extra copy of the E sheet cut along a line between the 10th and 11th columns meant that the names could be placed alongside the column to be coded. It was then possible to run through the Aye voters inserting 'A's appropriately and then to do the same for the Noes. 'Z' values were inserted first of all and after that 'A', 'N', 'B' and 'M' values were installed. All the remaining values had to be 'X' and these were filled in en masse. The final task was to produce the 'F' sheets and these were tackled independently. This proved to be quite an efficient arrangement and the process was considerably enhanced by the acquisition of the services of Mrs. Susan Thomas to do the bulk of the coding. Her degree in English and History provided her with a background knowledge of the period (through Trollope etc.) and this

meant that she found the encoding interesting. As a result, not only was the final product very accurate, but also Mrs. Thomas had been able to contribute some interesting insights from her firsthand knowledge of the data.

The punching of the data from the 162 coding sheets was achieved quite rapidly and without enormous amounts of error being introduced. The cards were punched in six batches, A sheets, then B sheets etc. The cards were then read onto the University of Bath computer and the error-removing process began, the steps of which were to:-

(i) Obtain a print of the data.

(ii) Match acronyms from one batch to another, collating the output for each M.P. This was done by sorting the acronyms into alphabetical order for each successive batch and looking for mismatches, of which there were a few in each batch. The faulty acronyms were then corrected and the data for the M.P.s merged together.

(iii) Search through the  $m_{ij}$  values to discover impossible values and compare the printout and coding sheets to identify the correct code.

(iv) Develop a program to produce the total of Ayes and Noes in each division and compare this with the published version. Where there were errors, the program produced a facsimile of the original division list so that the culprits could be identified. This occurred in about one half of the divisions. Errors occurred in punching rows of the coding sheets and, rarely, in coding. Typically codes for neighbouring divisions were transposed.

(v) Produce special range-error programs for the F sheets, which were particularly accurately punched, and check through the listing of the E sheet cards to identify nonsensical spellings, of which again there were few.

This process was quite time consuming, yet still not as exhaustive as it could have been, for each computerised division list could have been compared with the original. No doubt the final version still contained a few errors; those that cancelled out and did not affect the totals of Ayes and Noes would certainly not have been spotted. However the overall impression we gained was that the encoding had been very accurate and the punching quite accurate. Certainly, in the face of such large numbers, we felt that it was unlikely that the residual error was of sufficient importance to vitiate the results of our study. It has been our experience that multidimensional scaling is sensitive to error in data and able to show up curious behaviour and subsequently we have had our maps to use. That Col. Samuel Auchmuty Dickson (DICK), an alleged liberal, consistently appeared amongst the conservatives in the maps seemed unexpected, but recourse to the original data in this and other cases showed that the error was not in the encoding but presumably in his party label.

The next step was to change the format of the data to make it more amenable to use by computer and conserve space where possible. The intermediate blanks were removed, as were the repetitions of acronym and trailing year and sheet identification. Two files were produced, one containing acronyms and biographical information, the other acronyms and voting behaviour. The voting file contained just under 120,000 characters.

Three additional descriptive files were generated. The first contained the official division titles. The second contained a list of division subjects as shown in Table 4.3.2. Each division had to fall into at least one of these subject matter categories, but often more than one was appropriate. For example, taxation, government spending and defence often overlapped. The third file contained Valerie Cromwell's

TABLE 4.3.2

Categories of Divisions

<u>No.</u>	<u>Category of Subject Matter</u>	<u>Total Such Divisions in 1861</u>
1.	General, e.g. Queen's Speech, procedural matters.	26
2.	Foreign policy. .. .. .	4
3.	Taxation, revenue, pressure to economise, government spending. .. .. .	52
4.	Social problems. .. .. .	14
5.	Electoral arrangements, e.g. voting qualifications, constituency boundaries. .. .. .	22
6.	Religious and ecclesiastical matters. .. .. .	15
7.	Defence. .. .. .	23
8.	Miscellaneous; personal matters. .. .. .	2
9.	Railways, roads, harbours. .. .. .	12
10.	Irish matters. .. .. .	22
11.	Local government and revenue. .. .. .	12
12.	Scottish matters. .. .. .	12
13.	Education, universities, schools. .. .. .	15
14.	India. .. .. .	9
15.	Business regulation. .. .. .	10
16.	Legal reforms and rationalisation. .. .. .	4
17.	Civil service reform. .. .. .	7
18.	Fisheries. .. .. .	2

allocation of divisions into categories, but no completely objective approach could be adopted.

It was hoped that these processes would make widely available a useful political and historical source. This encoded data has provided a valuable, machine-readable source for a wide range of researchers. To date, complete sets of printed Commons' Division Lists have only been accessible in a very limited way in London. Our magnetic tapes enable relevant information to be obtained much more quickly and economically than from the original printed lists. For example, information about tellers has been derived with ease. The tapes have been lodged with the Social Science Research Council Survey Archive. In addition other tapes containing the data have been located at the University of Bath and the University of Sussex.



#### 4.4 A Preliminary Approach

##### Preliminary Analysis of M.P.s

The raw data was used firstly to compile lists of M.P.s with respect to various biographical details. A series of such lists corresponded to each separate value taken by each variable defined in the F sheet encoding described in Section 4.3.

Apart from being of intrinsic interest, these lists provided a coding check and enabled the formation of several subsets of M.P.s that were later to be used in the multidimensional scaling analyses. Specifically, five sets were formed, each of which was a subset of one of these lists, chosen to satisfy a minimum voting criterion.

Thus we formed:-

- (a) A group of the 79 most frequently voting past & present officers
- (b) " " " 37 " " M.P.s in the regular army
- (c) " " " 58 " " " " militia
- (d) " " " 95 M.P.s used in (b) and (c)
- (e) " " " 62 " " " " " with Irish constituencies.

It was hoped that set (a) would be informative for all divisions, but that the three sets (b), (c) and (d) would be especially interesting with regard to defence matters, and set (e) interesting in Irish and religious divisions.

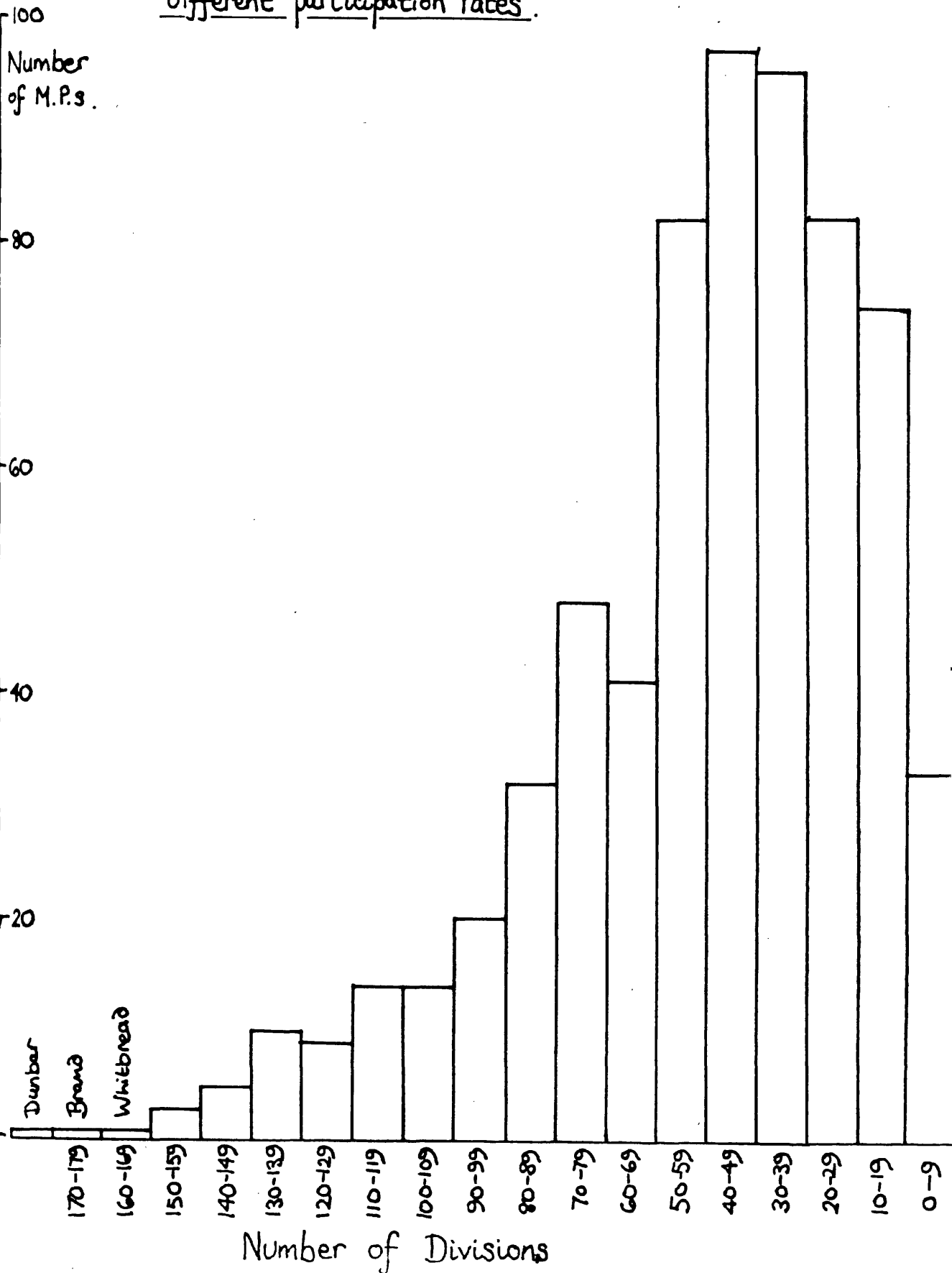
The next area covered was voting participation by individual M.P.s. All 662 were formed into a league table according to the number of votes that they registered. This number ranged in value from 182 out of 187 (Sir William Dunbar, an active liberal whip) to 0 (eight different M.P.s). Thirty-three M.P.s recorded less than 10 votes. A histogram of voting frequency is shown in Fig. 4.4.1.

That the three most prolific voters were three prominent liberals, including the two party whips, enabled us to be more confident in our estimate of the position taken by each party in each division. To do this we determined the vote made by:-

- (i) The majority of conservatives participating
- and
- (ii) The majority of liberals, radicals, reformers and whigs participating in each division.

This was uniformly straightforward for conservatives because we discovered that they were much more united as a group. However the liberals were more often divided and to check that we were obtaining the opinion of the centre of the party we compared the majority position with that of the senior whips and cabinet members and confirmed that these votes coincided. For example, there was only one discrepancy for the regularly voting Henry Brand. As a useful means of identifying probable dissidents we ordered each M.P. in the two groups we have just defined, of conservatives and of liberals with their allies, according to the number of times they voted against the majority of their party colleagues, and provided a break-down of the

Fig 4.1 Histogram showing  
different participation rates.



disagreements into the numbers coming in each of the eighteen categories of division subject matter. The liberal and allies group was much more split. Altogether 130 liberals disagreed at least ten times, with one disagreeing 75 times out of 87 (the incorrigible Colonel Dickson). By contrast only 51 conservatives disagreed at least ten times and the most disagreeing voice was that of John Pope Hennessy (40 out of 139). Liberal disagreements were often on matters of defence (category 7). This exercise enabled us to form two groups for subsequent use in multidimensional scaling.

(a) The 95 most frequently dissident liberal and allied M.P.s

(b) The 70 " " conservative M.P.s.

#### Preliminary Analysis of Divisions

The raw data was used to compile facsimiles of the original division lists, copies of which were made along with the full division title and division category descriptions. The divisions were also checked with the original lists in order to ensure that the totals of votes for and against matched the correct figure. This enabled the formation of the histogram shown in Fig. 4.4.2 which shows for each category of division the minimum, mean, maximum and sample size of votes for that category. The division attracting most votes was in category 3, finance. The category with highest mean voting was the fifth, electoral arrangements.

Four further analyses of the divisions were produced.

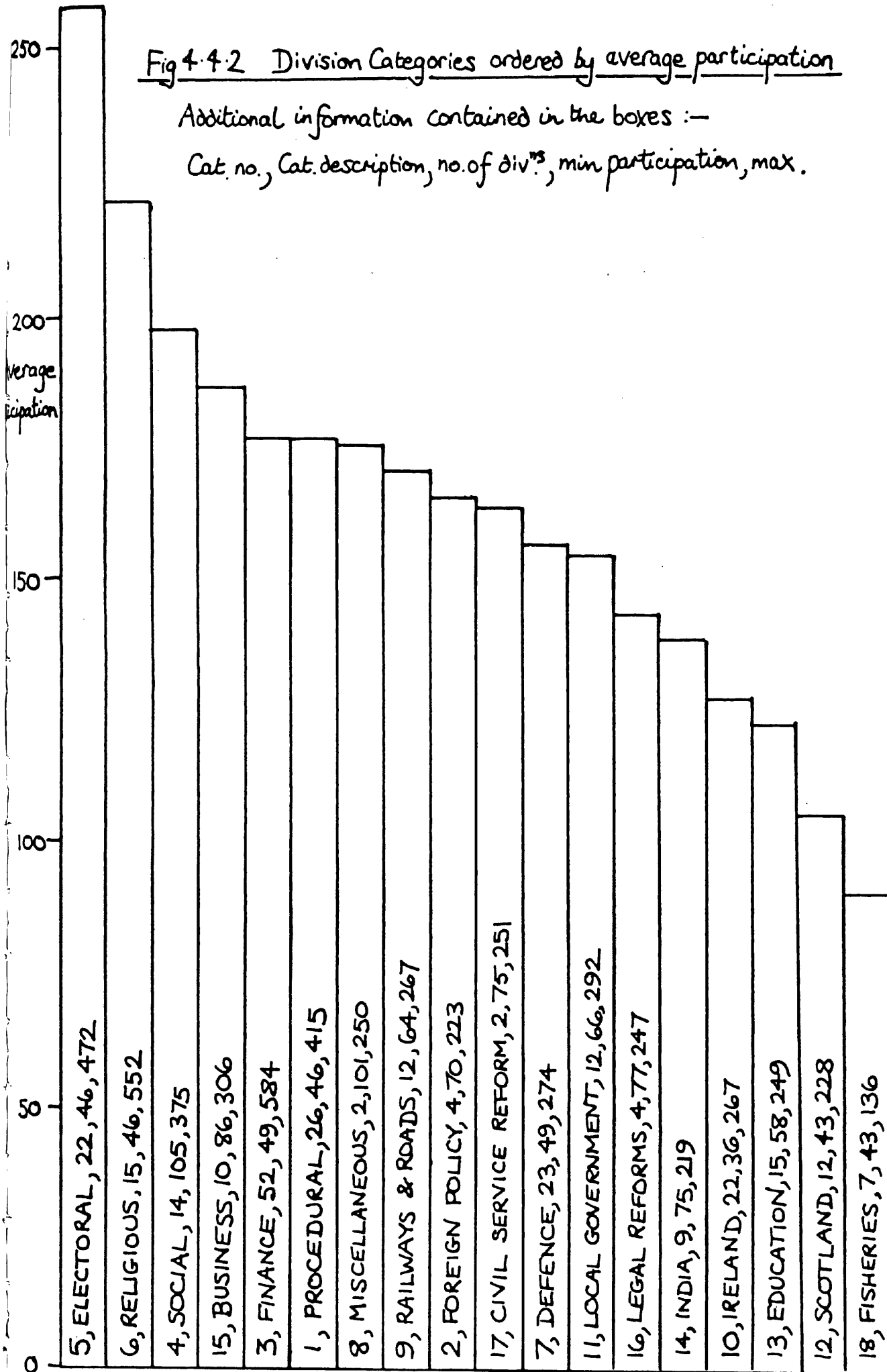
1. For each division successively the total of votes for and against was broken down into party contributions.

Thus, we obtained for the first division:-

Fig 4.4.2 Division Categories ordered by average participation

Additional information contained in the boxes :-

Cat. no., Cat. description, no. of div<sup>'s</sup>, min participation, max.



	<u>Ayes</u>	<u>Noes</u>
Liberals	38 (41%)	55 (59%)
Conservatives	0 (0%)	49 (100%)
Liberal-Conservatives	0 (0%)	24 (100%)
Reformers	7 (100%)	0 (0%)
Radicals	1 (100%)	0 (0%)
Whigs	0 (0%)	1 (100%)

Both the liberal and conservative majorities were on the side of the Noes. Suggestively the reformers and radicals were firmly on the side of the Ayes, the whigs and liberal-conservatives firmly against them. It seemed that even after the first division only, evidence was being gained for the political structure suggested in Section 4.1.

2. On the basis of the majority positions a table was set up to try to get a first impression as to which categories of division might be causing most inter-party disagreement. The figures, shown in Table 4.4.3, suggested, for example, that there was often disagreement concerning social and electoral problems, but relatively more agreement over financial matters.

3. Each division was classified into one of eighteen exhaustive and mutually exclusive cells according to the following three variables:

- X:- majorities agree, X=0  
majorities disagree, X=1
- Y:- liberal majority was 50- 70%, Y=0 (where the intervals  
" " " 70- 90%, Y=1 had closed  
" " " 90-100%, Y=2 right hand ends)
- Z:- Conservative majority was 50- 70%, Z=0  
" " " 70- 90%, Z=1  
" " " 90-100., Z=2.

TABLE 4.4.3

<u>Category</u>	<u>Number of Times Liberal &amp; Conservative Majorities Agree</u>	<u>Number of Times Liberal &amp; Conservative Majorities Disagree</u>
1. Procedural	10	16
2. Foreign Policy	4	0
3. Finance	30	22
4. Social	1	13
5. Electoral	5	17
6. Religious	6	9
7. Defence	14	9
8. Miscellaneous	0	2
9. Railways & Roads	7	5
10. Ireland	12	10
11. Local Government	6	6
12. Scotland	8	4
13. Education	5	10
14. India	1	8
15. Business	3	7
16. Legal Reform	3	1
17. Civil Service	0	2
18. Fisheries	6	1

For example, cell (0,2,2) corresponded to divisions in which liberals and conservatives had clear voting patterns which agreed, and contained the five divisions numbered 104, 114, 135, 155 and 187. Strong disagreement, represented by (1,2,2), was found in 21 divisions. The most full cell was (1,1,2) which showed consistency among conservatives, less among liberals, disagreement between the parties and contained 24 divisions. Division one, as shown above, fell into (0,0,2).

4. For each particular category of division the distribution of divisions into the above 18 cells was shown. Placing  $(x,y,z)$  in position  $9x + 3y + z + 1$  in the following vectors, we expressed the distribution for category 4 divisions (social) as

(0,0,0,0,0,0,1,0,0,0,1,1,1,1,2,4,1,2)

This contrasted with category 3 divisions (finance) for which we obtained

(5,7,3,1,4,5,1,2,2,4,1,0,1,4,1,1,5,5)

These distributions were presented in the tabular form:

		Y=0	Y=1	Y=2	
	(Z=0	1	4	7	
X=0	(Z=1	2	5	8	
	(Z=2	3	6	9	(The integer
					in each location
	(Z=0	10	13	16	represents the position
X=1	(Z=1	11	14	17	in the vector)
	(Z=2	12	15	18	



#### 4.5 The Measurement of Dissimilarity

A crucial aspect of the use of techniques designed to analyse similarity data is the measurement of the similarity values themselves. In this section we provide a brief justification of the practice that has been used throughout this particular study, that of using a Jaccard coefficient.

It was our aim to construct a measure of similarity of voting behaviour for every pair of M.P.s. For each pair we compared their two vectors or voting profiles, individual components of which were any one of the six values defined in Section 4.3. Our first simplification of this specification was to regard 'B' and 'M' votes which corresponded to tellers as equivalent to 'A' and 'N' votes respectively. There seemed little harm in this. More drastically, but in keeping with the desire expressed in Section 4.1 that we should only deal with positive evidence of attitudes, we chose to ignore those components for which either of the pair registered an 'X' or 'Z' vote. Thirdly we also maintained our expressed desire to regard each individual division as having equal importance in the final derivation of the coefficient. It would have been possible to weight divisions by participation rates, for example, but this would have contradicted the feeling that all divisions provided an indication of opinion that required equal respect. In practice it was not clear that a weighted coefficient would substantially alter the rank ordering of the pairs.

Thus we were left with four possible combinations of values in the remaining components of the vector:

	<u>First M.P.</u>	<u>Second M.P.</u>
	Aye	Aye
	No	Aye
	Aye	No
	No	No

The next condition was that the reversal of the entire set of votes in a division should not be allowed to affect the measure of similarity so that the values 'Aye' and 'No' were important only in that they reflected either agreement or disagreement. Thus the problem was reduced to a comparison of the number of agreements and the number of disagreements. It then seemed natural to use the Jaccard similarity coefficient

$$\frac{\text{No of agreements}}{\text{No. of agreements} + \text{No. of disagreements}}$$

which had the useful property of lying in the closed interval from 0 to 1. The corresponding dissimilarity value was the difference from 1. The denominator could equally well have been written as the number of divisions in which both M.P.s participated. This raised the question as to how to define the coefficient if the pair had no divisions in common. We chose not to define it and indeed, if the common number of divisions was less than five then we also declined to produce a value. Five was chosen as a minimum acceptable value to allow any meaningful interpretation. The occasional absence of similarity values did not prevent us from using any of the scaling techniques with the exception of the algebraically based classical method. This was not a great hardship, for the similarity values had not been designed to be nearly-Euclidean (for example, they were bounded above by 1) and we only used classical scaling to generate a starting configuration for ordinal scaling. Plotting dissimilarity against distance in the final configuration after scaling showed that the dissimilarity values were very widely spread over the possible range.

#### 4.6 Definitions of Groups Used in the Analyses

In the following sections statistical analyses are applied to several groups of M.P.s. It is the purpose of this short section to provide concise definitions of these groups for reference purposes. The definitions depended upon three 'league table' orderings of M.P.s, as introduced in Section 4.4.

(a) The league table of M.P.s with regard to their total participation in the session. (Highest = most voting.)

(b) The league table of conservative M.P.s with regard to their total number of votes against the conservative majority in divisions. (Highest = most dissenting.)

(c) The league table of liberal, radical, reformer and whig M.P.s with regard to their total number of votes against the liberal majority in divisions. (Highest = most dissenting.)

Seventeen groups of M.P.s were defined:-

1. Cohort 1: 1-100 in (a) above
2. Cohort 2: 51-150 "
3. Cohort 3: 101-200 "
4. Cohort 4: 151-250 "
5. Cohort 5: 201-300 "
6. Cohort 6: 251-350 "
7. Cohort 7: 301-400 "
8. Cohort 8: 351-450 "
9. Cohort 9: 401-500 "
10. Cohort 10: 451-550 "
11. Liberal Dissenters: 1-95 in (c) above
12. Conservative Dissenters: 1-70 in (b) above
13. Irish Based: The top 69 Irish based M.P.s in (a) above
14. Office Holders: The top 79 past or present office holders in (a) above

15. Regular Army: The top 37 M.P.s in the regular army in (a) above
16. Militia: The top 58 M.P.s in the militia in (a) above
17. Military: The union of regular army and militia.

In what follows we focus attention on the results for Cohort 1, which is used to illustrate the type of results that have been obtained. Results for other groups are briefly discussed.

#### 4.7 Results Obtained by Single-Link Clustering

Single-link clustering has been applied to Cohorts 1 to 8, liberal and conservative dissenters, Irish members and office holders. The analyses were only performed for similarities based on all divisions. Temporarily excluding the two dissenting groups, we may summarise the typical order of cluster formation for the rest as follows. The successive stages were obtained by gradually increasing the constant defining the clusters. (See Section 1.8 for more details.)

Stage 1. A large cluster of similarly voting liberal M.P.s was formed. These M.P.s voted very similarly to the members of the cabinet.

Stage 2. As this first group increased in size another cluster of conservative M.P.s formed.

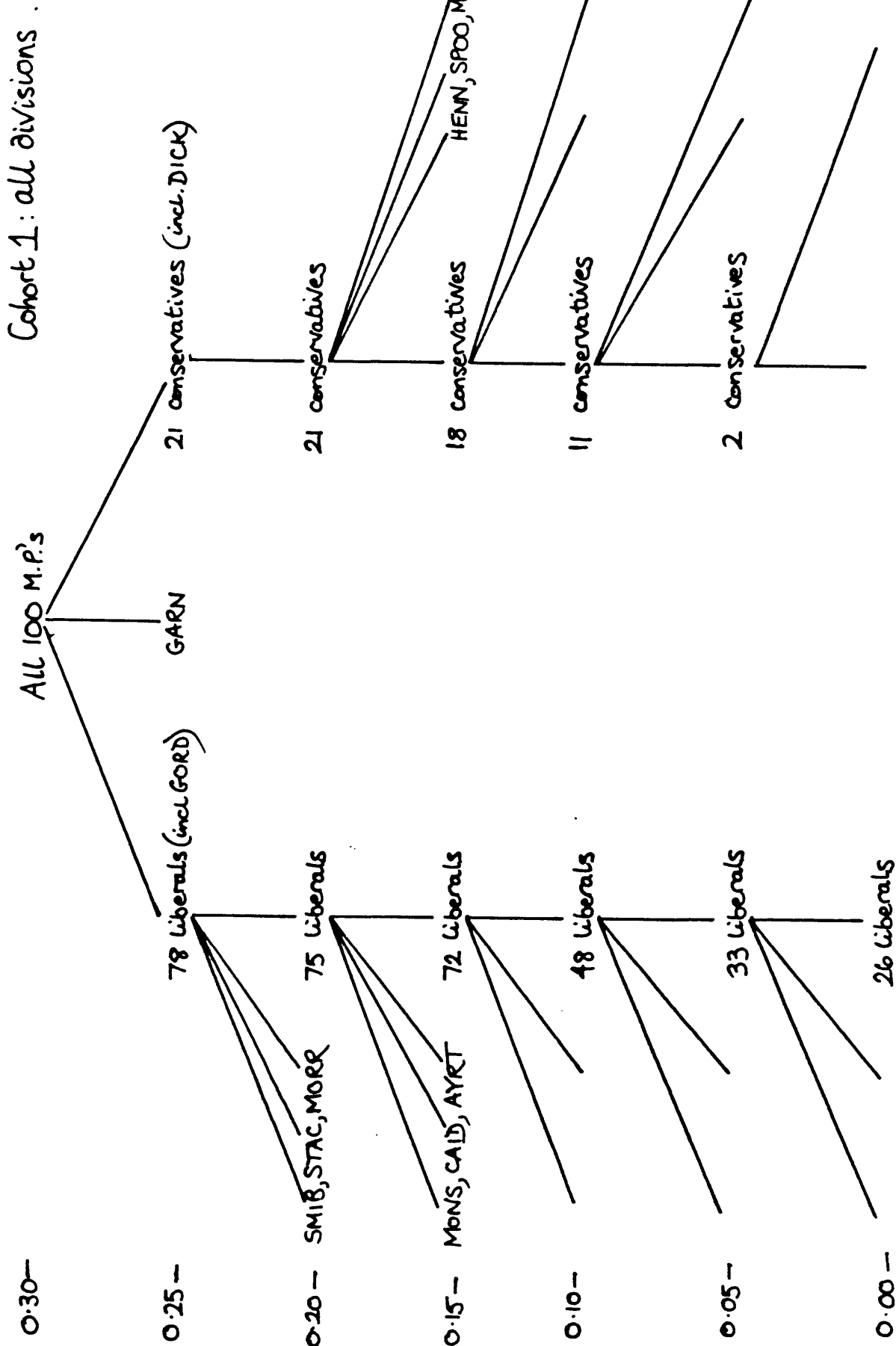
Stage 3. These two clusters both increased in size. Other clusters were uncommon, small and soon became engulfed in the two main groups. The growth of the conservative group was faster and thus more quickly completed. Meanwhile the liberal group began to include those M.P.s, known to have radical views, who appeared among the dissenters.

Stage 4. The liberal and conservative groups amalgamated, leaving a few unattached individuals.

An impression of the resulting dendrogram is given for Cohort 1 in Figure 4.7.1.

Thus the historical suggestion was that there existed a strongly united liberal cabinet with supporters which faced opposition from both the less cohesive conservative group and the radical sympathisers. These radicals were more akin to the liberals than conservatives.

Dissimilarity



The other two groupings had a similar structure to one another. For both dissenting groups a large cluster emerged that engulfed all temporary smaller ones leaving just a few anomalous individuals. For the conservatives this cluster started with orthodox conservatives who had voted frequently and thus their proportion of dissenting votes was small. Gradually the more extreme conservatives joined this group until only very abnormal voters remained. With the liberals the original members were radicals and the remaining few at the end were those who voted rather as conservatives.

Having completed this project these results became more clearly understood with the hindsight allowed by having used scaling methods. It was then possible to retrace the formation of clusters through the solution provided and see how each newly introduced individual related to the other group members. The only information provided by single-link clustering was the identity of the specific pair of M.P.s whose voting similarity caused the link to be formed. There was also a suggestion from the scaling maps that the tendency for large clusters to dominate the analysis was caused by the chaining effect of single-link. Other cluster analysis techniques, such as average-link, would not be so prone to this problem. Perhaps they would have identified a separate group of radicals. With single-link in this application, pairs of M.P.s were often immediately united through a third member agreeing with them both on all common divisions. Certainly the method was very efficient computationally, and could have coped with a single clustering of a much larger group had this been attempted. However, on balance, the results were not nearly as interesting as those to which we now turn.

#### 4.8 Results Obtained by Ordinal Scaling

Sixty-four combinations of groups of M.P.s and sets of divisions were examined to test the usefulness of ordinal scaling when applied to legislative data. An array showing these combinations is provided in Table 4.8.1. Cohorts 1 to 5 and the two dissenting groups were used in conjunction with all of the larger categories of division. Cohorts 6 to 10 voted less, so that scaling on proper subsets of the set of divisions was not attempted. The military groups were used in conjunction with defence divisions; the Irish-based group was examined for divisions relating to religion, finance and Ireland itself; office holders were also monitored for their voting on finance.

An outline of the computational arrangement used for these scalings is provided in Table 4.8.2. The input requirements were merely specifications of the required configuration dimensionality, the M.P.s to be used, the divisions to be used and a list of output options. The versatility of the Honeywell Multics system enabled temporary store to be used for all the intermediate files, and allowed jobs to be submitted at pre-specified, inexpensive times of the day. Thus the effort involved in running the programs was minimal. Extra graphical facilities were provided by a Tektronix 4014 graphics terminal with an attached hard copy unit, and a compatible Calcomp plotter allowing four ink colours.

Similarity values were produced according to the Jaccard coefficient of Section 4.5. Thus similarities were only defined for a pair when they voted in at least five common divisions. The number of values to which each M.P. contributed was calculated. If an M.P. contributed to ten or less values his positioning was regarded as dubious and he was removed from some of the plotted



TABLE 4.8.1

Combinations of M.P. Groups and Division Categories

	ALL Divisions	Procedural (1)	Financial (3)	Social (4)	Electoral (5)	Religious (6)	Defence (7)	Irish (10)	Religious & Irish (6 & 10)
Cohort 1	✓	✓	✓	✓	✓	X	✓	✓	X
Cohort 2	✓	✓	✓	✓	✓	X	✓	✓	X
Cohort 3	✓	✓	✓	✓	✓	X	✓	✓	X
Cohort 4	✓	✓	✓	✓	✓	X	✓	✓	X
Cohort 5	✓	✓	✓	✓	✓	X	✓	✓	X
Cohort 6	✓	X	X	X	X	X	X	X	X
Cohort 7	✓	X	X	X	X	X	X	X	X
Cohort 8	✓	X	X	X	X	X	X	X	X
Cohort 9	✓	X	X	X	X	X	X	X	X
Cohort 10	✓	X	X	X	X	X	X	X	X
Liberal Dissenters	✓	✓	✓	✓	✓	X	✓	✓	X
Conservative Dissenters	✓	✓	✓	✓	✓	X	✓	✓	X
Irish Based M.P.s	✓	X	✓	X	X	✓	X	✓	✓
Office Holders	✓	X	✓	X	X	X	X	X	X
Regular Army	X	X	X	X	X	X	✓	X	X
Militia	X	X	X	X	X	X	✓	X	X
Military	X	X	X	X	X	X	✓	X	X

TABLE 4.8.2

A Summary of the Computational Arrangement for Ordinal Scaling

Input

- Specify (i) The identification numbers of the M.P.s to be used  
(ii) The divisions to be considered (via a format statement)  
(iii) The number of solution dimensions required.

Program

- (i) Produces all possible similarity values (five common divisions needed).  
(ii) If all values defined, forms starting configuration by classical scaling.  
(iii) Otherwise uses stored final configuration from ordinal scaling when applied to all divisions, for a starting configuration.  
(iv) Runs ordinal scaling.

Output

(i) To the Line Printer

- (a) A list of those M.P.s (if any) who had ten or less similarities defined with other M.P.s, with their actual number.  
(b) If classical scaling was used, a report.  
(c) A report of the ordinal scaling progress and results.  
(d) The final configuration.

(ii) To the Tektronix 4014 Graphics Terminal

- (a) A plot of dissimilarity values against final configuration distances.  
(b) For two-dimensional configurations, plots of final configuration both with and without those M.P.s who had ten or less similarities defined.  
(c) For three-dimensional configurations, perspective plots of the final configuration both with and without those M.P.s who had ten or less similarities defined, with minimum spanning tree from single-link clustering superimposed.

(iii) To the Calcomp Plotter

Just as to the Tektronix 4014, except that individual M.P.s were identified by ink-colouring according to party allegiance.

configurations. However the values were still used in the scaling iterations. If an M.P. voted in less than five of the available divisions he could not have had any values defined. His position would not then have been adjusted in the iterations, and the final position would have been arbitrary and meaningless. This would also have applied if the M.P. had voted five or more times, but still not enough to find five common divisions with any other member. The reliability of the final position was thus dependent upon the number of votes cast, via the number of similarities defined.

Several aspects of this procedure need justification. The choice and definition of similarity measure was defended in Section 4.5. The choice of ten as a minimum sensible number of similarities was based upon the results of Section 3.4 concerning the accuracy of configurations produced from subsets of the similarity matrix. This was a conservative estimate, and it was thus felt that the final configuration would still be improved by the inclusion in the iterations of those M.P.s for whom some values were defined, but not enough to satisfy this requirement, and who thus would not appear on the final plots. Indeed the basis of the acceptance of an M.P.'s position was the existence of more than ten values with other M.P.s, some of whom might not have been plotted.

Our results suggest that an attractive extension of technique would be to use different definitions of the M.P. groups. For example, if each cohort had included a number of very regularly voting M.P.s, then fewer positionings would have had to be regarded as dubious. This would have been advantageous, especially if the added regulars had been widely spread across the spectrum of opinion, and would have thus provided a useful comparison between

plots. The extent of the problem may be gauged from Table 4.8.3 which provides the numbers of M.P.s failing to produce more than ten values in the similarity matrix. One hundred M.P.s made up each cohort. Some division categories had high turnouts and large numbers of divisions. It was the other categories that caused the problems.

It was necessary to take precautions that would enable the ordinal method to converge at, or very near to, the global optimum. In the case of all values being defined, classical scaling was used to generate a starting configuration. This procedure usually speeded up the convergence and added reliability to the final values (see Section 3.5). Additionally the eigenvalue spectrum was helpful in suggesting how many dimensions might be appropriate. Classical scaling was nearly always possible when a group of M.P.s was analysed on all divisions of the House. Cohort 10 was the exception. The final ordinal scaling configuration started from the classical scaling output was reliable, and it was used as a starting configuration if certain values were missing. The overall impression gained from the results on all divisions was that party allegiance and grouping was the dominant factor, and that this was likely to be the case on smaller categories of division. The use of the configuration for all divisions had the merit of resolving the group of M.P.s into this party structure, ensuring that the worst forms of local minimum behaviour were unlikely.

Ordinal scaling was based upon the starting configuration we have just defined, using the standard global ordering of dissimilarity values, the primary treatment of tied values and up to fifty iterations. Occasionally more iterations were used if convergence did not seem near. However under the arrangements we have described this was quite rare. Progress reports of the

TABLE 4.8.3

Number of M.P.s for Whom Ten or Less Similarity Values were Defined

<u>Group</u>	<u>Division Type and Category Number</u>						
	All Divisions (1-18)	Procedural (1)	Financial (3)	Social (4)	Electoral (5)	Defence (7)	Irish (10)
Cohort 1	0	0	0	17	0	0	3
Cohort 2	0	3	0	26	0	6	31
Cohort 3	0	12	0	41	0	38	72
Cohort 4	0	23	0	63	2	64	92
Cohort 5	0	42	0	84	5	72	95
Cohort 6	0	-	-	-	-	-	-
Cohort 7	0	-	-	-	-	-	-
Cohort 8	0	-	-	-	-	-	-
Cohort 9	0	-	-	-	-	-	-
Cohort 10	1	-	-	-	-	-	-

- means that the combination was not attempted

classical scaling, and the ordinal scaling iterations were provided along with the final configuration.

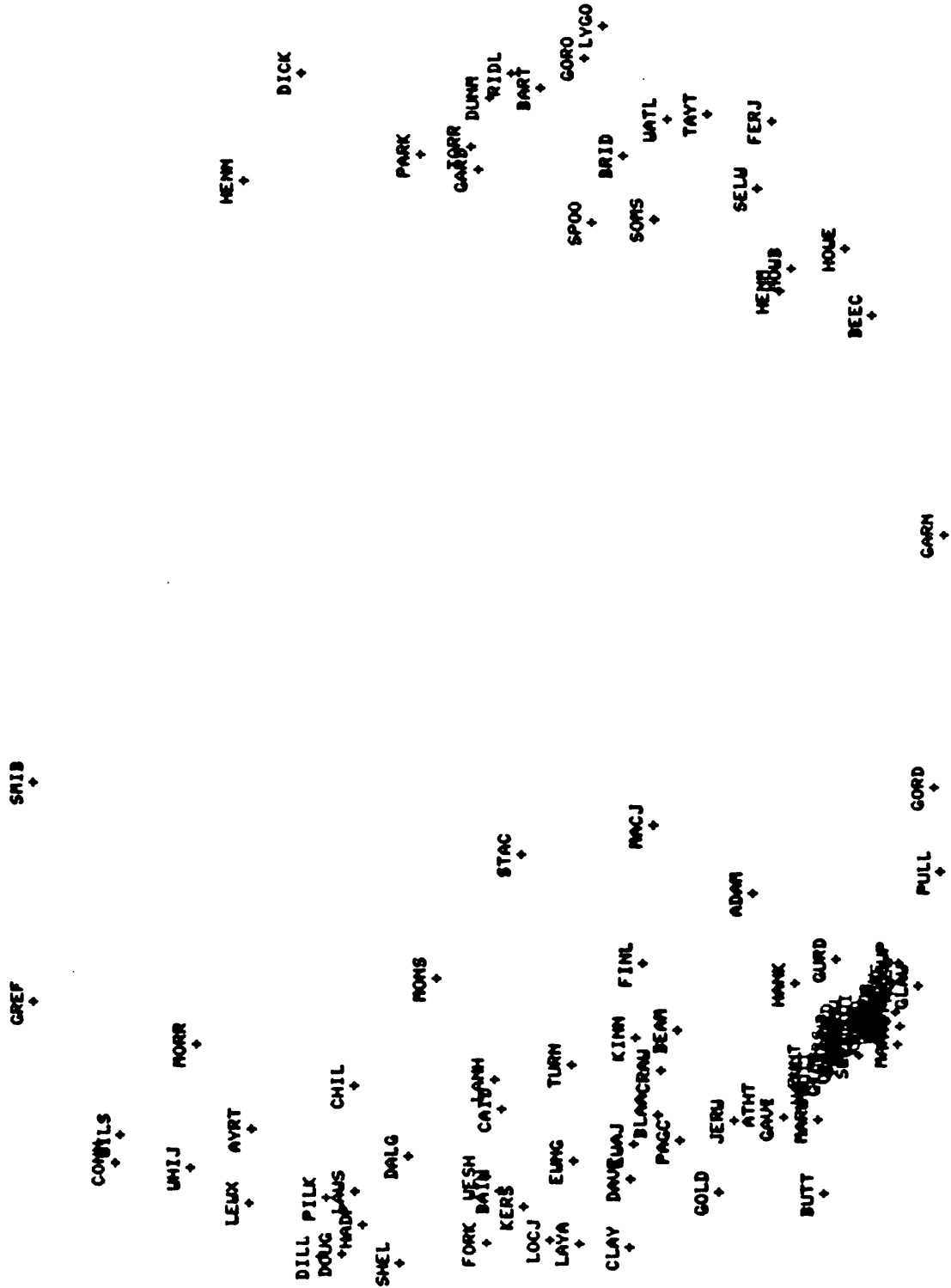
The graphical facilities greatly enhanced the appreciation of the final configurations. The S.S.R.C. Spatial Data Project Tektronix 4014 terminal allowed efficient development of plotting routines, and this was complemented by the Calcomp plotter which provided permanent versions, coloured to highlight party groupings. The ability to colour proved valuable in interpreting the M.P. clusters. Plots were produced of the final configuration with and without any dubiously placed M.P.s, the scatterplot of dissimilarity and configuration distance values, and for three-dimensional configurations a pair of perspective views with M.P.s linked by chains in the single link spanning tree. This last option was not extensively used in this application because three dimensions were rarely used and the number of M.P.s in each group was prohibitively large.

Three plots of configurations relating to Cohort 1 are provided. The first, Fig. 4.8.4, is based upon all divisions and we follow its generation in detail.

All dissimilarity values were defined, so that classical scaling was used to provide the starting configuration. When the output configuration was rescaled to nine dimensions the loads on each dimension were as follows:-

Dimension 1		0.55
"	2	0.23
"	3	0.04
"	4	0.04
"	5	0.03
"	6	0.03
"	7	0.03
"	8	0.03
"	9	0.03

FIG. 4.8.4 Ordinal Scaling of Cohort 1 on All Divisions



The sum of all 100 eigenvalues was 9.990 which compared with:

the sum of the first	eigenvalue	being	6.762
" " " two	eigenvalues	"	9.570
" " " three	"	"	10.04

The magnitude of the first	eigenvalue	was	6.762
" " second	"	"	2.808
" " third	"	"	0.471
" " one hundredth	"	"	-0.675

Thus inspection of the eigenvalue spectrum, the trace criterion and the magnitude criterion all suggested that the underlying structure should be represented in two dimensions.

The corresponding two-dimensional solution from ordinal scaling converged rapidly.

<u>Iteration</u>	<u>Step</u>	<u>Slope</u>	<u>Stress</u>
<u>Number</u>	<u>Size</u>	<u>Size</u>	<u>Value</u>
0	—	0.00365	0.14019
5	0.02375	0.00025	0.11720
10	0.01411	0.00024	0.11705
15	0.00100	0.00004	0.11696
20	0.00047	0.00002	0.11695
25	0.00028	0.00001	0.11695
30	0.00020	0.00001	0.11695
35	0.00008	0.00000	0.11695
40	0.00006	0.00000	0.11695
45	0.00008	0.00001	0.11695
50	0.00004	0.00000	0.11695



At this stage we defend the presentation of this solution as the final version, rather than solutions of any other dimensionality. It would be difficult to justify the use of more dimensions because classical scaling indicated so strongly that the underlying structure was two-dimensional, and any extra dimensions would be much more complicated to grasp. That the two-dimensional optimal configuration changed little from the classical scaling version reinforces this view. Moreover, both dimensions are important. It might be appealing to see if a one-dimensional solution would unwrap the 'horse-shoe' of Figure 4.8.4 , but we dispel such ideas by demonstrating that each dimension has a simple interpretation that would be lost by such an exercise.

For each division we have already defined whether the positions taken by the majority of each party agreed or not. For each individual M.P. we then defined the two values,  $x$  and  $y$  , as:-

$x$  = proportion of conservative-type votes in divisions of disagreement in which he participated. .

$y$  = proportion of minority-type votes in divisions of agreement in which he participated.

The agreement between the ordinal scaling configuration and the configuration produced by this means was quite extraordinary, and demonstrated that those M.P.s who would have been supposed to be most separated in an unfolded one-dimensional solution, actually adopted a similar dissenting attitude to divisions in which the majority of both parties agreed. The liberal cabinet was characterised by  $y=0$  , indicating allegiance to the majority in divisions of agreement, and  $x=0$  , showing liberal support. Apart from this superposition the M.P.s were arranged in a manner very similar to that obtained by ordinal scaling. The same surprise

positionings occurred and the same individuals were identified as dissidents.

This representation of the voting behaviour of M.P.s in all divisions of the session presented a clear pattern which was to a large extent repeated for other Cohorts. As was expected, the bulk of the liberal cabinet and office holders voted in a similar way. Only three were at all distant from the cluster formed by the rest. The leaders of the conservative opposition voted less frequently and less coherently. For example, the leader of the opposition, Disraeli (DISR), did not appear until Cohort 4. The radical liberals were even less coherent as a group, though quite separate from other liberals. In addition to these three sets it was easy to identify individuals voting unusually. We refer to four. Colonel Dickson (DICK) was described by Dod as "liberal; and in favour of civil and religious liberty". He represented Limerick County. He was placed further away from the cabinet than almost all conservatives. In the light of this evidence it is interesting to observe that he was a member of the Carlton Club and was destined to back the vote of censure on Palmerston in 1865. Close to Colonel Dickson on the extremes was John Pope Hennessy (HENN), "a supporter generally of Lord Derby". He was another Irish member and the first Roman Catholic conservative elected. Augustus Smith (SMIB), lessee of the Scilly Isles and described as a liberal "in favour of a wide extension of the franchise", was very isolated even from the loose group of radical liberals. A conservative who was placed close to the liberal cabinet (as voting inspection justifies) was Charles William Gordon (GORD), although nothing in his career suggested such political sympathy. Of the six liberal-conservatives, three opted for liberal, and three for conservative, positionings,

revealing their true colours.

The other two plots provided deal with the twenty-two electoral reform divisions (Figure 4.8.5) and the twenty-three defence divisions (Figure 4.8.6). The configuration for electoral reform shows a similar form to that for all divisions, but extended in the x-direction. Here the suggestion is that there is an obvious two-party split. Exceptions are Augustus Smith (SMIB, close to the conservatives), the liberal William Garnett (GARN, even closer) and Alexander Finlay (FINL, alongside), a Commissioner of Supply. More striking is the configuration for defence divisions. This is the most different from that for all divisions, whichever Cohort is considered. A very large though scattered liberal dissenting group was close to a number of conservatives, whilst other conservatives were closer to the cabinet than their supposed party colleagues. For example, three former conservative office holders, Joseph Henley (HENM), John Mowbray (MOWB) and Colonel Taylor (TAYT) were close to or in the group containing the liberal cabinet, whilst their positions on all divisions were quite different. Much of the differentiation among liberals and among conservatives that may be seen in the all-divisions configuration is thus accounted for by the defence divisions.

We summarise the results for other categories of division. In Category 3 (taxation, pressure to economise etc.), although the government was very closely bunched because supply had to be approved if it was to survive, other M.P.s, both supporters and opposition, were scattered. As 52 of the 187 divisions fell in this category it necessarily played an important part in the all-division maps. The small number of divisions concerning social problems (Category 4) produced a more scattered configuration. On Irish matters (Category 10)

FIG. 4.8.5 Ordinal Scaling of Cohort 1 on Category 5 (Electoral Reform) Divisions

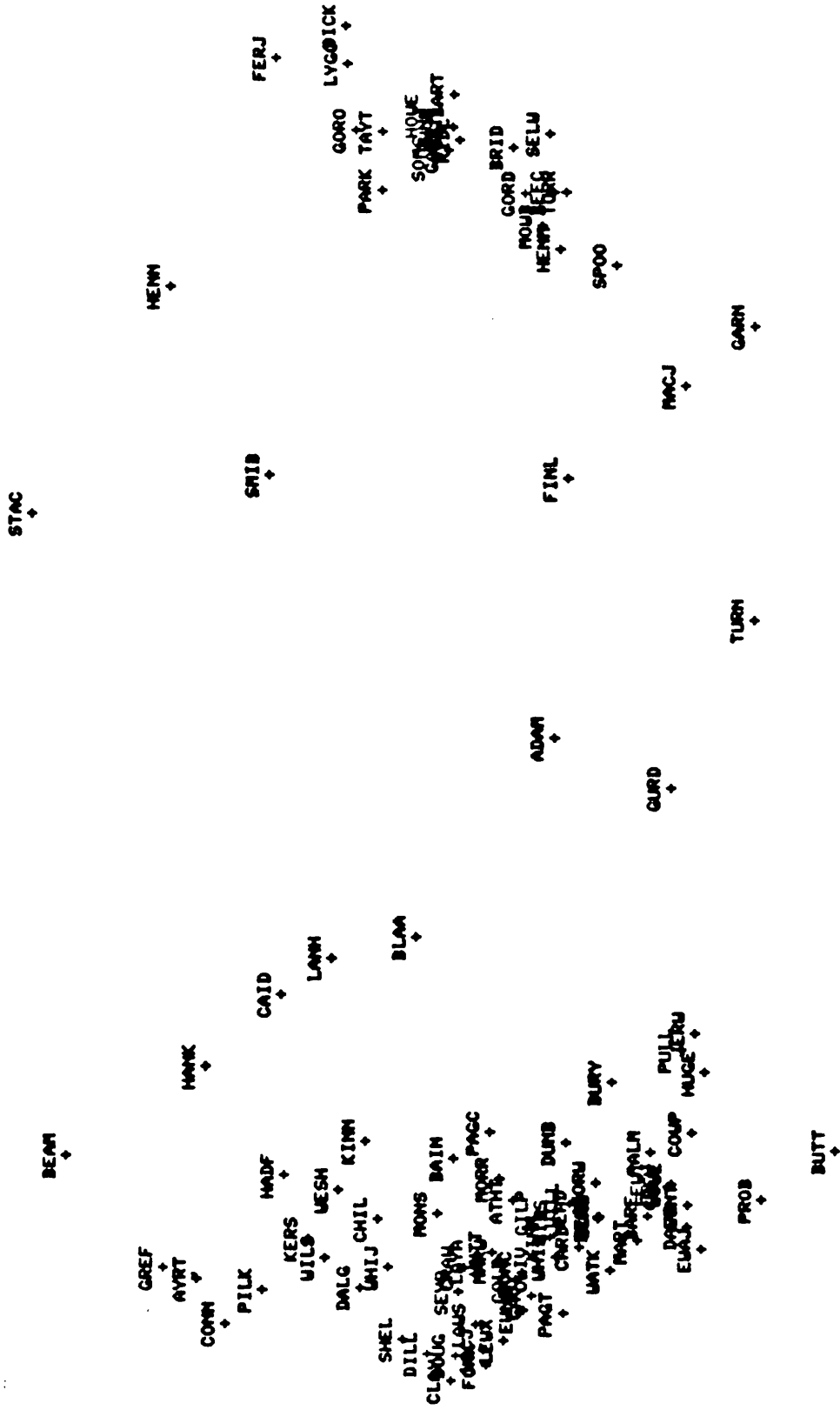


FIG. 4.8.6 Ordinal Scaling of Cohort 1 on Category 7 (Defence) Divisions



liberal members were quite at variance, although some of the dissenters rejoined the cabinet.

These and similar exercises highlighted the contribution of different categories to the map for all divisions.

It would have been an attractive proposition to break down the voting more than we have described, but the number of divisions falling into other categories was a limiting factor. Another consideration had to be the popularity of voting in different categories, for low participation again reduced the scope of the method. However those we have considered may be seen to have highlighted interesting and useful differences in typical voting behaviour, and have contributed to our understanding of the total picture. A complete set of the maps produced has been submitted to the S.S.R.C. Survey Archive to be lodged together with the magnetic tape containing the original data.

#### 4.9 Results Obtained by Least Squares Scaling

Cohort 1 was used to test the effectiveness of least squares scaling when applied to voting data. The test was based on votes recorded in all of the available divisions, and a solution configuration was obtained in two dimensions, for a comparison with our earlier results.

The form of least squares scaling that was used was that suggested by Sammon (1969) in which squared differences between distance and dissimilarity were weighted by the inverses of the dissimilarity values (see Section 1.5). This treatment required a convention for the treatment of zero dissimilarity values, of which there were several. A cut-off value was defined such that all dissimilarities less than 0.01 were assigned weight 100. Just as for ordinal scaling, it was necessary to provide a starting configuration, and this was chosen to be the configuration produced by classical scaling, suitably normalised so that its sum of squared interpoint distances was equal to the sum of squared dissimilarity values. Our implementation provided values of the objective function and its slope for each iteration of the Fletcher-Reeves conjugate gradient minimisation. The algorithm terminated upon satisfying either a maximum iteration or convergence criterion. In this case it was the iteration criterion, but inspection of the printed values indicated that the minimisation was close to convergence. The computer time consumed was then of the same magnitude as that required by the corresponding use of ordinal scaling.

	<u>Objective Function</u>	<u>Slope</u>
After Classical Scaling	$7.637 \times 10^3$	$7.003 \times 10^3$
After Least Squares Scaling	$5.964 \times 10^1$	$3.107 \times 10^1$

The resultant optimal configuration is displayed in Figure 4.9.1.

There were both similarities and differences between the solutions obtained by ordinal and least squares scalings. On a coarse level it was true that nearly every M.P. could be found in the same region in the two configurations. However at a finer level there were differences caused by the introduction of the weightings and stronger assumptions. The interpoint distances had to be better approximations to the smaller dissimilarity values because these had greater weight. Thus to minimise the objective function those M.P.s with little disagreement were drawn together and spurious near neighbours were thrust apart. In terms of the configuration these effects meant that the liberal cabinet was even more compact, and the radical group was more diffuse (compare Figures 4.8.4 and 4.9.1). The impression we gained was that these refinements corresponded to the political cohesion of these groupings. As such the contribution of least squares scaling was helpful in highlighting these effects.

The values that were taken by the Jaccard coefficient were sufficiently spread over the possible range to make least squares scaling a possibility. Classical scaling suggested that two dimensions would account for most of the variation (see Section 4.8) in this case, so the dissimilarity values were expected to be close to linear with configuration distance. Under such conditions the resultant configuration enhanced understanding of the underlying structure. The effect was to concentrate upon small values, accurately positioning nearly identically voting M.P.s, whilst still managing to describe small increases. In summary, the stronger assumptions of least squares scaling made it less prone to degenerate solutions than ordinal scaling.



FIG. 4.9.1 Least Squares Scaling of Cohort 1 on All Divisions



#### 4.10 Economies in Use of Similarities

An obvious advantage would have been gained if the entire set of M.P.s had been able to be scaled together. This would have enabled the comparison of the voting habits of all M.P.s simultaneously. However the corresponding set of similarity values would have been far too large to handle with our computational resources. In some cases many of the values were undefined and scaling then corresponded to using a subset of the similarities. Such subsets consisted of the whole possible range of values, rather than (say) the smallest third. The results of Section 3.4 on random subsets of the similarity matrix were then relevant. However, in other cases the available values were more than could be handled in themselves and so some selection procedure was required. Three were tried.

Firstly, when producing maps of overlapping sets of M.P.s the positions of the M.P.s were matched by Procrustean transformations. This corresponded to considering only similarity values that were located in blocks along the leading diagonal of the complete matrix. It was therefore rather crude, but straightforward. The question arose as to how to treat the overlapping M.P.s, for they had two natural positions. It was decided to attach more weight to the M.P.'s position in the Cohort of higher voters, for it was felt that this would be the more reliable. Thus the procedure was to treat Cohort 1 as the target, fit the overlap with Cohort 2 and add the new M.P.s in Cohort 2 to Cohort 1 after they had experienced the same translation, rotation and scale change. This produced an extended target configuration of 150 M.P.s. Cohort 3 was fitted similarly and so on. There was room for instability in this process, and it was caused by only being able to compare M.P.s far apart in

the voting list by way of a succession of intermediate positions. However, for all its inadequacy, this technique provided help in comparing maps. As an illustration, we give Procrustes statistics for the first three matches based on all divisions.

<u>Target</u>	<u>Fitted</u>	<u>Procrustes</u>
	<u>Configuration</u>	<u>Statistic</u>
Cohort 1 (100)	Cohort 2 (50 overlap)	0.01400
Cohorts 1 & 2 (150)	Cohort 3 (50 overlap)	0.01195
Cohorts 1, 2 & 3 (200)	Cohort 4 (50 overlap)	0.01520

It may also be noted that Procrustes statistics could have been used to compare the same sets of M.P.s for their positions on different sets of divisions. This was not done.

Secondly, and thirdly, maps were produced based on selected subsets of the entire similarity matrix, according to high values (that is, small dissimilarities) and random positions in the matrix, conditional upon a minimum number of values for each M.P. When dealing with large values a good starting configuration was a requirement and the only ones available were generated by Procrustes fitting or the configuration obtained from a random sample. Trials were made with either six or ten as the minimum number of values for each M.P. The existence of solutions for subsets of the M.P.s that were being considered enabled checks to be made against poorly positioned M.P.s. These trials treated a maximum of 250 M.P.s although in principle this could well have been extended. The usefulness of such large maps is dependent upon their accuracy and the ease with which the information can be visualised, and going beyond 250 would have caused problems in this latter respect. The trials did not suggest a significantly different political

interpretation, but they demonstrated the feasibility of this approach. The main advantage that it had over Procrustean fitting was that the number of 'links' required to join each pair of M.P.s was small for the number of 'links' leaving each M.P. This efficiency could have been enhanced by stipulating a design for the choice of similarities used. As it was, it enabled all of the scaling to be completed in one run rather than several.

#### 4.11 Conclusions

Perhaps the main conclusion to emerge from this study is that ordinal scaling is well adapted to dealing with the type of data provided in division list analysis. Several features that were common to the maps enable us to make this claim. Firstly we were able to produce interesting and illuminating interpretations of the data that could be appreciated in a straightforward, visual manner. In particular the availability of coloured computer graphics helped to highlight the positionings of M.P.s of different party loyalties. The scaling maps we submitted were all two-dimensional, and this corresponded to our finding that the patterns underlying the dissimilarities could often be represented adequately in that way, whilst one-dimensional solutions missed some structure. Secondly the success of scaling was manifest at different levels of similarity value. Thus in any one map we found it straightforward to identify groups of M.P.s, whose members voted differently from those of other groups, but at the same time minor differences within those groups were also distinguishable. It was particularly easy to pick out those individual M.P.s whose pattern of voting behaviour was aberrant, for it seemed that multidimensional scaling was well suited to the isolation of eccentric or unexpected behaviour. It was possible to demonstrate the activity of sub-groups of voters, in this case the various forms of radical liberals. Thirdly the technique enabled us to compare with ease the maps produced by scaling equivalent sets of M.P.s on different sets of divisions. This process identified marked changes in the composition of voting groups when different issues were at stake. We have illustrated this by considering the behaviour of the first one hundred M.P.s on all divisions, electoral reform and defence divisions.

We were able to cope with one feature of the House of Commons' division data that was unusual in scaling applications, namely sheer size. In many fields of research, 40 objects would have been regarded as providing a large problem. Here, we have been confronted with 662, or at least 530, that number which corresponded to M.P.s who voted 'sufficiently often to be interesting'. We tackled this problem by producing maps of sets of 100 M.P.s which overlapped and could thus be compared. While 100 was not by any means an upper limit, it would certainly have been very hard to go beyond 150-200, at least with a full system of similarities.

We have seen the limited usefulness of single-link clustering, which gave an indication of the group identities without spelling out the inter-personal voting relationships. We have demonstrated the feasibility of applying least squares scaling to this type of data, formed as it was according to a nearly euclidean dissimilarity function. The results represented minor but helpful modifications of the ordinal scaling configuration. We have shown the advantages of a thorough attempt to correct the original punched data and to undertake several routine analyses of the data. Additional refinements have been the comparison of overlapping sets of M.P.s by Procrustes rotation, and the occasional use of less than the entire system of similarities.

Much can be derived from cluster and scaling analyses of voting data. It is to be hoped that their use in this area will become widespread.

#### 4.12 References

- ANDERSON, L. F., WATTS, M. W. and WILCOX, A. R. (1966).  
*Legislative Roll Call Analysis*. Evanston.
- AYDELOTTE, W. O. (1963). Voting patterns in the British House of Commons in the 1840's.  
*Comparative Studies in Society and History*, 5, pp. 134-163.
- AYDELOTTE, W. O. (1966). Parties and issues in early Victorian England.  
*Journal of British Studies*, 5, pp. 95-114.
- AYDELOTTE, W. O. (1972). The disintegration of the Conservative Party in the 1840's: a study of political attitudes.  
In *The Dimensions of Quantitative Research in History* (Aydelotte, W. O., Bogue, A. G. and Fogel, R. W. : Eds.)  
London: Oxford University Press, pp. 319-346.
- AYDELOTTE, W. O. (1977). Constituency influence in the British House of Commons, 1841-47.  
In *The History of Parliamentary Behaviour* (Aydelotte, W. O.: Ed.)  
Princeton: Princeton University Press.
- BERINGER, R. (1978). *Historical Analysis: Contemporary Approaches to Clio's Craft*. New York: Chichester: Wiley.
- BERRINGTON, H. (1968). Partisanship and dissidence in the Nineteenth-Century House of Commons.  
*Parliamentary Affairs*, 21, pp. 338-374.
- CROMWELL, V. (1980). *Computer Analysis by Multidimensional Scaling of House of Commons' Division Lists (1861)*.  
End of grant report to the Social Science Research Council.
- DOD (Charles, Roger, Phipps) (1861,1862). *The Parliamentary Companion*.
- HARTIGAN, J. A. (1972). Direct clustering of a data matrix.  
*Journal of the American Statistical Association*, 67, pp. 123-129.
- HARTIGAN, J. A. (1974). Block voting in the United Nations.  
In *Exploring Data Analysis: the Computer Revolution in Statistics* (Dixon, W. J. and Nicholson, W. L.: Eds.) London.
- HARTIGAN, J. A. (1975). *Clustering Algorithms*.  
New York: Wiley (esp. chapters 14 & 16)
- HATZENBUEHLER, R.L. (1972). Party unity and the decision for war in the House of Representatives, 1812. *William & Mary Quarterly* Ser. 3, 29, pp.371-390.
- HEYCK, T. W. (1974). *The Dimensions of British Radicalism*.  
Urbana: University of Illinois Press.
- HEYCK, T. W. and KLECKA, W. (1973). British Radical M.P.s 1874-1895: New evidence from discriminant analysis.  
*Journal of Interdisciplinary History*, 4, pp. 161-184.

- LOVEDAY, P. (1975). *Parties in Papua New Guinea 1972-1974*.  
*Institute of Development Studies Discussion Paper No. 82*.  
Brighton, 1975.
- LOVEDAY, P., MARTIN, A. W. and PARKER, R. S. (Eds.) (1977).  
*The Emergence of the Australian Party System*.  
Sydney: Hale and Iremonger.
- MONYPENNY, W. F. and BUCKLE, G. E. (1916).  
*Life of Benjamin Disraeli*. London: John Murray. 6 vols.
- MOWBRAY, Rt. Hon. J. R. (1900). *Seventy Years of Westminster*.  
London: Blackwood.
- SAMMON, J. W. (1969). A nonlinear mapping for data structure analysis.  
*I.E.E.E. Trans on Computers*, 18, pp. 401-409.
- VEITCH, L. G. and JAENSCH, D. H. (1974). A procedure for the analysis  
of Colonial Australasian legislatures.  
*Political Science New Zealand*, 26, pp. 12-27.
- WAHLKE, J. C. and EULAU, H.(Eds.) (1959). *Legislative Behaviour:  
a Reader in Theory and Research*. Glencoe, Illinois: Free Press.



C H A P T E R F I V E

AN APPLICATION IN LINGUISTICS AND ETHNOLOGY

	<u>PAGE</u>
5.1 Introduction .. .. .	209
5.2 Materials and Methods .. .. .	210
5.3 Results .. .. .	219
5.4 Conclusions .. .. .	229
5.5 References .. .. .	232

## 5.1 Introduction

This chapter describes an application of multidimensional scaling and related techniques in the field of linguistics. The project was undertaken in collaboration with Mr. Andrew Baring, an anthropologist and ethnologist, who has a particular interest in the central areas of Sudan.

It was in the course of writing a history of the central Sudan that Mr. Baring became dissatisfied with the classification of African languages proposed by Greenberg (1978), and generally accepted by ethnologists. In plotting the geographical distribution of 300 languages ranging from North-east Nigeria to West Sudan according to Greenberg's classification, it seemed to him that the resultant groups were not related to ecology, land formation or even likely history. These groups had been obtained by lexical and grammatical comparison of the languages. Phonetics had not been used. As with all such comparisons a large amount of hard work is needed to establish similarities. In order to provide a quick test of Greenberg's hypothesis, a simple phonetic measure of similarity between languages was devised, that could be made from a dictionary or word list. This used just the first sound in each word, and could be speedily obtained from an alphabetically-arranged list. To his surprise there seemed to be similarity in the distribution of first sounds for related languages. A pilot study of twelve languages accorded well with lexical and grammatical patterns. In order to extend this further, the University of Bath was approached and this project was conceived. By its completion over 550 languages had been analysed.

In (5.2) we discuss the materials and methods used; (5.3) contains some sample results and these are discussed; (5.4) contains conclusions.

## 5.2 Materials and Methods

### Language Groups

The 554 languages and dialects studied were partitioned into 26 families as shown in Table 5.2.1. For most of the languages only one word list, vocabulary or dictionary was available. However for others several such sources were easily obtained. Altogether well over a thousand sources were used, varying in size from lists of just a few hundred words to comprehensive modern dictionaries. Many of the languages have long ceased to be spoken and are used to help understand the evolutionary development. The families were formed according to traditional classifications. The original collection of languages was largely African. When this was seen to produce interesting results the range was extended to include the better understood Indo-European languages in order to test further this phonetic method.

### Phonetic Groups

Traditional phonetic studies of language consider the phonetic structure of entire words. They are laborious and time consuming to conduct. In restricting attention to just the first sound this process is made much simpler because the number of sounds recorded is reduced and, above all, advantage can be made of the ordering of words in dictionaries according to this very criterion. The hypothesis is that similar languages will have similar proportions of words of each phonetic type. Two immediate problems arise. Firstly there are well established processes of sound shift that may be frequently observed in cognate languages. Thus words beginning with F in one language may start with V in another (Father, Vater etc.). Many other possible shifts exist. To avoid missing such similarities between

TABLE 5.2.1

The Language Groups

<u>Group No.</u>	<u>Description</u>	<u>No. of 'Languages'</u>
1	West Atlantic (African) )	15
	)	
2	Voltaic )	17
	)	
3	Kwa ) Greenberg's Niger-Congo	19
	)	
4	Eastern Adamawa )	11
	)	
5	Bantu )	38
6	Sudanic	40
7	Semitic )	20
	)	
8	Ethiopian-Semitic )	10
	)	
9	Cushitic ) Greenberg's Afro-Asiatic	23
	)	
10	Chadic )	11
	)	
11	Berber & Ancient Egyptian )	9
12	Iranian	44
13	Indian	24
14	Dravidian & Munda	20
15	Sino-Tibetan/Mon-Khmer	28
16	Malayo-Polynesian	25
17	Asiatic	29
18	Uralic	22
19	Miscellaneous (mostly other Indo-European)	17
20	Slavic	12
21	North European (except English)	22
22	English	16
23	Celtic	8
24	Romance	23
25	Eastern Sudanic	26
26	Gurage & Neighbours	25
TOTAL		554

the overall sound of a pair of languages, groups of sounds were formed. Their profiles across the nine groups (Table 5.2.2.) were compared rather than across the twenty-six letters of the alphabet. The groups were chosen according to known phonetic structure in order to maximise the interchange that might occur within groups, and minimise the interchange between groups. Phonetic theory and a little experience quickly established the final group formation. Secondly some languages abound in commonly used prefixes, others form their plurals at the beginning of the word. Where there was evidence of this, efforts were made to ensure that only the root word was counted.

#### Procedure with a Dictionary

When a dictionary, vocabulary or list had less than about 2,000 entries a complete census of the (unprefixed, unpluralled) words was carried out. Beyond that level it was found that the results were scarcely changed by counting the number of half-pages occupied by words with any particular starting letter. Care was taken with the three forms of 'C' and two of 'X'. The resulting estimate of the word profile was expressed as the percentage in each of the nine phonetic groups. When there were several sources for each language the final values were averages of the different sources, weighted for size of list. Some sample figures are presented in Table 5.2.3 which describes one language selected from every other language group as set out in Table 5.2.1. The crudeness of the technique did not warrant the recording of percentages to greater than integral accuracy.

TABLE 5.2.2.

The Phonetic Groupings

Group

<u>No.</u>	<u>Letters</u>	<u>Description</u>
1	A, E, I, Y	front and palatal vowels
2	B, P, F, V	voiced and unvoiced labial stops and dentals
3	Ch, J	palatals
4	C (hard), K, Q, G, H, X (sometimes)	gutteral fricatives, laryngeal stops
5	L, R	liquids
6	M, N	nasals
7	Z, S, Sh, C (soft), X (sometimes)	sibilants
8	T, D	dentals
9	O, U, W	rounded back vowels

TABLE 5.2.3.

Some Sample Data

		<u>Phonetic Group Number</u>								
<u>Language</u>		1	2	3	4	5	6	7	8	9
<u>Language</u>	<u>Group No.</u>									
Fulani	1	9	16	8	18	8	11	8	14	8
Yoruba	3	39	11	2	9	5	5	6	8	14
Kikuyu	5	12	2	6	25	9	24	0	16	6
Hebrew	7	16	9	0	30	9	15	16	6	0
Beja	9	13	11	1	29	5	11	11	14	3
Coptic	11	7	7	12	19	5	12	19	9	9
Sanskrit	13	15	23	2	17	5	12	11	11	4
Vietnamese	15	2	12	9	18	9	13	7	28	2
Japanese	17	13	6	7	27	2	12	13	12	7
Basque	19	32	12	3	14	5	7	12	4	9
Danish	21	11	24	0	14	8	7	16	10	10
Welsh	23	19	13	1	33	8	7	3	15	1
Jur	25	17	10	12	15	10	7	0	13	16

### Stability Within Languages

A question of immediate concern was whether two different dictionaries of the same language would produce similar phonetic profiles. Dictionaries of English abound and this was the obvious first test case. Fifteen forms and dialects of English are shown in Table 5.2.4, some from quite obscure sources. The three modern 'languages' used (English, Modern English and American English) show very much the same pattern, with a largest discrepancy of two percent between any of the twelve contributory dictionaries. The others show a marked difference in the first component between mainstream and dialect English. Evidence of a time trend may also be seen (Old Icelandic, Anglo Saxon and Medieval English are low on Phonetic Group 2 and high on Group 9). These effects showed clearly in the scaling configuration. These findings were taken as a justification of the underlying idea of using first sounds, and of the choice of phonetic groups. However English was unlikely to be a representative case, so similar exercises were conducted by gathering as many dictionaries as possible of Hausa, Hebrew and Chinese. Once more there seemed considerable stability of the profile between dictionaries. The most variable language encountered was Etruscan and this was not surprising since it is not properly understood. Compilers of Etruscan word lists tended to emphasise different aspects of the language.

### Methods

A list of the methods employed is provided in Table 5.2.5. Once a combination of languages had been selected and gathered in a computer file, simple commands were provided, sufficient to perform the required combination of techniques. Most of these techniques were introduced



TABLE 5.2.4

The Varieties of 'English'

<u>Variety</u>	<u>Phonetic Group Number</u>								
	1	2	3	4	5	6	7	8	9
American English	14	21	2	17	8	8	13	11	7
Anglo Saxon	11	16	1	20	7	7	13	11	14
Australian English	5	23	3	23	8	9	11	10	8
Black English	5	24	5	21	7	6	14	11	6
English	14	20	2	16	9	7	14	11	8
Hobson-Jobson	6	22	7	22	5	11	12	11	3
Jamaican English	6	26	6	20	6	9	10	10	7
Medieval English	13	19	2	17	9	7	12	10	10
Modern English	12	22	3	17	8	8	12	11	6
Obsolete English	16	22	3	17	8	6	11	11	6
Old Icelandic	13	14	0	20	8	7	15	10	13
South African English	8	23	1	20	7	10	14	11	6
Scottish English	6	20	2	21	8	6	17	12	8
Somerset English	6	23	2	21	9	7	14	11	6
Sussex English	6	19	4	21	10	6	18	11	5

TABLE 5.2.5

The Set of Techniques Used

1. Partition likelihood clustering (standard search).
2. Partition likelihood clustering (extended search).
3. Formation of similarity matrix (euclidean distance or information radius).
4. Single-link clustering, from (3).
5. Principal component analysis.
6. Two-dimensional ordinal scaling from (3) and (5).
7. Three-dimensional ordinal scaling from (3) and (5).
8. Display of similarities against configuration distances for (6) and (7).
9. Configuration from (6) plotted with partition likelihood clusters  
superimposed.
10. Configuration from (6) plotted with single-link clusters superimposed.
11. Configuration from (6) plotted with colours for different language groups.
12. Configuration from (7) as a pair of perspective plots.
13. Histograms of frequency of phonetic group values for a set of languages.

in the first chapter. A few comments about their use in this context should be sufficient.

The more extensive search for the optimal partition in the partition likelihood clustering method was not often used. Its only occasional refinements made it difficult to justify the extra computing resources for a moderately large (greater than fifty) number of languages.

Similarities were usually formed by euclidean distance between profiles. The resulting scaling plots were quite similar to those obtained from the information radius when both were used. The anthropologist found the similarity values from euclidean distance of interest in themselves.

The addition of clusters to configuration plots was done by hand. The three-dimensional perspective plots were only used for small numbers of languages, otherwise the effect was lost in the overall confusion. Simple histograms of the frequency distribution of specific values for the phonetic groups were quite useful in highlighting outlying behaviour and typical structure for any language family.

The relative success of these techniques and the anthropologist's assessment are discussed in the context of two examples that follow in the next section. The first reproduced some known structure between languages, the second formed part of the evidence suggesting a new hypothesis about a pre-Indo-European mediterranean source, akin to Afro-Asiatic.

### 5.3 Results

#### Dialects of Gurage

Gurage is a region of southern Ethiopia, north-west of Lake Zway (Fig. 5.3.1), south-west of Addis Ababa. Its people derive from Sidamo, Tigré and Harar, surrounding areas. For centuries before the conquest by Ethiopia in 1875 it had links with that country even though it consisted of independent tribal units. In consequence the dominant language, Gurage, is Ethiopian-Semitic in nature. However this could be described as a dialect cluster with three main groups. The first, eastern Gurage, contains Selti, Wolane, Ulbarag, Inneqor and Zway, all related to Harari. The five dialects certainly in western Gurage are Chaha, Eza, Ennemor, Endegen and Gyeto. Muher, Masqan and Gogot are sometimes linked with this group. Soddo and Aymellel, which are related to Gafat, constitute northern Gurage. The only literature in Gurage is a Chaha catechism written in Ethiopic characters. The vocabularies contain a number of Sidamo words, reflecting the peoples' earlier migration. The unity of the whole group is still open to doubt, as east and west dialects are largely mutually unintelligible.

Leslau (1979) has published a set of individual dictionaries treating twelve of the Gurage dialects (Table 5.3.2). He has used the same word lists for each. This comparability, allied with the advantages conferred by a standard hearing and writing of the dialects, allows a useful test of the sensitivity of the leading sound hypothesis. Firstly we consider the twelve dialects themselves, and then we introduce other representatives of the Ethiopian-Semitic and Chadic families for comparison.

Fig. 5.3.3. contains the two-dimensional configuration

Fig 5.3.1 The Geographical Distribution of Gurage Dialects.

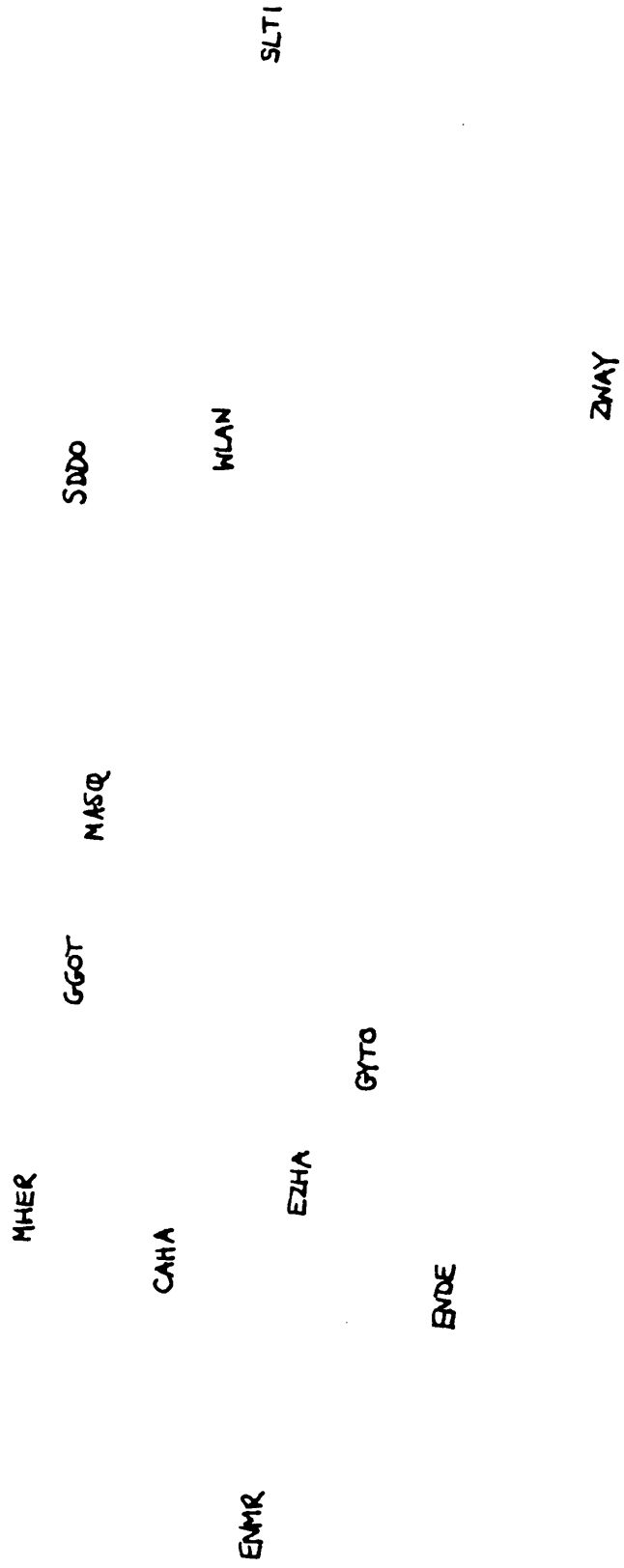


TABLE 5.3.2

The Dialects of Gurage

<u>Dialect</u>		<u>Phonetic Group Number</u>								
		1	2	3	4	5	6	7	8	9
Chaha	(W)	22	11	6	21	1	11	10	11	7
Endegen	(W)	21	11	5	21	1	12	10	10	9
Ennemor	(W)	23	11	5	20	1	11	10	10	9
Eza	(W)	21	11	6	21	1	11	10	11	8
Gyeto	(W)	20	11	6	21	1	11	11	11	8
Gogot	(?)	21	11	5	20	3	11	10	12	7
Masqan	(?)	20	11	5	21	3	10	11	12	6
Muher	(?)	22	12	6	21	2	10	11	11	7
Soddo	(N)	19	9	4	22	3	12	12	12	6
Selti	(E)	16	11	5	24	4	11	12	12	7
Wolane	(E)	18	10	5	22	4	12	11	11	6
Zway	(E)	17	11	4	20	4	12	12	12	8

Fig 5.3.3 Ordinal scaling of Gurage dialects.



produced by multidimensional scaling. The final stress was 3.2%. The principal component analysis (equivalently classical scaling) used to derive the initial configuration indicated that the variation was most significant in the first phonetic group, with more moderate contributions coming from the fourth, fifth, seventh and ninth groups. Inspection of the original figures confirms that western Gurage dialects are higher in groups one and nine, lower in groups four, five and seven. This is reflected in the final configuration which has a dominant east-west dimension. Indeed the correspondence between Fig. 5.3.1 and Fig. 5.3.3. is quite striking. Single-link clustering was not so striking. One large group emerged from the western dialects, swallowing others on its way. Thus the final three clusters were a group of ten dialects, with Zway and Selti by themselves. In contrast both search algorithms for the partition likelihood clustering technique identified the last three clusters as:-

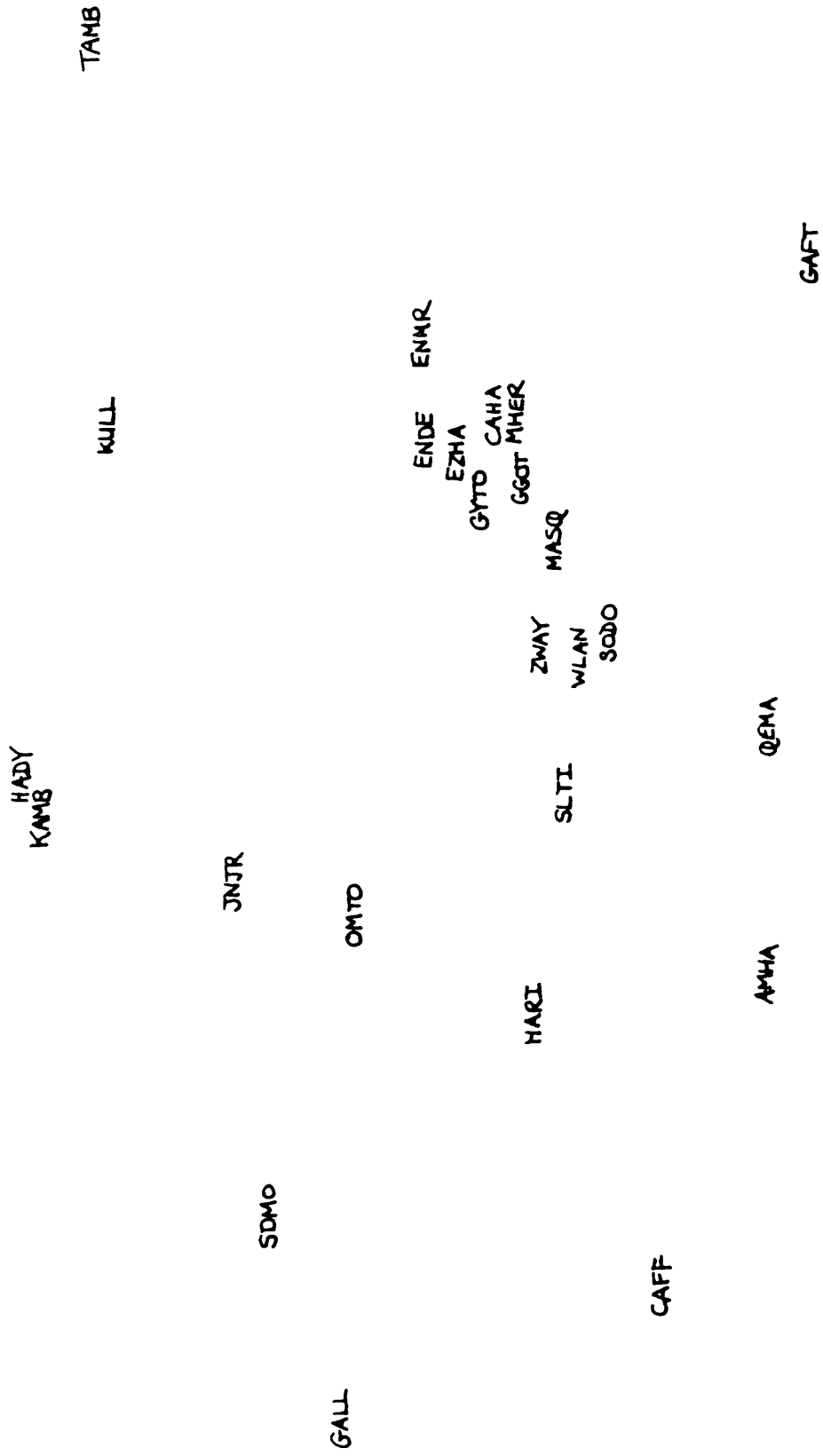
- (i) The five western dialects
- (ii) The three unknown dialects
- (iii) The four northern and eastern dialects (themselves separated at an earlier stage).

At the penultimate stage Masqan and Gogot were joined with the eastern group, Muher with the western. This may be seen to correspond better with the geography of Fig. 5.3.1 than with traditional attempts at classification.

Fig. 5.3.4 contains the two-dimensional configuration produced by multidimensional scaling when thirteen neighbouring languages from Ethiopian-Semitic and Chadic were added. Some of these are located in Fig. 5.3.1. The final stress was 8.8% (4.2% in three dimensions). The Gurage dialects may be seen to possess the same inter-relationships as when studied alone. Gafat is closest to the northern dialects,



Fig 5.3.4 Ordinal Scaling of Garage  
Dialects and Surrounding Languages



as expected. Harari and Amharic are closest to the eastern group. Sidamo, Galla, Kambata and Hadiyya, all to the south geographically, are quite distinct as a group but closest to the eastern dialects. Gangeru and Tembaro are also distinct, partition likelihood clustering putting them closest to the western group. Again geography matches the scaling configuration well: Gurage has been placed in its context. Principal component analysis demonstrates the importance of the first and fourth phonetic groups. Single-link clustering again evolves as one dominating cluster centred on Gurage. Partition likelihood clustering suggests the existence of a north and eastern Gurage group related to Ethiopian-Semitic and surrounding eastern Chadic languages, and a western Gurage group related to surrounding western Chadic languages.

The Gurage dialects seem related as a group, quite distinct in themselves. The overall impression is of the close relation between dialect and geographical proximity. Yet this has been obtained from a supply of dictionaries showing little variation in phonetic structure. These results encourage extensions to larger regions and time spans, where no such clear interpretation is available. We now turn to a much larger set of dictionaries with little prior indication as to the structure that would emerge.

#### Language around the Mediterranean

The following exercise in multidimensional scaling is based upon a set of 100 languages selected from the families that would be involved in the evolution of language around the Mediterranean. They come from the following families:-

A	Sudanic (19)	F	Berber & Ancient Egyptian (7)
B	Semitic (15)	G	Uralic (3)
C	Ethiopian-Semitic (7)	H	Miscellaneous Indo-European (11)
D	Cushitic (7)	I	Celtic (6)
E	Chadic (3)	J	Romance (22)

Two- and three- dimensional solutions (stress values 16.4 and 9.3 respectively) were obtained following principal component analysis, as before. Percentages of the trace corresponding to successive dimensions were 42, 19, 14, 11, 6, 4, 3, 2, 0 respectively. Good convergence was then obtained. The two-dimensional solution is plotted in Fig. 5.3.5. The first character of each plotted symbol indicates the family as above, the remaining three characters provide an index number. Each of the ten families may be identified in a restricted location of the solution. To the left there is a compact cluster of Romance languages centred on Latin (J-15) and French (J--6). Below these are the Miscellaneous Indo-European languages, mainly Greek, classical and modern. At the bottom are the Sudanic languages. At the top on the right hand side are Berber and Ancient Egyptian languages. In between lie the other African languages: Chadic centrally; Semitic nearer to Berber; Ethiopian-Semitic and Cushitic nearer to Sudanic. Uralic lies between Romance and Berber/Egyptian, Celtic slightly more central. There are a few exceptions, but it is possible to draw a line through the plot that separates northern Mediterranean from southern with just five exceptions.

These exceptions are Old Breton (I-23), Cornish (I-25), Welsh (I-27), Akkadian (B-39) and Sumerian (H-99). The first three are quite separate from the other Celtic languages including Irish (I100) and Gaelic (I-26). This corresponds to the usual division of Celtic into Goidelic and Brithonic. Akkadian is normally considered to be Semitic, but possibly Indo-European. Its position



here is quite anomalous and may reflect a poor source. It is a characteristic of the least understood languages that their dictionaries and word lists are most variable. This could also apply to Sumerian, Hittite (H-34) and Etruscan (H-30). Albanian (H-28) is surprisingly placed, but quite near the Uralic group.

Despite these exceptions the groups are quite well defined. Conventional classification of languages according to Greenberg does not propose any hierarchical structure for the families it defines. This analysis suggests some similarities that could be used in such an exercise. Generally the degree of clarity of the relationships is greater for European sources, then Northern African and finally other African languages. The anthropologist concerned found evidence from this and similar maps for the existence of a pre-Indo-European source around the Mediterranean and for the similarity of many Sudanic and Semitic languages that correspond to Greenberg's Saharan and Afro-Asiatic families.

#### 5.4 Conclusions

The rudimentary form of measurement used to assess each language could not be expected to yield results as convincing as more elaborate lexical and grammatical comparisons. The unsophisticated idea demands scepticism in the interpretation of results. However empirically the scaling plots conformed to accepted patterns when well understood groups of languages were studied. English and Gurage were two small, successful examples. The patterns for European languages seemed sensible, agreeing with established evolutionary theories. But recognisable success in describing African languages is more elusive for two major reasons. Firstly there is a far inferior body of dictionaries to use. These tend to be limited word lists, often compiled by amateur linguists who would hear and write language quite differently from one another. Thus comparability is less often achieved. Secondly there is a far less coherent picture of the history and development of the African languages against which to gauge the success of this method. It would be possible to conclude that better results were obtained when dealing with languages that were quite similar. This could reflect the greater knowledge of their development, or it could be that the measures are only really comparable locally.

Not all languages seemed to fit into the expected patterns. There were several anomalous positionings which did not seem to be caused by sound shifts or prefixes. For example, Welsh and Bini (Kwa) were far removed from any other languages (and each other!). Bini has a very large number of words in phonetic Group 1, starting with vowels. The tendency is then to find the remainder of the languages compressed in the plot. These sorts of problems emphasise that this approach provides no substitute for careful lexical comparison of languages,

a technique that is becoming within the range of a computer-based treatment.

Assessment of the different statistical techniques that were applied to these data must take into account their usefulness for (and appeal to) the consumer, here the anthropologist. Single-link clustering was found to be useful in principle, but difficult to interpret by itself and highly prone to chaining with these data. By contrast, multidimensional scaling plots were easy to understand and stress values gave some indication of success of fit, although they were often quite high. The extra dimensions that would reduce the stress were difficult to present, although experiments with perspective views of small three-dimensional configurations and knitting-needle models of larger ones assisted a little. Consistently the most popular technique was partition likelihood clustering which was regarded as particularly accurate. Loglikelihood plotted against number of clusters helped to assess the minimum sensible number of clusters. (See Chapter 6 for an example of this). It was felt that the most useful combination of techniques was two-dimensional multidimensional scaling for its visual presentation and partition likelihood clustering for its accuracy. A popular initial approach for a particular language was just to provide an ordered list of its dissimilarities with others, thereby highlighting its nearest neighbours. Another initial display was a set of nine histograms for each language family, intended to demonstrate the typical values taken by the family across the phonetic groups. The suggestion that each family should have its own distinct pattern prompted the use of an average set of figures, representing the family, in some of the analyses.

Despite its lack of sophistication, this phonetic measure of language has several benefits. It is simple to calculate and

therefore allows consideration of a large set of sources. It is objective in that different researchers would obtain the same measure from the same source, without having to decide subjectively whether two words for the same concept were cognate. It allows for the descent, growth, formation and convergence of languages quite naturally, in that the state of the language is reflected in the dictionary. It permits direct statistical analysis, the clarity of which has been demonstrated. Extensions could include analyses, as suggested in Chapter 4, that would allow comparability between plots through common languages so that a set much greater than one hundred in number could be considered. Subsets of a large dissimilarity matrix could also be used.

Only a few results have been presented here. It is intended to publish a much larger selection in "A Phonetic Experiment in Linguistic Classification", a work being prepared by Andrew Baring which will give a complete bibliography of the dictionaries that have been used.



5.6 References

GREENBERG, J. H. (1978). *Universals of Human Language (4 volumes)*.  
Stanford University Press: Stanford, California.

LESLAU, W. (1979). *Etymological Dictionary of Gurage (Ethiopia)*.  
University of London Press: London.

C H A P T E R   S I X

AN APPLICATION IN TAXONOMY

	<u>PAGE</u>
6.1    Introduction            .. .. .	234
6.2    Preliminaries: Selection and Measurement of Specimens .. .. .	239
6.3    The Measurement of Similarity    .. .. .	245
6.4    Results Obtained from Single-Link Clustering    ..	250
6.5    Results Obtained from Ordinal Scaling            .. ..	252
6.6    Results Obtained by Partition Likelihood Clustering	255
6.7    Discussion and Conclusions    .. .. .	263
6.8    References            .. .. .	265

## 6.1 Introduction

The genus *Colisa* was first named by Cuvier and Valenciennes in 1831. It consists of five species that inhabit the Indian subcontinent including Burma (see Figure 6.1.1). Additionally these species are kept as aquarium fish, for they are easy to obtain, to maintain even in poor and overcrowded conditions, and easy to breed. Table 6.1.2 provides details of the five species including their natural location. Henceforth we shall refer to them by the abbreviated forms, Species A through to Species E.

An evolutionary theory for the development of these species has been suggested by various authors, including Liem (1963) and Dawes (1978). For some time geologists (e.g. Pascoe, 1920) have advocated the existence of an ancient river, the Indobrahm, that ran east to west across the north of India and connected the headwaters of the Indus, Brahmaputra and Irrawaddy. The evidence for this is based upon alluvial deposits, but other biological support exists, including the existence of river dolphins, common to more than one river. If this river existed then the river systems mentioned in Fig. 6.1.1 and Table 6.1.2 would have been confluent. The genus *Colisa* belongs to the family Anabantidae, and Anabantids would have existed at that time (Sanders, 1934). However lowering water levels would have separated the rivers, isolating populations of fish and allowing visible changes to occur in them as time progressed. These changes became sufficiently marked to allow the classification into different species that has been described in Table 6.1.2. A factor that could have accelerated this process would have been the different strengths of water flow in the various rivers.

Several attempts have been made to establish the relationship

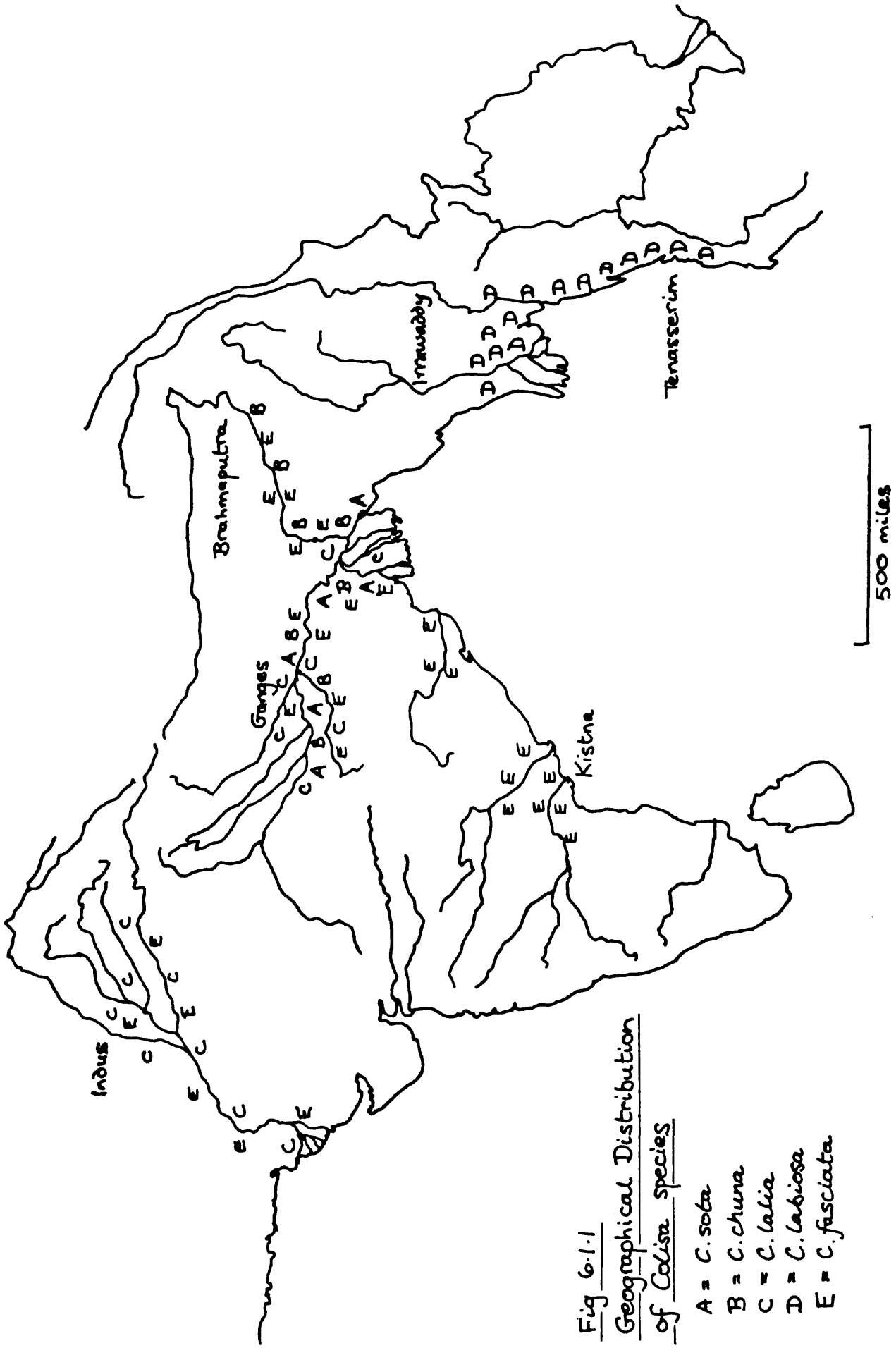


Fig 6.1.1  
Geographical Distribution  
of *Calisa* species

- A = *C. soba*
- B = *C. chuna*
- C = *C. talia*
- D = *C. labiosa*
- E = *C. fasciata*

TABLE 6.1.1.2

<u>Colisa Species</u>	<u>When Named</u>	<u>Named By</u>	<u>Popular Name</u>	<u>Location</u>
Species A = <i>C. sota</i>	1822	(Hamilton-Buchanan)*	_____	R. Ganges
Species B = <i>C. chuna</i>	1822	(Hamilton-Buchanan)	Honey Gourami	R. Ganges and R. Brahmaputra
Species C = <i>C. lalia</i>	1822	(Hamilton-Buchanan)	Dwarf Gourami	R. Ganges and R. Indus
Species D = <i>C. labiosa</i>	1878	(Day)	Thick-lipped Gourami	R. Irrawaddy and R. Tenasserim
Species E = <i>C. fasciata</i>	1801	(Bloch and Schneider)	Giant or Striped Gourami	R. Indus, R. Ganges, R. Brahmaputra and R. Kistra

Bracketing indicates that the species has been subsequently renamed.

*Trichopodus* and *Trichogaster* were the previously used genus names.

\* Details obtained from Tate-Regan (1909)

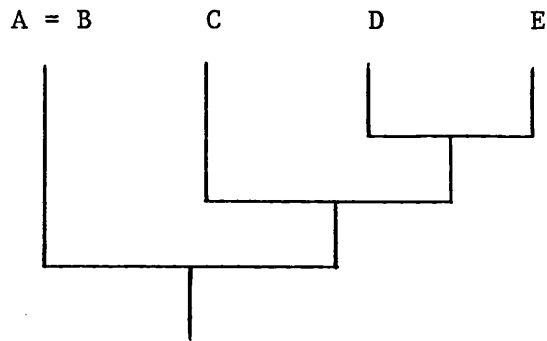
between various Anabantid genera and species. These have been based upon a study of skeletal features (Liem, 1963) or behavioural factors (Vierke, 1975). However no statistical studies have been attempted. The advantages conferred by a statistical analysis of morphology would include the ability to examine preserved specimens, which often have to form a large component of a study, and the larger number of objects and variables that could be treated. The important discriminatory variables could then be determined.

Two pairs of species of *Colisa* seemed particularly closely related. It has been suggested by John Dawes that Species A and Species B may be one and the same. Males of Species B have a completely different colouring during breeding, and it is approximately in this form that the nine specimens of the alleged Species A are to be found in the Natural History section of the British Museum. Further evidence for the existence of just one species is provided by Day (1878) who collected these nine specimens and regarded them as just a variety of Species B. Recent catalogues of Indian fish only include the one species (Sen, 1978; Sheri and Saied, 1975).

The other species that seem closely related are Species D and Species E. The differences between them are quite small. Species D has a more pointed posterior region in the dorsal fin with a corresponding greater surface area providing locomotory power. Small colour quality differences also exist.

Extra tests for correct classification may be provided by hybridisation experiments, which are easily set up with *Colisa*. True species will not hybridise, or if they do the hybrids are sterile or only partially viable. Species B stubbornly refuses to breed with any other, even with Species C which is of similar size.

Species A is not therefore a B/C hybrid. Hybrids have been produced between Species C and Species D and E. However these have always been male and sterile (Pinter (1960); Dawes (1978) ). Fertile hybrids of both sexes have only been produced between Species D and Species E. Their patterns of courtship are very similar, as are their other breeding habits. For example, whilst Species B has its own courtship display akin to standing on its tail, Species C, D and E produce a fine mist of bubbles under the nest from their gills as they shake their heads. At most two such shakes have been seen in Species C, but Species D and E average four to five. Heartbeat rates in embryos and young fry also support the similarity of D and E and differences with C. The overall suggestion that we test is summarised by the tree diagram in Figure 6.1.3.



(6.1.3)

Morphology has always formed a fundamental part of classification. In this chapter we investigate the above assertions using scaling and clustering techniques based upon morphological parameters. This enables one hundred and fifty six preserved specimens to be considered.

## 6.2 Preliminaries: Selection and Measurement of Specimens

One hundred and fifty six specimens have been analysed by John Dawes for this study. For each one the following procedure was adopted.

A. An X-ray of the specimen was taken.

B. A colour or black-and-white photograph was taken.

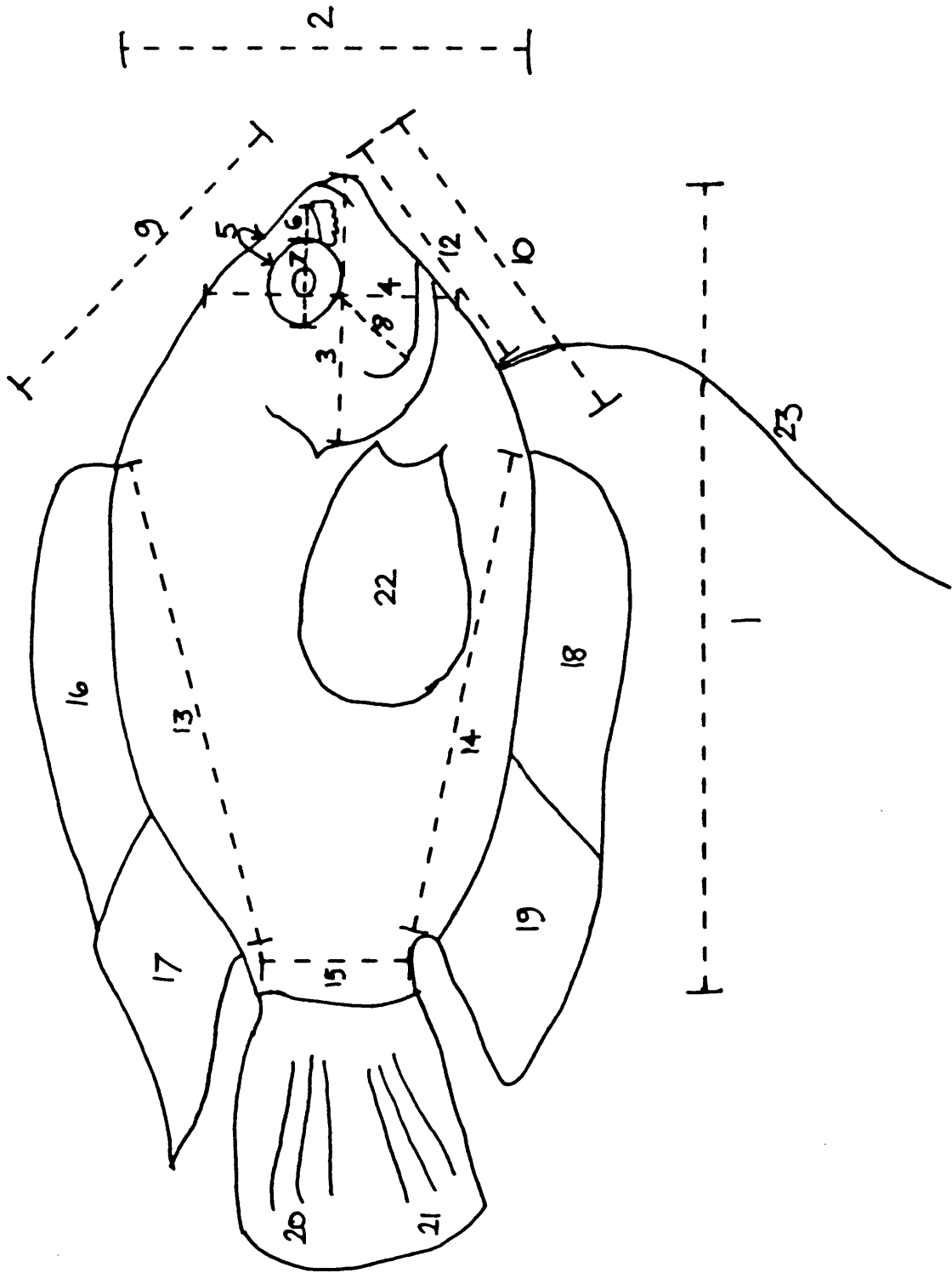
C. The following morphological parameters were measured

(see also Fig. 6.2.1):-

1. Standard length, measured from snout to caudal peduncular crease.
2. Maximum body height.
3. Length of head, measured from snout to posterior margin of opercular bone.
4. Height of head, measured perpendicularly from the isthmus.
5. Interorbital width.
6. Length of preorbital.
7. Length of orbit.
8. Depth of cheek.
9. Length from snout to base of first dorsal spine.
10. Length from snout to base of first anal spine.
11. Length from snout to most anterior point of pectoral fin base.
12. Length from snout to origin of pelvic fin.
13. Length of dorsal fin base.
14. Length of anal fin base.
15. Height of caudal peduncle measured at the caudal peduncular crease.
16. Number of dorsal fin spines.
17. Number of dorsal fin rays.
18. Number of anal fin spines.
19. Number of anal fin rays.



Fig 6.2.1 Fish Parameters.



20. Number of caudal fin rays in the dorsal lobe.
21. Number of caudal fin rays in the ventral lobe.
22. Number of pectoral fin rays.
23. Number of pelvic fin spines and rays (combined).

D. Other information was collected concerning the sex of the specimen, the name of the collector, the locality from which the specimen was collected, the catalogue number, the date of collection or registration and any further notes which were thought relevant about observable abnormalities, broken or torn fins, missing spines or rays.

Many of the specimens were borrowed from international museums and collections. Some came from the Zoological Survey of India, others from the Museum National D'Histoire Naturelle in Paris, others from the Zoologisches Museum der Humboldt-Universität in Berlin, and a large supply came from the Natural History section of the British Museum. This was done with the help of Dr. P. H. Greenwood and Mr. G. Howes of the latter establishment's Zoology department.

Borrowed specimens were preserved in 70% alcohol. Aquarium specimens were fixed in 10% formalin and then transferred to 70% alcohol. The measurements were made with adjustable mathematical dividers and rounded off to the nearest 0.5 mm. Any finer measurement would not have been justified by the nature of the parameters.

The total collection of 156 specimens consisted of:-

- |    |   |
|----|---|
| 9  | from species A (the nine originally collected by Day) |
| 29 | " " B (including two very immature specimens)         |
| 38 | " " C   |
| 25 | " " D   |
| 41 | " " E   |

2 specimens labelled as B, but almost certainly C  
2 D/E hybrids  
2 C/D hybrids  
8 miscellaneous specimens, labelled G, H, I which are  
*Macropodus opercularis*, *M. chinensis* and *M. cupanus cupanus*  
respectively.

The collection included some syntypes and a holotype.  
However the holotype of Species E appeared not to be a member of  
that species at all. Rather it was suspected to be a *Macropodus*  
and accordingly other specimens of this genus were introduced to  
try to confirm this suspicion.

Many of the specimens were damaged so that not all of the parameters  
could be measured. Ninety-two of the specimens that could be  
completely measured were used in many of the analyses. Where more  
than twenty complete specimens were available in one species, twenty  
were randomly selected to form this group. It was usually possible to  
determine the sex of the specimen by the extent of pointedness of the  
dorsal and anal fins (males are more pointed), but where life colours  
were still present, these were also used. Representatives were taken  
from both sexes whenever this was possible.

It will sometimes be useful to refer to the composition of a  
set of specimens in terms of the number of representatives from each  
species. This will be done by producing a vector with six  
components corresponding to Species A, B, C, D, E and others  
respectively. Thus the total set is described by (9, 29, 38, 25,  
41, 14) and the set of 92 completely measured specimens by (9, 20,  
20, 20, 20, 3).

A small subset of the data set is provided in Table 6.2.2.  
to give an impression of the range of values taken. The labels

TABLE 6.2.2

	<u>Variable</u>																						
<u>Specimen</u>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A 1	2.95	1.40	1.20	1.15	0.50	0.15	0.45	0.35	1.40	1.40	1.10	1.10	1.85	2.00	0.50	17	9	19	13	9	9	12	2
A 9	2.75	1.30	1.00	0.85	0.45	0.15	0.35	0.35	1.20	1.25	1.05	0.90	1.70	1.85	0.50	18	8	19	13	9	9	11	2
B 10	2.40	1.00	0.85	0.65	0.40	0.15	0.30	0.25	1.05	0.95	0.90	0.80	1.30	1.50	0.40	19	8	21	13	8	9	10	2
B110	3.10	1.55	1.00	0.95	0.50	0.20	0.40	0.40	1.35	1.35	1.05	1.10	1.95	2.00	0.50	17	7	18	11	8	8	10	2
C 14	3.50	1.80	1.20	1.20	0.50	0.25	0.40	0.50	1.55	1.55	1.20	1.15	2.20	2.30	0.70	17	8	17	15	10	9	10	2
C 60	2.55	1.35	0.95	0.80	0.40	0.20	0.35	0.35	1.30	1.25	0.95	1.00	1.40	1.40	0.45	15	9	17	15	10	9	11	2
D 23	5.95	2.95	1.85	1.40	0.75	0.30	0.50	0.60	2.20	2.35	1.85	1.65	3.45	4.20	1.00	17	9	17	20	10	9	11	2
D122	4.25	2.05	1.35	1.00	0.60	0.25	0.40	0.50	1.70	2.20	1.50	1.45	2.40	2.70	0.65	17	9	18	17	10	9	11	2
E130	4.40	1.85	1.50	1.20	0.65	0.30	0.45	0.45	1.75	2.10	1.55	1.45	2.60	2.75	0.65	16	12	16	16	9	9	11	2
E 33	7.15	3.00	2.20	1.80	1.05	0.40	0.60	0.75	2.80	3.10	2.20	2.20	4.30	4.50	1.20	16	10	17	16	9	9	12	2

Variables 1 - 15 are measured in cms; 16 - 23 are counts.

for a specimen include the species from which it is supposed to derive and a serial number.

The analysis was performed using single-link clustering (Section 6.4), ordinal scaling (Section 6.5) and partition likelihood clustering (Section 6.6). As for our other studies, programs were set up to analyse specified groups of specimens and sets of variables. The choice of similarity coefficient was varied; this forms the basis of Section 6.3, along with a discussion of the correlation between the variables, and treatment of standard length.

### 6.3 The Measurement of Similarity

A problem that required immediate attention was how to deal with the variability in size of the specimens when measuring their similarity. Six correlation matrices were produced for the set of variables based upon each of the five species individually, and finally their combination. Each of these showed that the variables that were based upon measurement of length (Nos. 1-15) rather than counts (Nos. 16-23) were highly positively correlated. For example, when considering all specimens, all such continuous variables had correlations in excess of 0.85. The picture was not so clear for the discrete variables, but spine counts tended to be slightly negatively correlated with the body parameters, whilst ray counts were positively correlated with these continuous variables. This is developed later.

These findings were a combination of two factors. Firstly, the large correlations were caused predominantly by size which acted to increase these measurements proportionally (at least to a first approximation). But also there were genuine effects, particularly amongst fin counts.

The process of growth is certainly not one of dilation, and so we anticipated difficulty in comparing different aged specimens from this set of variables. But since it was necessary to be able to classify specimens independent of their size, parameters 2 to 15 were divided by parameter 1, the standard length. The other parameters were left unaltered, since it seemed that these were not so influenced by the standard length.

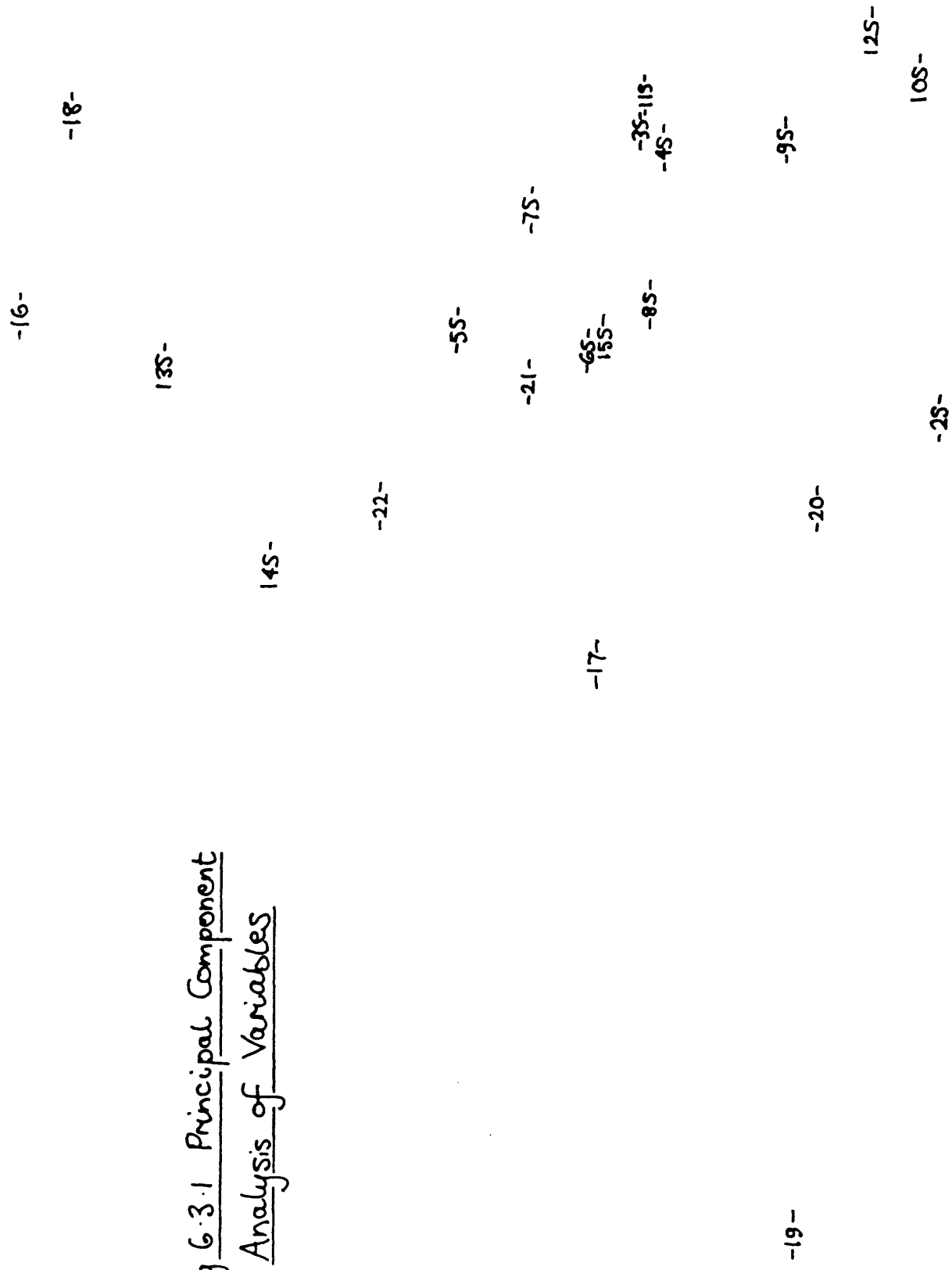
Re-evaluating the correlation matrices showed that a substantial amount of the correlation among variables had been removed, but that some variables were still highly correlated.

This was hardly a surprise, since some measurements are quite similar. For example the lengths from the snout to the anal fin and pelvic fin almost overlap (see Fig. 6.2.1) and so may be expected to be correlated.

The standard length was the largest of all measurements, so that in order to bring the ranges and variances of the standardised body parameters and fin counts into approximate agreement, the former values were multiplied by 30. This set of values then constituted the final description for use in partition likelihood clustering. Additionally some runs of single-link clustering and ordinal scaling were made using euclidean distance, based on these values, and averaged for the number of variables that were used in making the comparison. If this was done a criterion concerning the minimum number of variables for which both specimens of a pair could be measured had to be satisfied. A classical scaling start was possible when the set of specimens was chosen so that all pairs satisfied this criterion.

In this same initial analysis we looked further at the relationship among the measurements by performing a principal component analysis of the variables based upon the specimens for which complete data were available. The loads on the first four dimensions were 51%, 17%, 9% and 6% respectively. The first two dimensions are plotted in Fig. 6.3.1. This enabled us to determine the variables which explained most of the variation between specimens, given that this was dependent upon scaling factors used. Under the arrangements described above the ten variables with most variance were:-

Fig 6.3.1 Principal Component  
Analysis of Variables





1. Number of anal fin rays .. .. (19)
2. " " " " spines .. .. (18)
3. Height of body (standardised) .. .. (2S)
4. Number of dorsal fin rays .. .. (17)
5. Snout to pelvic fin (standardised) .. .. (12S)
6. Snout to anal fin (standardised) .. .. (10S)
7. Number of dorsal fin spines .. .. (16)
8. Length of anal fin base (standardised) .. (14S)
9. " " dorsal " " (standardised) .. (13S)
10. Snout to dorsal fin (standardised) .. .. (9S)

A trailing 'S' indicates that the value has been divided by standard length.

The way in which these variables tended to discriminate between the species is summarised in the following association of variables with the species for which they had relatively large values and small values.

<u>Variable</u>	<u>Variable is Large for Species</u>	<u>Variable is Small for Species</u>
(13S,14S,16,18)	A & B	C & E
(2S,9S,10S,12S)	C	A, B, D & E
(19)	D	A & B
(17)	E	A & B

In order to provide a check against similarities being unduly weighted by the presence of many correlated variables, Mahalanobis  $D^2$  statistics were also used to provide a measure of similarity. Section 1.11 outlined forms of this coefficient that were used, so that in various analyses either the scale effects or the

correlation effects, or both, could be removed. This approach was adopted with the original data both with and without the standardisation provided by division of variables by standard length. Values were supplied to single-link and ordinal scaling programs.

For all methods analyses were carried out on the body parameters alone, on the fin counts alone and on all parameters combined.

The mean length of specimens increased from those of A through B, C, D to E. Species E specimens were about three times longer than those from A. These procedures eliminated excessive dependence upon size. Some dependence was no doubt left, but that was inevitable and that real differences existed justified this further, because it would have been a feature that would have immediately discriminated between species.

6.4 Results Obtained From Single-Link Clustering

We illustrate the results obtained from applying single-link clustering by using the standard set of 92 completely measured specimens defined in Section 6.2 with body and fin parameters separately.

For both analyses the chaining effect proved a severe problem as clusters developed. We may demonstrate that the initial attachments were often in agreement with the supposed subdivision into species by the following displays. For each pair of species the number of occurrences of a link in the minimum spanning tree being formed by specimens from these species is recorded.

Body Parameters

A	7					
B	1	15				
C	0	2	17			
D	0	3	1	22		
E	2	1	1	11	15	
Other	0	0	0	2	0	1
	A	B	C	D	E	Other

Fin Parameters

A	4					
B	7	22				
C	0	3	16			
D	0	0	2	22		
E	1	1	5	11	13	
Other	0	3	1	0	0	1
	A	B	C	D	E	Other

The extra links are caused by truncation error in writing out the Mahalanobis distance calculation. This left open the possibility of equalities in similarity.

Most specimens start by linking to another of the same species. There is already a suggestion of similarity between A and B (fin) and D and E (body and fin). Other early cluster formation often agreed with the supposed subdivision as well.

However one large cluster soon emerged that tended to swallow the others. Thus considering body parameters, one cluster of form (7,17,18,18,20,0) soon developed, leaving eight single specimens (A,A,B,B,B,D,D,G) and two clusters of two elements (I,I) and (C,C). The picture was similar for fin parameters. Of course, this could have provided a perfectly satisfactory description of the data, but the more sophisticated technique of partition likelihood clustering, based upon more information, was able to demonstrate that this description was inadequate. Details are provided in Section 6.6. Nor could using another of the variety of similarity coefficients mentioned in Section 6.3 improve this situation. The position improved slightly upon amalgamating the body and fin parameters.

## 6.5 Results Obtained from Ordinal Scaling

We illustrate the results obtained from ordinal scaling by discussing results for the standard 92 specimens based on all parameters. Here the body measurements have been standardised and Mahalanobis distances are used. The final configuration is plotted in Fig. 6.5.1.

A classical scaling was performed first. Loadings on the first six dimensions were 83%, 7%, 4%, 2%, 2% and 1% respectively. This suggested that one dimension would be sufficient, but the first two dimensions were used to start the ordinal scaling so that a local minimum solution would be less likely, and because there was no additional difficulty in plotting the configuration. The ordinal method converged rapidly to the region of 7.4%, and then bumpily to a final value of 7.389%. We discuss the final configuration.

Species A. These nine specimens were fairly closely grouped. A4 was found in the main mass of B's and A3 was more distant from the others. These findings concurred with the partition likelihood results for A3 and A4.

Species B. Eighteen of the twenty were grouped together, B10 and B11 were separate. These two formed cluster 'C3' in Table 6.6.1.

Species C. At least sixteen of the twenty were well grouped in the configuration. C17, C15 and C114 were closer to other E specimens, C112 was separate and by itself.

Species D. The entire group was located in a narrow region, with several E specimens.

Species E. Eleven formed a clear group. Six were contained within the galaxy of D specimens at Stage 87 of the partition likelihood clustering (see 6.6.1). Three were quite separate and nearer A's, B's and C's. These three E126, E129 and E133 are



all identified as the three E's that do not fit a complete D/E union in 6.6.1, where a reason is given in terms of their morphology.

Macropodus Specimens. These were together and apart from all *Colisa* specimens, closer to C's than others.

Scaling demonstrated the broad division into species that became confused in single-link clustering. It also identified particular specimens that failed to conform to this supposed classification. A specimens seemed separable from B's, D's were contaminated with some E's. The picture was far from clear. We turn to the use of partition likelihood clustering which, as we shall see, confirms the unexpected positionings and adds weight to the feeling that the different species groups are largely separable, despite their contiguity in the scaling solution.

## 6.6 Results Obtained by Partition Likelihood Clustering

In Section 1.9 we observed that the method of partition likelihood clustering and those methods like it would have a special appeal for the taxonomist. In terms that he would understand we may state the problem as follows. "We are given  $N$  specimens from a particular genus and are required to produce the partition of these  $N$  specimens into  $K$  species which is most acceptable or likely."  $K$  is allowed to vary over the entire range from 1 to  $N$ , although the end points are trivial. It is most convenient to start from  $K=N$  and successively reduce its value by one, and this corresponds to a process of merging species. However it will possibly be beneficial to relocate individuals after such a merging. We have measurements from our specimens to use to form our judgements. For certain specimens the measurements will be more reliable, perhaps because of their greater size, and this must be taken into account in the analysis. As we have seen this consideration arises quite naturally.

The conceptual framework is attractive, but in addition we now claim that the method works very well in practice. The basis for this assertion is the set of results obtained from partition likelihood clustering of *Colisa*. We have considered our standard set of 92 specimens, from which complete data was available covering all 23 parameters. Results have been obtained based on the body measurements, the fin counts and both combined. For comparison with Section 6.5 we present first the results based upon all of the parameters simultaneously. The last stages of the clustering are presented schematically in Table 6.6.1 and the value of the loglikelihood function for successive stages is plotted in Figure 6.6.2.



TABLE 6.6.1 Partition Likelihood Clustering: The Final Stages

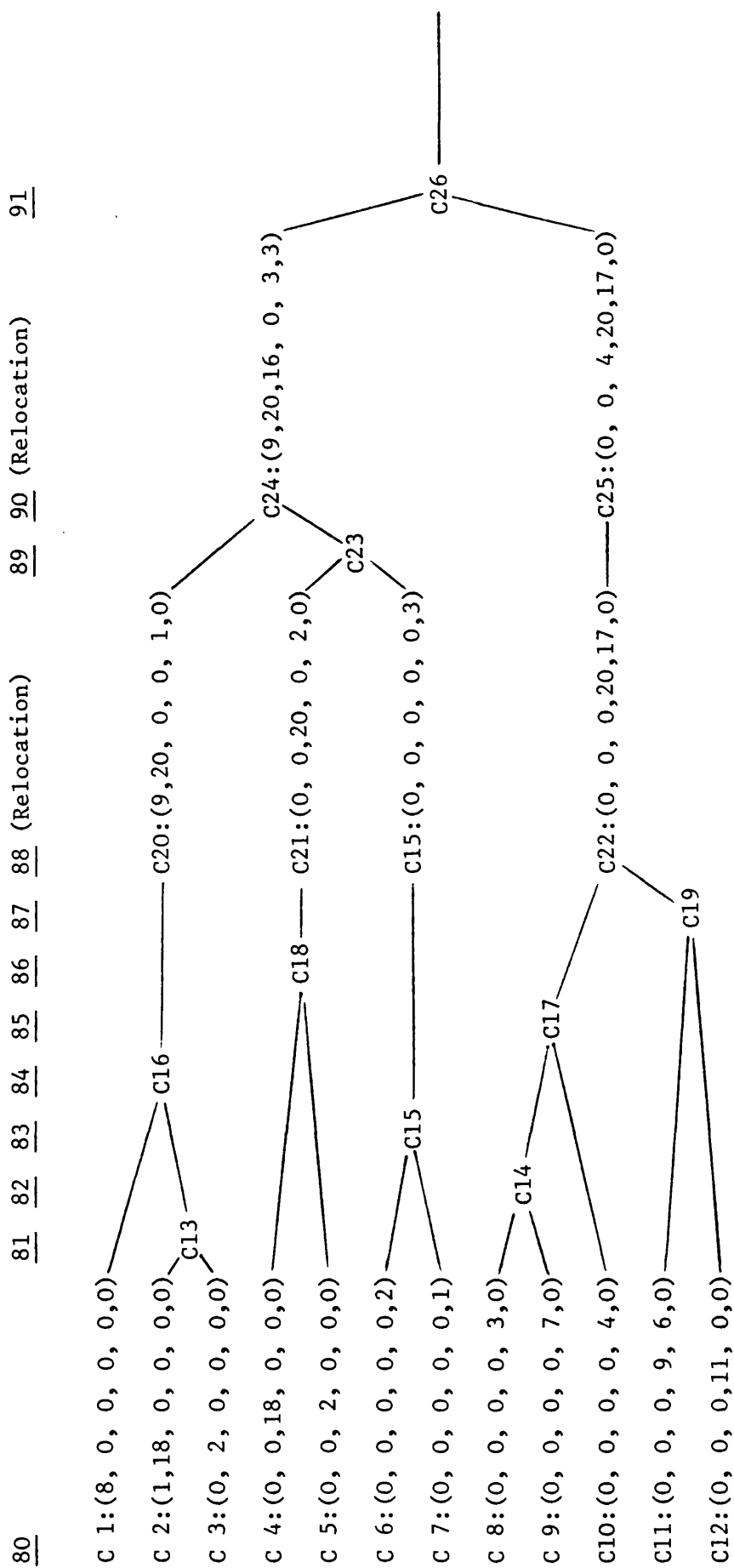
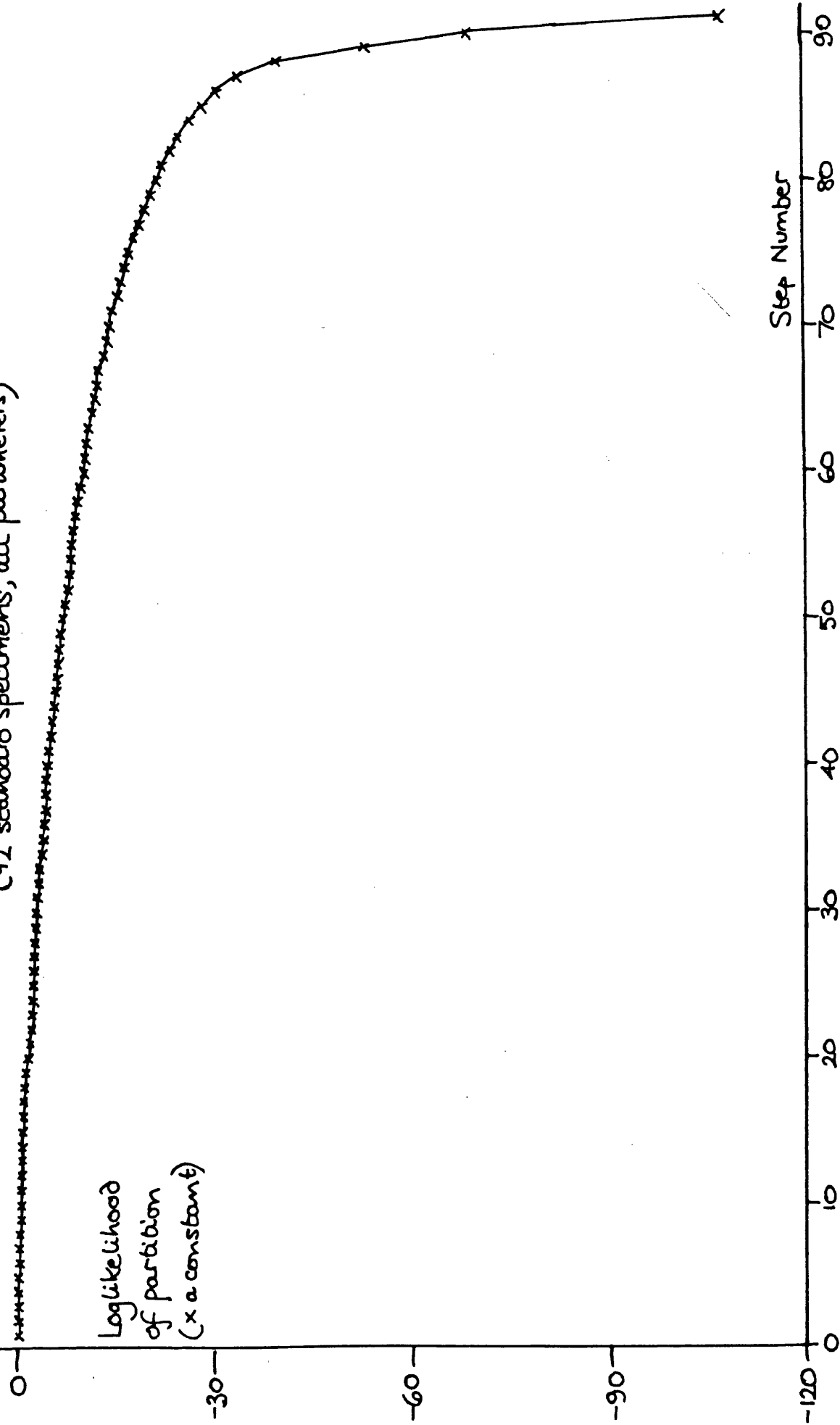


Fig 6.6.2 Partition Likelihood Clustering  
(92 standard specimens, all parameters)



The first seventy-one stages were simple unions of already existing groups. The twenty-one clusters that existed after these mergings had the following structure:

(2A), (5A), (2A + 2B), (1B), (5B), (2B), (8C), (4C), (6C), (2C),  
(3D), (3D), (6D), (3D + 1E), (4D + 1E), (1D + 4E), (3E), (7E),  
(4E), (2I); (1G)

where, for example, (1D + 4E) represents a cluster of one specimen from Species D with four from Species E. Thus the only clusters that contained specimens from more than one species followed the suggestions for classification that were made in Section 6.1.

During the next nine stages four relocations were required. The clusters that emerged are shown in Table 6.6.1. Clusters are identified in the table by abbreviation where 'C9' stands for cluster nine etc. Their structure was as follows:

Purely Species A	'C1'	8 specimens
" " B	'C3'	2 "
" " C	'C4', 'C5'	20 "
" " D	'C12'	11 "
" " E	'C8', 'C9', 'C10'	14 "
A/B combinations	'C2'	19 "
D/E "	'C11'	15 "
Purely Species G	'C7'	1 "
" " I	'C6'	2 "

The method then produced the following results:-

Stage 81: The entire B group was united. One A member remained with them.

Stage 82: Two purely E clusters united.

- Stage 83: The G and I clusters united.
- Stage 84: All A's and B's were united.
- Stage 85: Another purely E cluster joined the cluster from Stage 82, totalling fourteen specimens.
- Stage 86: All C's were united.
- Stage 87: All D's were united. At this point the initial classification hypothesis (Section 6.1) seemed to be quite accurate. A and B were together, C alone, D and E separable but with a few specimens in common, all *Macropodus* specimens being quite distinct.
- Stage 88: All D's and E's were united with the exception of three E's which left on relocation, two to the C group, one to the A/B group. The similarity between D and E was thus further established. The three E specimens that did not fit well into this scheme were re-examined. Each had an unusually low number of anal fin rays for an E specimen and it was noted that these (which were from the aquarium) all had a membrane between the last ray and the caudal fin. The membrane had taken the place of anal fin which would normally have contained more spine and ray.
- Stage 89: The *Macropodus* specimens united with the C group.
- Stage 90: Contrary to the hypothesised tree diagram classification of Section 6.1 the C group amalgamated with A/B. Four C specimens were relocated to join D and E.
- Stage 91: The process complete, just the one universal cluster remained.

In summary, the method worked very satisfactorily in so far as it agreed with the hypothesised classification. It was seen to be more powerful in discrimination than either of the other two methods

that had been applied to these data, for such clean results did not emerge from single-link clustering or ordinal scaling. This was achieved at a correspondingly greater cost in computer resources, and it is difficult to conceive that the method could be applied to data arrays that were much larger. However the results for body parameters and fin parameters which follow demonstrate that quite clean results were obtained with less data. The form of presentation is deliberately made comparable.

### Body Parameters

Stage 71: Final 21 clusters contained: (2A), (7A), (2B), (4B), (6B), (6B), (1B + 1C), (1B + 4D + 1E), (3C), (5C), (11C), (1D), (5D), (1D + 5E), (2D + 1E), (2D + 4E), (3D + 2E), (2D + 3E), (4E), (2I), (1G).

Stage 80: (9A), (8B), (10B), (1B + 8C), (1B + 4D + 2E), (12C), (7D), (2D + 5E), (2D + 8E), (5D + 5E), (2I), (1G).

Stage 81: (2I) + (1G) = (2I + 1G).

Stage 82: (8B) + (10B) = (18B).

Stage 83: (1B + 4D + 2E) + (2D + 5E) = (1B + 6D + 7E).

Stage 84: (1B + 8C), (12C), (18B) relocated to (1B + 17C), (18B + 3C).

Stage 85: (9A) + (18B + 3C) = (9A + 18B + 3C).

Stage 86: (7D), (2D + 8E), (1B + 6D + 7E) relocated to (1B + 6D + 8E), (9D + 7E).

Stage 87: (5D + 5E), (1B + 6D + 8E), (9D + 7E) relocated to (1B + 11D + 12E), (9D + 8E).

Stage 88: (9D + 8E), (9A + 18B + 3C), (2I + 1G) relocated to (9A + 18B + 3C + 2E), (9D + 6E + 2I + 1G).

Stage 89: Clusters are relocated to be (9,18, 3, 3, 9,0)  
(0, 1, 0,13,11,0)  
(0, 1,17, 4, 0,3)

Stage 90: Clusters are relocated to be (1, 1, 0,13,13,0)  
(8,19,20, 7, 7,3)

For body parameters alone, slightly different impressions were formed, indicating whether similarity between A's and B's or D's and E's was more strongly based upon body or fin morphology. In fact A's and B's were well separated in the early stages, D's and E's were well mixed. Again the C group joined the A/B combination but there were more stray specimens from D and E with them at the end.

#### Fin Parameters

Stage 71: Final 21 clusters contained: (2A + 6B), (2A + 1B),  
(5A + 1B), (5B), (7B), (4C), (3C), (3C), (6C), (1C), (1C),  
(1C + 5E), (1C + 2D + 1E), (7D + 3E), (6D + 1E), (5D + 2E),  
(4E), (2E), (2E), (1G), (2I).

Stage 80: (2A + 8B), (7A + 5B), (7B), (13C), (2C), (1C + 9E),  
(1C + 2D + 1E), (3C + 4D + 2E), (14D + 4E), (4E), (1G), (2I).

Stage 81: (2C), (1C + 2D + 1E), (1C + 9E) relocated to  
(3C + 2D), (1C + 10E).

Stage 82: (1G) + (2I) = (1G + 2I).

Stage 83: (2A + 8B) + (7A + 5B) = (9A + 13B).

Stage 84: (3C + 4D + 2E), (14D + 4E), (3C + 2D) relocated to  
(3C + 9D + 2E), (3C + 11D + 4E).

Stage 85: (9A + 13B) + (7B) = (9A + 20B).

Stage 86: (1C + 10E), (4E), (13C) relocated to (11C), (3C + 14E).

Stage 87: (11C), (3C + 9D + 2E), (3C + 11D + 4E) relocated to  
(14C + 1D), (3C + 19D + 6E).

Stage 88: (14C + 1D), (3C + 14E), (3C + 19D + 6E) relocated to  
(5C + 19D + 6E), (15C + 1D + 14E).

Stage 89: (9A + 20B), (5C + 19D + 6E), (15C + 1D + 14E)

relocated to (9A + 20B + 1C + 1E), (19C + 20D + 19E).

Stage 90: (9A + 20B + 1C + 1E) + (1G + 2I) =

(9A + 20B + 1C + 1E + 1G + 2I).

For fin parameters alone, A's and B's were a little more mixed, D's and E's more separate. Thus at Stage 85 a cluster of nine A's and thirteen B's merged with one of seven B's to achieve the grand union. At Stage 88 the D and E specimens were still quite distinct. Thus it would seem that A's and B's differed more on body parameters, D's and E's on fin counts. This time the C group merged with E's (Stage 88) and subsequently D's. The *Macropodus* specimens were clearly separated when considering fin counts.

## 6.7 Discussion and Conclusions

From the taxonomical point of view these analyses of morphological parameters have added weight to the proposed classification of Section 6.1. Certainly the suggested similarity between Species D and E has been supported. Likewise Species A and B have always been seen to be alike, but there is a suggestion that the group of nine A specimens does have an identity of its own. Whether this is caused by them all being males, by their having been collected from one location, or by a genuine specific difference is unresolved. It can certainly be claimed that they strongly resemble B specimens. The graph of loglikelihood against step number (Figure 6.6.2) for all parameters showed that after the existence of four clusters (A/B, C, D/E and *Macropodus*) a great strain was required to make any further mergings. This has added further support to the classification.

Other scalings were based on specimens for which not all of the parameters were available. These demonstrated the holotype of Species E to be central to the *Macropodus* group of eight specimens, and quite distant from other known *Colisa*. The hybrids were also interestingly located, intermediately between the groups formed by their parents.

These methods have also highlighted the relative importance for discrimination of each of the measured variables, and in particular the two sets provided by the body and the fins.

From the statistical point of view we have seen the usefulness of partition likelihood clustering. Its value has been the extra confidence that it enabled to be attached to suspected groupings in a scaling solution. A measure for the strain required to accommodate each new cluster has been useful. Single-link clustering,



although fast and efficient, was again not so informative. An important extra advantage of the partition approach was its natural formulation and, thus, appeal to the non-mathematician.

6.8 References

- DAWES, J. A. (1978). Aquarium fish in evolutionary biology. *Journal of College Science Teaching*, 8, pp. 237-241.
- DAY, F. (1878). *Fishes of India*, I, (London).
- LIEM, K. F. (1963). *The Comparative Osteology and Phylogeny of the Anabantoidei (Teleostei, Pisces)*  
University of Illinois Press, Urbana.
- PASCOE, E. H. (1920). The early history of the Indus, Brahmaputra and Ganges. *Quarterly review of the Geological Society*, 75, pp. 138-157.
- PINTER, H. (1960). Kritische Betrachtung zur Systematik einiger Colisa-Arten. *Die Aquarien-und Terrarien Zeitschrift (Datz)*, 13, pp. 198-201.
- SANDERS, M. (1934). Die fossilen Fische der altertiaren Susswassera-blagerungen aus Mittel-Sumatra. *Verhandelingen van het Geologisch-Mijnbouwkundig Genootschap voor Nederland en Kolonien*, 11, pp. 1-444.
- SEN, T. K. (1978). Fishes of Assam - with scientific, local and English names. *Seafood Export Journal*.
- SHERI, A. N. and SAIED, T. (1975). Revised list of freshwater fish fauna of Pakistan. *Pak. J. of Agric. Sc.*, p. 74.
- TATE-REGAN, C. (1909). The Asiatic fishes of the Family Anabantoidae. *Proc. Zool. Soc. Lond.*, p. 786.
- VIERKE, V. J. (1975). Beiträge zur Ethologie und Phylogenie der Familie Belontiidae (Anabantoidei, Pisces). *Z. Tierpsychol.*, 38, pp. 163-199.

C H A P T E R   S E V E N

AN APPLICATION IN NUTRITION

	<u>PAGE</u>
7.1    Introduction to the Data    ..    ..    ..    ..    ..	267
7.2    Results Obtained by Multidimensional Scaling    ..	272
7.3    References    ..    ..    ..    ..    ..    ..    ..    ..	281

## 7.1 Introduction to the Data

In this chapter we describe one additional application of ordinal scaling, namely to a study of regional trends and variations in dietary intake. The project was carried out with Michael Nelson, a nutritionist at the Medical Research Council Environmental Epidemiology Unit of the University of Southampton. An interest of this unit is the association between diet and morbidity due to various complaints such as gallstones, renal stones, diabetes etc. The nature of this work was exploratory, as it was based upon a routinely published set of data known to have many weaknesses. However it was felt worthwhile to make such an attempt because all measurements of intake are liable to serious error, and are expensive.

The data that were used were compiled for the Ministry of Agriculture, Fisheries and Food and published in their annual report by the National Food Survey Committee under the title "Household Food Consumption and Expenditure". Reports dating back to 1958 and up to 1979 were used. Each year the field workers sampled a number of private households in Great Britain. In 1978, 7,173 such households were used. Foods which entered into the household food supply intended for human consumption were recorded. Sweets, alcoholic drinks, soft drinks and foods eaten away from home were not recorded. Sampling took place throughout the year to minimise seasonal variations, and was based upon selected constituencies within standard regions of Great Britain, intended to be representative of the country as a whole. However only a limited number of localities were considered and these varied from year to year. Thus in 1978 Wales was represented by Merthyr Tydfil and Llanelli, the South West by Bristol North West, Wells, North Somerset and Taunton, but these

would have changed by the next year. This implied that comparisons between individual years would be prone to large amounts of error, because a different, small, potentially unrepresentative sample was used each time. Additional complications were provided by changes in the standard regions used. The Food Survey Committee thus compiled a large mass of information concerning consumption, income, prices, expenditure, individual foods, regions, age, social class and nutritional value. It was with the regional consumption of individual foods that we were concerned.

The abbreviations that were used for the regions are defined below, with the period for which the region was used.

EM = East Midlands (1967-79)

EW = East and West Ridings of Yorkshire (1960-66)

LO = London (1958-79)

MI = Midlands (1958-66)

NM = North Midlands (1958-66)

NO = North (1960-79)

NR = North and Yorkshire (1958-59)

NW = North West (1958-79)

SA = South East and East Anglia (1967-79)

SC = Scotland (1958-79)

SE = South East (1958-66)

SW = South West (1958-79)

WA = Wales (1958-79)

WM = West Midlands (1967-79)

YH = Yorkshire and Humberside (1967-79)

The boundary changes allowed the timespan from 1958 to 1979 to be split into six convenient periods. These periods were used

in the analyses in the hope that an aggregate of years would smooth out the fluctuations between individual years caused by the sampling plan. There were nine regions defined in the first time period, ten in all subsequent ones. The best approximation to the effect of the boundary changes was as follows:-

	<u>PERIOD</u>					
	<u>1958-59</u>	<u>1960-63</u>	<u>1964-66</u>	<u>1967-70</u>	<u>1971-74</u>	<u>1975-79</u>
	SW	SW	SW	SW	SW	SW
	SE	SE	SE	SA	SA	SA
	LO	LO	LO	LO	LO	LO
	MI	MI	MI	WM	WM	WM
<u>REGION</u>	NM	NM	NM	EM	EM	EM
	WA	WA	WA	WA	WA	WA
	NW	NW	NW	NW	NW	NW
	NR	EW	EW	YH	YH	YH
		NO	NO	NO	NO	NO
	SC	SC	SC	SC	SC	SC

Thus in all there were fifty-nine combinations of region and period, each of which was described by the two characters for region and two denoting the first year of the period. Thus SW67 represented the South West in 1967-70.

Seventeen broad categories of food were used. The measurements related to household food consumption and gave annual averages in units of ounces per person per week, unless otherwise stated. In order to study vitamin A intake, two additional variables were included. These were liver and carrot consumption. Thus we considered:-

MILK = Total milk and cream (in pints or equivalent)

CHEE = Total cheese

CRMT = Total carcass meat

MTPR = Total other meat and meat products

FISH = Total fish

EGGS = Total eggs (number)

FATS = Total fats

SGPR = Total sugar and preserves

POTA = Total fresh potatoes

GVEG = Total green vegetables

OVEG = Total other fresh vegetables

PVEG = Total processed vegetables

FFRT = Total fresh fruit

FRTP = Total other fruit and fruit products

BRED = Total bread

CERE = Total cereals (less bread)

BEVE = Total beverages

LIVR = Total liver

CRTS = Total fresh carrots

A 59 x 19 matrix of values was formed by taking the annual average consumption for a particular time period and region for all of these foods. In order to give an idea of the range and variability of the values taken, four rows of this matrix are provided in Table 7.1.1, and presented in column form.

From the four rows displayed in Table 7.1.1 it may be seen that there is a considerable time trend, and also geographical difference. Thus during those years more and more meat products, processed vegetables and fruit products have been eaten. Geographically, Scotland has consumed more bread and potatoes, less carcass meat,

green vegetables and fresh fruit. These facts are emphasised in the multidimensional scaling solutions described in the other section of this chapter. The configurations show both time and geographical differences, with a convergence as more convenience foods are eaten. The foods which are causing the differences are also displayed.

---

TABLE 7.1.1

	L058	L075	SC58	SC75
MILK	5.47	4.79	4.96	4.87
CHEE	3.30	4.07	2.67	3.60
CRMT	20.98	18.38	14.16	14.49
MTPR	17.29	23.90	17.75	22.71
FISH	6.35	4.89	5.44	3.99
EGGS	4.76	4.01	5.00	4.37
FATS	10.93	10.70	10.72	9.95
SGPR	20.70	12.56	21.83	13.73
POTA	52.64	39.69	60.19	44.95
GVEG	19.60	13.31	5.73	6.94
OVEG	10.15	16.04	10.13	13.25
PVEG	7.18	15.72	7.70	13.09
FFRT	27.89	23.01	16.07	14.67
FRTP	7.33	7.69	4.65	5.92
BRED	40.87	29.63	50.51	37.49
CERE	22.80	23.57	27.55	23.24
BEVE	3.78	3.08	2.81	2.52
LIVR	0.97	0.87	0.62	0.64
CRTS	2.33	2.72	3.41	3.19

---



## 7.2 Results Obtained by Multidimensional Scaling

Two scaling solutions are reported in this section. The first considers the region/period combinations with foods as the variables, to look for geographical and temporal trend; the second considers the foods with region/periods as variables in order to determine what is causing the differences shown up in the first configuration. Results of similar quality from the same technique have been based upon vitamin A-rich foods only, but are not reported here. A simultaneous solution of the problem formed, for example, by correspondence analysis (Hill, 1974) or a biplot (Gabriel, 1971) might have been attractive. However the results as they stand offered suggestions of the underlying mechanisms and proved to be easily understood by the nutritionists involved. The other methods would have required considerably more explanation and sophistication.

### (1) Region/Period Analysis

Some initial transformation of the data was necessary to ensure that the differences between region/periods were not dominated by largely consumed food variables, such as bread or potatoes. It was felt a priori that each of the variables should be considered equally potentially important in making up a difference. There is no reason to assume that an absolutely large quantity of any food must be eaten to change the relative risk of any disease. For example, small quantities of vitamin C are sufficient to eradicate scurvy. Accordingly the data were transformed so that the variables should have equal means and variances. Each variable was of independent interest in its contribution to overall similarity, and correlations, although they certainly existed, were not considered worthy of particular, individual attention. Thus a Euclidean distance was used to

measure dissimilarity between region/periods. The starting configuration for ordinal scaling was produced by principal component analysis, which for this case was formally equivalent to classical scaling, but simpler because it involved the inversion of a smaller matrix. The loadings on the first nine dimensions were

37%, 26%, 10%, 7%, 5%, 4%, 3% 2%, 2% respectively.

These indicated that at least two dimensions were required, but that a third was probably unnecessary and unreliable. The corresponding solution converged rapidly to that provided in Figures 7.2.1 and 7.2.2. The two plots show progress through time, and progress geographically. The final stress value was 12.8%.

Figure 7.2.1 shows points corresponding to each region (or its equivalent) connected by lines following progress through time. Each of these lines had a trend from the bottom left of the plot to the top right as time passed. The lines tended to converge, suggesting that the dietary habits of regions of Britain were becoming more uniform. In addition the lengths of segments of the line were seen to be in approximate proportion to the number of years' difference between the mid-points of the periods in question. Thus the last two segments, which covered the longest intervals of time, tended to be the biggest. The overall impression was that the plot showed a clear time dimension, with a stable geographical configuration being translated across the plot. The rate of change was approximately proportional to elapsed time.

Figure 7.2.2 highlighted the geographical contributions. The regions making up each time period were connected. In each period the pattern was similar. There was a gradient from London at one

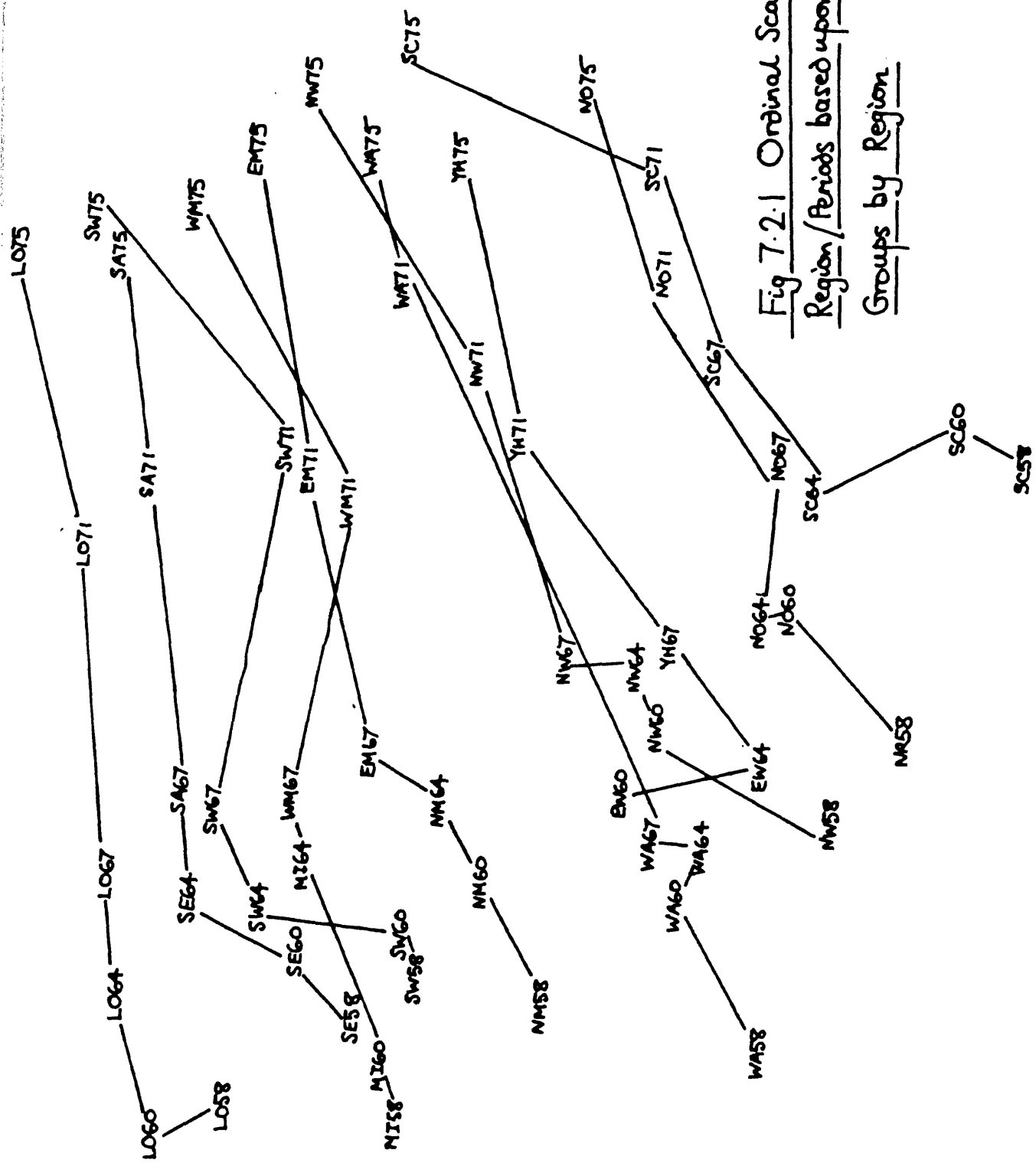


Fig 7.2.1 Ordinal Scaling of  
Region/Periods based upon Foods :  
Groups by Region



extreme, through the southern regions, the midlands, the north west, with Wales somewhere near, and Yorkshire, leading to the extreme north and then Scotland. As we have seen there was also a convergence effect bringing the extremes closer together. This gradient across the country was perhaps not too surprising, because many variables also show a similar pattern. Examples are mortality rates, social class measures and rainfall which are respectively low, high and low in London. Indeed the gradient corresponds quite well with latitude. That this should also apply to diet is to some extent accounted for by wealth differences, but it is also suggestive of a contributory influence to disease patterns.

(2) Food Analysis

Again an initial transformation of the data was required to ensure that we were not measuring just the differences in quantity of foods consumed. We did not want a single dimension from potatoes to liver, which would have provided no information about what was contributing to the regional differences. A slightly different device was used here. For each food the region/period values were transformed to their percentage contribution to the sum of all region/period values for that food. Thus the 19 x 59 data matrix was forced to have rows summing to 100.0. This ensured that all foods were comparable in magnitude, but that if a food had a particularly high coefficient of variation then this was preserved. Such variables would have contributed most to the region/period differences. The transpose of this matrix could have been used for the previous region/period analysis, but there we wanted to make no assumptions about the particular foods creating differences, and thus equated all variances. Again there were correlations among the variables. As we have seen, short distances

and small time differences would have brought about large correlations between region/periods. However each variable was felt to contribute its own important component towards overall similarity of foods, and accordingly a Euclidean distance was used. This generated no particular imbalance because regions and periods were all equally well represented.

This time the simplest initial configuration was obtained by inverting the 19 x 19 matrix required by classical scaling. The ordinal method converged quite quickly to a final stress value of 7.5%. The configuration is plotted in Figure 7.2.3. We turn to interpretation of the axes.

During the period covered by this study there was a general decrease in calorific intake per individual. This was the net result of a decrease in consumption of a lot of foods and the development of new foods that became more common. If we concentrate on the most recent of the six intervals used (1975-79) we can look at how many of the ten regions were consuming more than the overall average for all six intervals for any particular food. Doing this we found that foods for which at most one region consumed more than average

were:-

fish	beverages
sugars and preserves	eggs
fresh potatoes	fresh green vegetables
bread	

On the other hand those foods for which seven or more regions exceeded the average were:-

other meat & meat products	cheese
other fresh vegetables	carrots
processed vegetables	

These foods are marked on Figure 7.2.3. They were also identified by simple plots of the average national consumption for the different



time intervals. Looking at Fig. 7.2.3 one dimension was seen to describe these differences.

A pilot study of foods had been based upon the year 1978 only. This enabled the removal of the time factor so that the configuration of foods produced by ordinal scaling described differences between regions only. The configuration had one dominating dimension which ranged from green vegetables (high in the South) to potatoes (high in Wales and Scotland). In this pilot study liver and carrots were not used, but the other variables were ordered along this leading dimension as follows:-

- fresh green vegetables
- fresh fruit
- cheese
- other fruit and fruit products
- carcase meat
- other fresh vegetables
- milk and cream
- beverages
- cereals
- fats
- sugar and preserves
- processed vegetables
- eggs
- fish
- other meat and meat products
- bread
- potatoes

No doubt this ordering varied a little with time, but it would not have done so substantially. Comparison with Fig. 7.2.3



enabled us to identify another dimension, orthogonal to the first, which described foods accounting for the geographical differences. This hypothesis, based upon 1978, was easily extended to the other years of study by comparison with the original data. We can confidently claim to have identified the foods causing most regional diet difference.

(3) Conclusion

We have been able to explain and describe trends both temporal and geographical in dietary habits in Britain, by demonstrating a clear interrelationship between time, region and food type. This has been done by producing two ordinal scaling solutions, each in two dimensions, with two interpreted axes.

### 7.3 References

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, pp. 453-467.

HILL, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Applied Statistics*, 23, pp. 340-354.

APPENDIX: Index to Figures and Tables

	<u>PAGE</u>
Fig. 1.4.1      A plot of dissimilarity values against configuration distance .. .. .	10
Table 1.6.1    Eigenvalue spectrum from classical scaling of road distance data .. .. .	18
Table 1.6.2    Ordinal scaling iterations progress report for road distance data from a random start .. ..	20
Table 1.6.3    Ordinal scaling iterations progress report for road distance data from a classical scaling start .. .. . *   *   *   *   *	22
Table 2.4.1    Parameter ranges in random rankings studies *   *   *   *   *	62
Fig. 3.1.1      Binomial hyperplane dissimilarity plotted against euclidean distance; 50 hyperplanes ..	92
Fig. 3.1.2      Binomial hyperplane dissimilarity plotted against euclidean distance; 500 hyperplanes ..	93
Fig. 3.1.3      Expected Jaccard distance dissimilarity plotted against euclidean distance; 20, 50 and 100 discs of fixed radius 0.2 .. .. .	95
Fig. 3.1.4      Jaccard distance dissimilarity plotted against euclidean distance; 500 discs with radius distribution exponential mean 0.2 .. ..	97
Fig. 3.1.5      Wilkinson metric dissimilarity plotted against euclidean distance; 3,200 additional points in the larger disc .. .. .	99
Fig. 3.1.6      Jaccard distance dissimilarity plotted against euclidean distance; 1,000 discs with radius distribution exponential mean 0.2. Dissimilarities processed according to the assumption of an underlying bivariate normal configuration .. .. .	100
Fig. 3.1.7      Jaccard distance dissimilarity plotted against euclidean distance; 1,000 discs with radius distribution exponential mean 0.2. Dissimilarities processed according to the assumption of an underlying uniform configuration .. .. .	101
Table 3.2.1    Design of classical scaling simulations ..	103
Table 3.2.2    Mean values and sample standard deviations for procrustes statistics in the classical scaling simulations .. .. .	105

	<u>PAGE</u>
Fig. 3.2.3	Log mean procrustes statistic plotted against log hyperplanes for classical scaling simulations of binomial hyperplane model .. .. 106
Fig. 3.2.4	Log mean procrustes statistic plotted against log 'hyperplanes' for classical scaling simulations of independent binomial model .. 107
Fig. 3.2.5	Log mean procrustes statistic plotted against log discs for classical scaling simulations of Jaccard distance model .. .. .. 108
Fig. 3.2.6	Log mean procrustes statistic plotted against log discs for classical scaling simulations of processed Jaccard distances .. .. .. 109
Fig. 3.2.7	Log mean procrustes statistic plotted against level for classical scaling simulations of Wilkinson metric model .. .. .. 110
Fig. 3.2.8	Classical scaling eigenvalue spectra for the binomial hyperplane model in two dimensions .. 114
Fig. 3.2.9	Classical scaling eigenvalue spectra for the binomial hyperplane model in six dimensions 115
Fig. 3.2.10	Classical scaling eigenvalue spectra for the independent binomial model in two dimensions 116
Fig. 3.2.11	Classical scaling eigenvalue spectra for the independent binomial model in six dimensions 117
Fig. 3.2.12	Classical scaling eigenvalue spectra for the Jaccard distance model .. .. .. 118
Fig. 3.2.13	Classical scaling eigenvalue spectra for the Wilkinson metric model .. .. .. 119
Table 3.2.14	Performance of the trace and magnitude criteria .. .. .. 121
Table 3.3.1	Results of comparison of scaling methods .. 125
Table 3.4.1	Results of using random entries of similarity matrix .. .. .. 130
Fig. 3.4.2	Plots of procrustes statistic and stress against density for random entries of a similarity matrix .. .. .. 131
Table 3.4.3	Results of using large elements of similarity matrix .. .. .. 132
Fig. 3.4.4	Plots of procrustes statistic and stress against density for large elements of a similarity matrix .. .. .. 133

	<u>PAGE</u>
Table 3.5.1	Procrustes statistics arising from slightly different configurations .. .. . 138
Table 3.5.2	Procrustes statistics transformed to show accuracy of $\chi^2$ approximation .. .. . 139
Fig. 3.5.3	Kolmogorov-Smirnov test graph for appropriateness of $\chi^2$ approximation, $\xi=0.2$ .. .. . 141
Fig. 3.5.4	Kolmogorov-Smirnov test graph for appropriateness of $\chi^2$ approximation, $\xi=0.5$ .. .. . 142
Fig. 3.5.5	Kolmogorov-Smirnov test graph for appropriateness of $\chi^2$ approximation, $\xi=1.0$ .. .. . 143
Table 3.6.1	Appropriateness of approximations to procrustes statistics arising from slightly different squared distance matrices .. .. . 146
Table 3.6.2	Procrustes statistics after classical scaling and Sibson's approximations .. .. . 147
	* * * * *
Fig. 4.3.1	A sample coding sheet .. .. . 161
Table 4.3.2	Categories of divisions .. .. . 166
Fig. 4.4.1	Histogram showing participation in divisions 170
Fig. 4.4.2	Histogram showing mean, minimum and maximum participation in division categories .. .. . 172
Table 4.4.3	Frequency of agreement of party majorities by division category .. .. . 174
Fig. 4.7.1	Single-link clustering, an impression of the dendrogram for Cohort 1 on all divisions .. 181
Table 4.8.1	The extent of ordinal scalings of M.P. groups and division categories .. .. . 184
Table 4.8.2	A summary of the computational arrangement for ordinal scaling .. .. . 185
Table 4.8.3	The number of M.P.s for whom ten or less similarity values were defined by M.P. groups and division categories .. .. . 188
Fig. 4.8.4	Ordinal scaling of Cohort 1 on all divisions 190
Fig. 4.8.5	Ordinal scaling of Cohort 1 on Category 5 (electoral reform) divisions .. .. . 195
Fig. 4.8.6	Ordinal scaling of Cohort 1 on Category 7 (defence) divisions .. .. . 196
Fig. 4.9.1	Least squares scaling of Cohort 1 on all divisions .. .. . 200
	* * * * *

	<u>PAGE</u>
Table 5.2.1 The language groups .. .. .	211
Table 5.2.2 The phonetic groups .. .. .	213
Table 5.2.3 Some sample data .. .. .	214
Table 5.2.4 The varieties of 'English' .. .. .	216
Table 5.2.5 The set of techniques used .. .. .	217
Fig. 5.3.1 Gurage and its surrounding areas .. .. .	220
Table 5.3.2 The dialects of Gurage .. .. .	221
Fig. 5.3.3 Ordinal scaling of Gurage dialects .. .. .	222
Fig. 5.3.4 Ordinal scaling of Gurage dialects and surrounding languages .. .. .	224
Fig. 5.3.5 Ordinal scaling of languages around the Mediterranean .. .. .	227
* * * * *	
Fig. 6.1.1 Geographical distribution of <i>Colisa</i> .. .. .	235
Table 6.1.2 Species of <i>Colisa</i> .. .. .	236
Fig. 6.1.3 Suggested classification of <i>Colisa</i> species .. .. .	238
Fig. 6.2.1 Morphological parameters used .. .. .	240
Table 6.2.2 Some sample data .. .. .	243
Fig. 6.3.1 Principal component analysis of variables .. .. .	247
Fig. 6.5.1 Ordinal scaling of standard 92 specimens on all parameters .. .. .	253
Table 6.6.1 Partition likelihood clustering of standard 92 specimens on all parameters .. .. .	256
Fig. 6.6.2 Loglikelihood plotted against number of clusters for partition likelihood clustering of standard 92 specimens on all parameters .. .. .	257
* * * * *	
Table 7.1.1 Some sample data .. .. .	271
Fig. 7.2.1 Ordinal scaling of region/periods based upon foods; progress through time .. .. .	274
Fig. 7.2.2 Ordinal scaling of region/periods based upon foods; progress through space .. .. .	275
Fig. 7.2.3 Ordinal scaling of foods based upon region/ periods .. .. .	278
* * * * *	