UNIVERSITY OF STRATHCLYDE

PROBLEM STRUCTURING FOR THE ANALYSIS OF ARCHITECTURAL DESIGN DATA

ALAN H. BRIDGES, DipArch, MSc, ARIAS

A THESIS SUBMITTED TO
THE DEPARTMENT OF
ARCHITECTURE AND BUILDING SCIENCE
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

JULY, 1982

TABLE OF CONTENTS

PRELIMINARIES

CHAPTER 1 - THE PROBLEM DESCRIPTION

CHAPTER 2 - LITERATURE REVIEW

CHAPTER 3 - GENERAL STRUCTURE OF MAGIC

CHAPTER 4 - CLASSIFICATION

CHAPTER 5 - HIERARCHICAL CLUSTER ANALYSIS

CHAPTER 6 - EUCLIDEAN CLUSTER ANALYSIS

CHAPTER 7 - ORDINATION TECHNIQUES

CHAPTER 8 - NONLINEAR MAPPING

CHAPTER 9 - PRINCIPAL COORDINATES ANALYSIS

CHAPTER 10 - DISPLAY OF RESULTS

CHAPTER 11 - DATA CLUSTERING

CHAPTER 12 - EXAMPLES OF USE

CHAPTER 13 - CONCLUSION

FIGURES IN THE TEXT

ACKNOWLEDGEMENTS

The research for this thesis was carried out under a Science Research Council grant to Professor T. W. Maver at ABACUS, University of Strathclyde, Department of Architecture and Building Science. The thesis was later written up under the supervision of Professor Maver.

I am heavily indebted to Professor Maver for providing firstly the opportunity and facilities to undertake this work, and then the encouragement and support to write up the results.

Lastly, this thesis is dedicated to my parents who put up with many things for the sake of my education; and to Brenda and Jennifer who put up with me and my work whilst this thesis was finally being written.

ABSTRACT

One of the many problems in architectural design is the multivariate nature of the design problem. Typically this problem has been resolved by ranking the various design elements and working through a series of design modifications considering each in turn. Unfortunately architecture is concerned with complex situations in which many variables are simultaneously related and the "fragmentary" approach gives little insight into the basic relationships obscured within the design data. To achieve that insight the complex of variables must be studied as a whole.

This thesis describes a way of examining activity data sheets or other briefing data using a number of techniques based on multivariate statistical methods. The various techniques have been incorporated into a computer program called MAGIC - Multivariate Analysis by Graphical Interactive Computing. The program output is specially designed to produce diagrams to enable the designer to manipulate and investigate the design data easily and conveniently.

The thesis reviews the problem of architectural design and its place in design methods theory, and the relationship of MAGIC to other layout planning programs. The program structure is outlined and detailed descriptions of the analytical techniques presented, together with those graphical techniques developed to present the results. Finally the application of MAGIC is shown in two practical examples.

CHAPTER 1

THE PROBLEM DESCRIPTION

1.1 INTRODUCTION

One of the many problems in architectural design is the multivariate nature of the design problem. Typically this problem has been resolved by ranking the various design elements (adjacency, structural, servicing, environmental, etc., requirements) and working through a series of design modifications considering each in turn (figure 1.1). The design process thus appears as a branching tree where the various options are explored at each level and the "best" route through selected. Unfortunately architecture is concerned with complex situations in which many variables are simultaneously related and the "fragmentary" approach gives little insight into the basic relationships obscured within the design data. To achieve that insight the complex of variables must be studied as a whole.

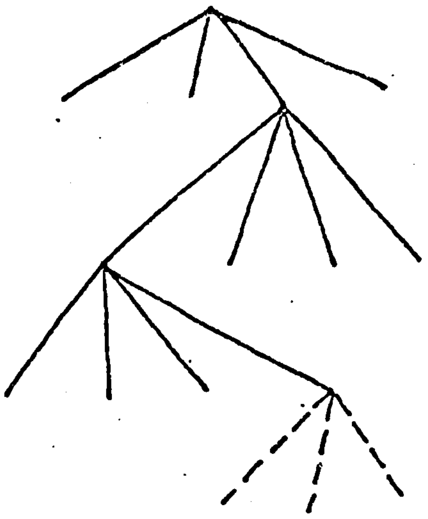
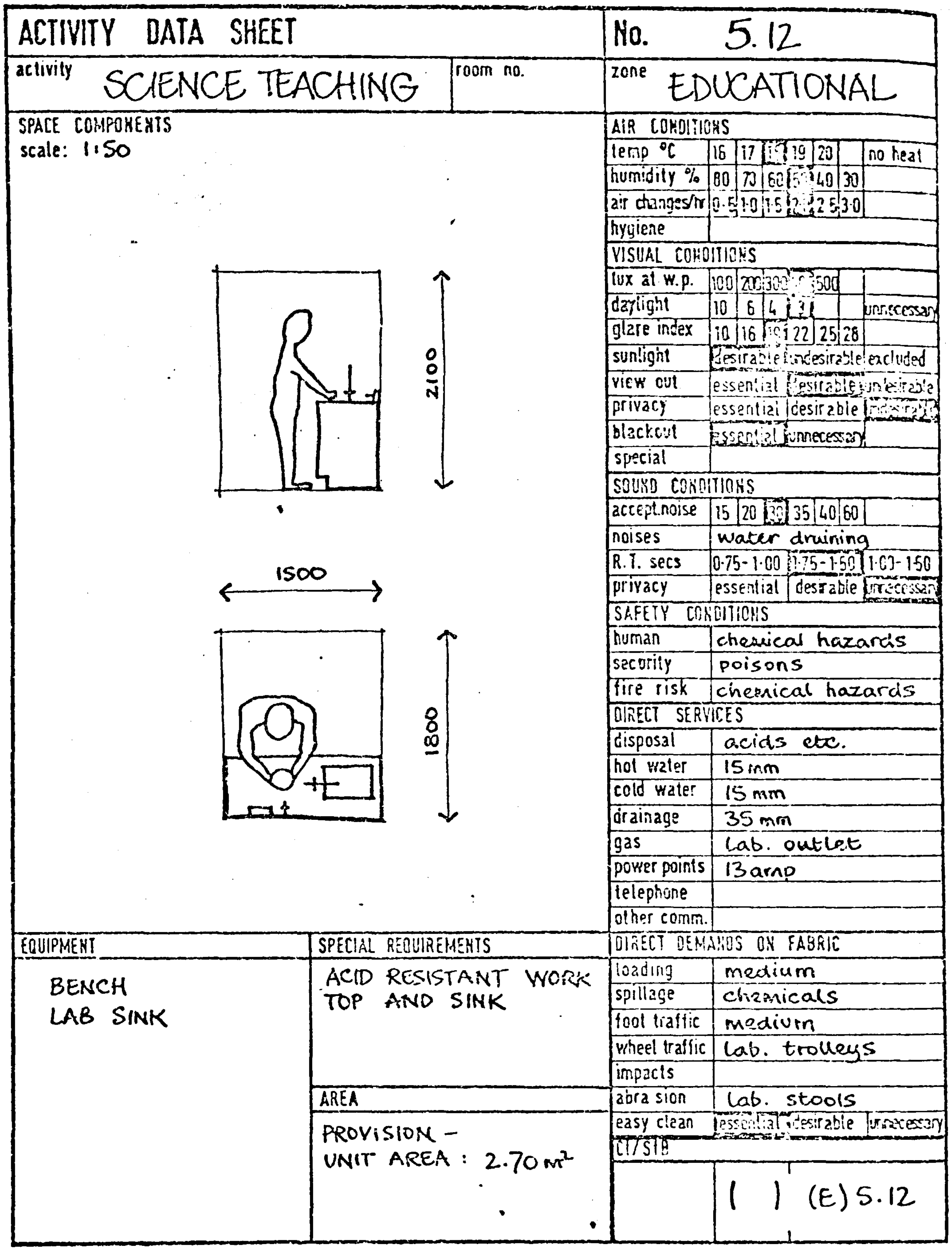
Designers ranking of importance in layout	Representation of design selection/modification	Action at each stage
adjacency requirements of activities		design layout to optimise adjacency requirements
structural requirements		modify first layout to take account of structural requirements select best compromise
servicing requirements		try various forms of servicing to fit plan developed so far. Compromise again.
and so on until all different aspects have been considered		finally check back against each set of requirements to make sure no major element has been too badly compromised by the successive modifications

Figure 1.1

Design Process Tree

Since the architectural layout problem is to arrange activities within a building the activities could be called associated if their physical demands require similar types of accommodation. To make clear the individual needs of the activities it is possible to use the Activity Data Method (Poyner 1966). All the activities to be accommodated may be listed, and then documented on an activity data sheet (a typical example is shown in figure 1.2). The lefthandside of the sheet describes the spatial requirements of the activity and the righthandside describes the characteristics required by that space to house that activity, i.e. temperature, light, service requirements, etc. The set of activity data sheets thus provide a comprehensive list of requirements to be met by the building. This thesis describes a way of examining this (or any other) data using a number of different techniques for analysing spatial and functional requirements to produce bubble diagrams and other design aids which may be of assistance to the designer in developing a plan layout.


2100
1500
1800

ACTIVITY DATA SHEET			No. 5.12
activity SCIENCE TEACHING		room no.	zone EDUCATIONAL

SPACE COMPONENTS
scale: 1:50

AIR CONDITIONS							
temp °C	16	17	[✓ 18]	19	20		no heat
humidity %	80	70	60	[✓ 50]	40	30	
air changes/hr	0·5	1·0	1·5	[✓ 2·0]	2·5	3·0	
hygiene							
VISUAL CONDITIONS							
lux at w.p.	100	200	300	[✓ 400]	500		
daylight	10	6	4	[✓ 3]			unnecessary
glare index	10	16	[✓ 19]	22	25	28	
sunlight	[✓ desirable]		undesirable		excluded		
view out	essential		[✓ desirable]		undesirable		
privacy	essential		desirable		[✓ undesirable]		
blackout	[✓ essential]		unnecessary				
special							
SOUND CONDITIONS							
accept.noise	15	20	[✓ 30]	35	40	60	
noises	water draining						
R.T. secs	0·75-1·00		[✓ 0·75-1·50]		1·00-1·50		
privacy	essential		desirable		[✓ unnecessary]		
SAFETY CONDITIONS							
human	chemical hazards						
security	poisons						
fire risk	chemical hazards						
DIRECT SERVICES							
disposal	acids etc.						
hot water	15 mm						
cold water	15 mm						
drainage	35 mm						
gas	Lab. outlet						
power points	13 amp						
telephone							
other comm.							
DIRECT DEMANDS ON FABRIC							
loading	medium						
spillage	chemicals						
foot traffic	medium						
wheel traffic	Lab. trolleys						
impacts							
abrasion	Lab. stools						
easy clean	[✓ essential]		desirable		unnecessary		
CI/SfB							
	( ) (E) 5.12						

EQUIPMENT	SPECIAL REQUIREMENTS	
BENCH LAB SINK	ACID RESISTANT WORK TOP AND SINK	
	AREA	
	PROVISION - UNIT AREA : 2.70 m²	

Figure 1.2

Activity Data Sheet

The various techniques have been incorporated into a computer program called MAGIC - Multivariate Analysis by Graphical Interactive Computing. The program can handle a mixture of different types of data and the graphical output and interaction facilities enable the designer to manipulate and investigate the design data easily and conveniently without usurping the designers own special expertise in the development of the final layout.

This introductory chapter reviews the problem definition and solution and its place in design methods theory. Also, as the solution techniques are based on a number of specialised statistical techniques the use and application of these methods in relation to the more commonly known and accepted theory of hypothesis-testing statistics is briefly discussed. Chapter 2 discusses a number of other layout planning programs and their relationship to MAGIC. Chapter 3 presents a summary of MAGIC showing how the program is controlled and the results displayed. The succeeding chapters then describe each of the analytical techniques incorporated in the program and the techniques developed to present the results. Finally the application of MAGIC is shown in two examples.

1.2 DESIGN METHODS

The origins of the design methods movement lie in the application of scientific techniques to a wide range of novel problems during World War II. The formalisation of these techniques formed the general subject area of what is now known as Operations Research (O.R.). In the 1950's these OR techniques were increasingly applied to management decision making and this formalisation of the "art of management" was one model that attracted the originators of architectural design methods. The early work is reported in Gregory (1967) and Jones (1970) and, although Alexander's "Notes on the Synthesis of Form" (1964) was influential, the classic text to emerge from this design science phase was "The Sciences of the Artificial" (Simon 1969). A number of the leaders of the movement later recanted, most notably Christopher Alexander (1971) and Christopher Jones (1977).

Apart from the ethical objections raised by Jones it became apparent that design problems were not that amenable to solution by scientific method. Rittel and Webber (1973) characterised design problems as "wicked" problems, as distinct from the "tamed" problems of science. Rittel (1973) further suggested that such design methods as existed were only the "first generation" design methods, and went on to outline the features of an emerging, more sophisticated "second generation". Second generation methods were characterised by Rittel as:

- assuming an equal distribution of knowledge about the problem (i.e. designers, users and others all have valid knowledge to contribute).

- embodying an argumentative process (i.e. influenced by different points of view rather than following a fixed method).

- casting the designer in a "midwife" role (i.e. there only to enable the interested parties to produce their own solution).

This generation of methods were prevalent during the design participation experiments of the 1970's (see, for example, Cross 1972).

The 1980's have already seen the emergence of a third generation of design methods. Broadbent (1979) has suggested that a common failing of the earlier generation methods was their prohibition of the designer's intuition. The third generation view is that designer inputs are a necessary part of any design method and the processes of Popper's (1963, 1968) "conjectures and refutations" model are seen as providing the mechanism for this. Hillier et al (1972) were amongst the first to discuss this approach to architectural design, and the method still enjoys some support (for example, Darke 1979). The method is not, however, without its critics: March (1976) suggests that its impact has been "pernicious", emphasising too much the superficial similarities of science and design. As a response to the continuing debate about scientific method philosophers such as Feyerabend (1975) have suggested that the only general methodological rule which could have universal validity in science (or design) is "anything goes". The problem in this approach to design methods hinges on the relationship of design to science, and, if an epistemologically coherent concept of science is still proving elusive then it seems unlikely that such a concept of design will develop satisfactorily. There now appears to be a growing body of opinion calling for science to be left to the scientists so that designers may get on with designing.

1.2.1 A Design Compromise

The central dilemma in the science and design problem concerns the relationship between present and future knowledge. March (1976) has summed it up as "Science investigates extant forms. Design initiates novel forms.". This dilemma may be side-stepped by recognising an interesting paradox - design is part of science whilst science is also part of design. A scientific experiment must be designed; equally, to engender any design, it must be initiated by the application of science. The relationship may be expressed diagrammatically (figure 1.3).

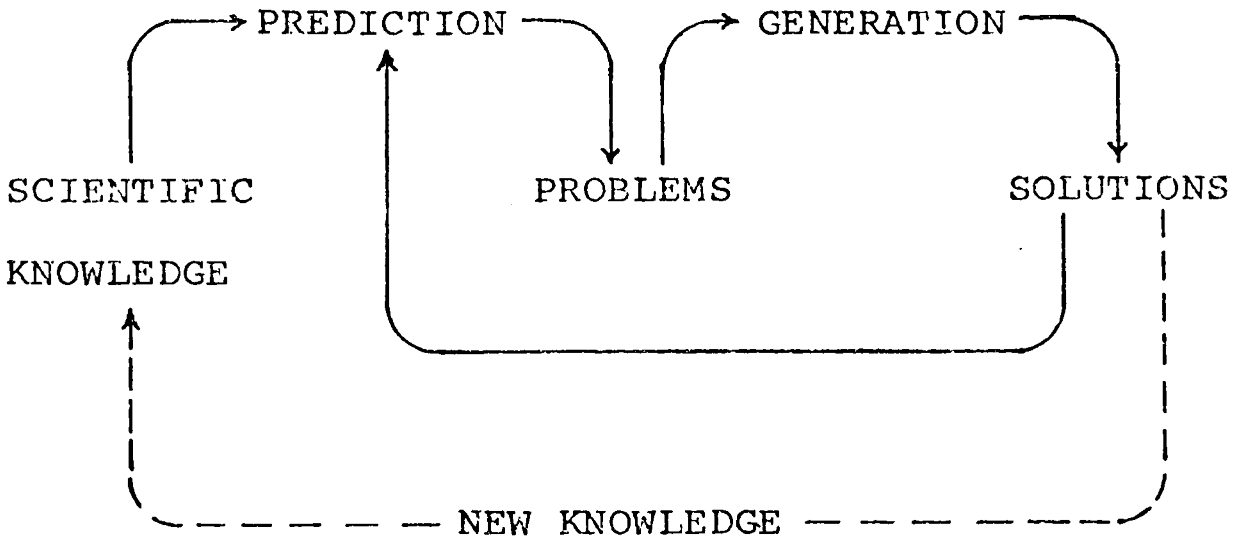


Figure 1.3

Relationship between science and design

As science reveals new knowledge the perception of problems is altered and the motivation for design changes. Furthermore, the impact of design on the corpus of knowledge is shown by the broken line linking future knowledge to science as existing knowledge.

Cross et al (1981) avoid the dilemma in a similar way, presenting design as a technological activity, and defining technology as "the application of scientific and other organised knowledge to practical tasks by social systems involving people and machines". This definition allows designers use of a variety of kinds of knowledge, from scientific knowledge of materials to craft experience. A number of other authors have developed the thesis that designing relies heavily on modes of thought which are neither "scientific" or "literary". Balchin (1972) coined the term "graphicacy" (as distinct from numeracy and literacy) to summarise those intellectual and practical skills concerned with nonverbal forms of communication. This approach is argued further by Archer (1979) in the context of defining design as a neglected central area of education. A related argument has been made by Ferguson (1977) who emphasises the role of "nonverbal thought" in technological development.

1.2.2 The Place Of MAGIC

This thesis has been written as a part-time occupation over a number of years. The initial impetus arose from an interest in the first generation Activity Data Method (Poyner 1966). This technique was developed to produce a detailed and comprehensive statement of the clients needs for the designer. It succeeded so well that the designer was completely overwhelmed with information. This caused the well-known break at the end of the Analysis stage in the then popular "Analysis - Synthesis - Appraisal" model when the designer, having purged his soul, put away the analysis to get on with the design. Believing that some useful information could emerge from the analysis the original idea behind MAGIC was to provide a means of interpreting or summarising in some useful way the body of data which may be available. MAGIC might now, alternatively, be considered a pioneer fourth generation design aid, making the numerate-literate subculture of the scientific-academic world accessable to the graphicate designer.

More seriously, the real utility of any model of the design process is not intrinsically bound up in the model itself. The value lies in the extent to which the model allows us to improve design teaching and practice. Architecture is a multidimensional problem and the solution area of any particular project is ill-defined. Furthermore a detailed knowledge of facts outside the universe of problem definition is needed to achieve a solution. To go full circle in the methodology debate, architecture is a classic example of an "ill-structured problem" (Simon 1973). The current methodologies recognise the need for interplay between two major contributing aspects of design:

- creativity and imagination (the "art" in design)

- recognising and satisfying formal constraints (the "science" in design)

MAGIC attempts to provide information on formal spatial requirements in a form suitable for the designer to work on creatively.

1.3 STATISTICAL METHODS

Architecture is concerned with complex situations in which many variables are simultaneously related, thus obscuring links and relationships. Furthermore, because of the intercorrelations systematic experiments comparing cause and effect in the relationships between seperate pairs of variables are not really possible: the complex of variables must be studied as a whole. In any case in the layout problem we are attempting to understand not just relationships between two variables but among sets of variables. The answer is not to be found in formal multivariate statistical techniques such as factor analysis - or any other technique which places too much reliance on numeric summaries of data based on distributional characteristics. Instead one must attempt to look at the overall pattern of the data. The approach is "exploratory" rather than "confirmatory", the underlying assumption of exploratory data analysis being that the more one knows about the data the more effectively that data can be used.

Although "data analysis" means the breaking down of data into its component parts, it is usually taken to mean the analysis of data by means of classical statistics alone i.e. by numerical summaries of the data to the exclusion of other methods of analysis. This tends to diminish the importance of the visual display of data and leads to a belief that a "statistic" is somehow more accurate or meaningful than a graphical representation. However even widely used statistical techniques may contain unreasonable hidden assumptions about the distributional nature of the data and the classical summary measures of data may conceal or even misrepresent the most informative aspects of certain data sets.

Exploratory data analysis is a method of examining a set of data from various angles and piecing together information about the system being studied. Such information may lead to a subsequent analysis that is refined and possibly more revealing, but Tukey (1977) makes the point "... to concentrate on confirmation, to the exclusion or submergence of exploration is an obvious mistake. Where does new knowledge come from?".

MAGIC uses a number of exploratory data analysis techniques, in particular a number of clustering methods. As this is a relatively undeveloped field of statistics, in the evaluation of the utility of these techniques efforts have been made to relate back to classical statistics wherever possible.

CHAPTER 2

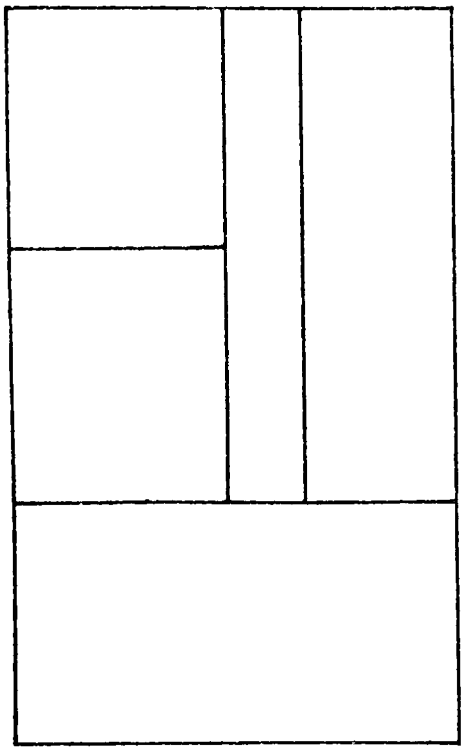
LITERATURE REVIEW

2.1 INTRODUCTION

The layout problem must have consumed more computer time than all other architectural applications put together. It was the first area to attract attention and has continued to exercise a fatal fascination ever since. Good reviews (and extensive bibliographies) of the progress in this field are to be found successively in Mitchell (1970a), Eastman (1972a) Mitchell (1975a) and Henrion (1978) This chapter briefly charts some of the main developments to set MAGIC in context.

2.2 FACILITIES PLANNING

The earliest layout programs were developed to allocate facilities to a floor plan divided into suitable modular areas. CRAFT and CORELAP (Armour and Buffa 1963, Lee and Moore 1967) are typical programs from this era. A floor plan layout is represented within CRAFT as a two-dimensional array of integers as in figure 2.1.



1	1	1	2	3	3
1	1	1	2	3	3
1	1	1	2	3	3
4	4	4	2	3	3
4	4	4	2	3	3
4	4	4	2	3	3
5	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	5	5

Figure 2.1

Integer array representation of floor plan

Each integer represents a square module of space of some defined dimension, and aggregations of such modules represent "rooms" and "departments". Letting S = (i_1, i_2, ...,i_m) be the set of modules to be located and R = (j_1, j_2, ..., j_m) the set of possible locations in the array, CRAFT attempts to allocate S to R to maximise some specified criterion. Vollmann and Buffa (1966) produced an overview of the problem, and several possible solution techniques were developed. The solution techniques are now generally referred to as "additive" (successively adding facilities trying to maximise the target criterion at each step) and "permutational" (allocating all the facilities and permuting their positions to try and achieve an improvement in the criterion). Nugent et al (1968) present a comparison of CRAFT with two earlier techniques (Hillier 1963 Hillier and Connors 1966) and a technique of their own.

Architects soon took an interest in these formalised planning techniques, particularly applied to the analysis of circulation patterns (Mosely 1963; Whitehead and Eldars 1964, 1965; Beaumont 1967). Other architectural layout programs were developed by Johnson (1970) Willoughby (1970) Mitchell (1970), Portlock and

Whitehead (1971) and Stewart and Lee (1972). Portlock and Whitehead (1974) later extended their technique to plan in three dimensions. Moore (1974) presents a general survey of facilities planning work to that data. Phillips (1969) compared a number of programs in an architectural context. Lew and Brown (1970) modified CRAFT for architectural use and Carter and Whitehead (1975a) looked at the effect of the quality of data on the plans produced. Gawad and Whitehead (1976) attempted to progress the technique by adding communication paths to the diagrammatic "idealised" layouts. Other sophistications enabling layout programs to work with realistically large problems are reported by Shaviv and Gali (1974).

Eastman (1972) presented a generalised formulation of the space planning problem. The EDRA 3 conference produced three papers outlining techniques which are receiving increasing attention today. Liggett (1972) discussed floor plan layout by implicit enumeration and Mitchell and Dillon (1972) and Frew et al (1972) introduced polyomino "pattern-building" techniques to the problem. Liggett, in particular, has continued to work on this problem and her recent publications include an efficient solution method for the quadratic assignment problem (Liggett 1980) and practical applications of the technique in office planning (Liggett and Mitchell 1981) Frew's work with polyominoes is extended in Shapira and Frew (1974) and related work reviewing the literature of polyominoes and formalising an architectural application is found in March and Matela (1974). Further approaches to achieving an efficient computational procedure to solve the facilities problem are contained in Juel and Love (1976) and Loomis (1977). Jackson (1977) presents a further architectural formulation.

2.3 GRAPH THEORETIC APPROACHES

The N-ominoes approach provides an interesting link with graph-theoretic based layout methods. A floor plan may be regarded as a planar graph, in which corners of spaces are nodes and walls are edges. The dual of the graph thus represents adjacencies. Procedures were developed for constructing a floor plan given the adjacency graph or matrix. Thus the adjacency graph became used for the solution of a class of layout problems which were specified in terms of required adjacency between spaces. Levin (1964) was the first to discuss floor plan layouts using graphs. Other early work based on graph representations is found in

Krejcirik (1969), Seppanen and Moore (1969), Grason (1969, 1970), Steadman (1970, 1973) Cousin (1970) and Pereira et al (1973). The formal graph-theoretic aspects of polyominoes are defined by Matela and O'Hare (1976). Foulds and Robinson (1976) present a graph theoretic solution to the plant layout problem, and a number of the previous authors combine (Mitchell et al 1976) to describe a set of algorithms to produce a limited set of plans.

Graphs have also been applied to the slightly different, but closely related, field of problem structuring. Alexander (1965) first proclaimed a city is not a tree provoking the interesting (if belated) response from Harary and Rockey (1976) that it is not a semi-lattice either. Other graph theoretic decomposition algorithms are described by Shaviv et al (1977, 1978).

2.4 STATISTICAL APPROACHES

A number of authors have looked at a statistical approach to problem stucturing. Rossi (1970) prepared a survey of classification techniques for the Department of Architecture at the University of Bristol, but presents no information on implementation or architectural application. Mitchell (1970) describes a clustering program, CLUMP3, similar in many respects to Milne s (1971) better known CLUSTR. CLUSTR operates on a binary interaction matrix to produce a semi-lattice structuring of the problem. A direct link with facilities planning is found in Carrie (1973) who uses a modified Nearest Neighbour clustering to obtain plant layouts from a single criterion "adjacency matrix".

Carter and Whitehead (1975b, 1976) describe an analytical program to derive clusters from an association matrix, plot a dendrogram and link to a layout stage. They conclude that the clustering approach produces better layouts than their previous "additive" or "permutational" facilities layout programs. Frew (1976) in a broad review also suggests clustering methods hold more promise than the heuristic and enumerative techniques.

Tabor (1976) produced one of the most complete surveys, covering not only the permutational and additive techniques and graphs, but also hierarchical classification methods, a "clumping" nonhierarchical method, and multidimensional scaling. This is all done in the context of the analysis of communication patterns and so is derived only from the trip matrices, but is the only paper to show some of the advantages of bringing a number of different techniques to bear on the same problem.

There is some history of the use of multivariate statistics in architectural applications in France Maroy and Peneau (1973) summarise a number of techniques and present a factor analysis mapping using a mixed data matrix. Ullrich and Braunstein (1977) describe the use of multidimensional scaling and cluster analysis to help clients structure their design requirements to prepare architectural briefs. Fortin (1978) uses a mapping algorithm to produce relationship diagrams from relationship matrices, and Roy (1979) uses classical multidimensional scaling to the same ends.

2.5 SUMMARY

A wide range of solution techniques have been applied to the layout problem and the associated problem-structuring question. None have satisfactorily solved the problem of the multivariate data. None have provided more than one solution technique operating on the same data set. None have provided any kind of user interface for effective interactive use by designers.

CHAPTER 3

GENERAL STRUCTURE OF MAGIC

3.1 INTRODUCTION

MAGIC (Multivariate Analysis by Graphical Interactive Computing) is an interactive, graphical computer program for space planning. The program is carefully designed to allow the architect to investigate a planning problem, and outputs information in diagrammatic form of sufficient generality not to inhibit the designer, whilst containing a distillation of information such that the final architect-produced design will closely meet the requirements of the organisation. Although of use in any layout analysis the program is illustrated here by a simple theoretical example. Examples of the practical use of the program are presented in chapter 12.

MAGIC is designed to operate during the early design stage analysis. That is, given almost any planning data, the program will analyse that data and present the results as relational diagrams for the designer to then use in the preparation of the final layout. As several types of data may be collected a computational problem may arise if incompatible data types are used together. To avoid this most analysis programs only allow the use of one variable - thus forcing all relationships to be expressed in terms of adjacency requirements or cost. This program deals with the problem in an entirely different manner which allows the use of different data types in such a way that the veracity of the computed output is maintained across a range of variables. Thus, in addition to the typical inventories of equipment and furniture, available office space, work station requirements, etc., it is possible to compile information on the required physical environments of the various activities and use this data in the analysis. Details of each of the computational techniques are presented in the chapters following this general description.

3.2 A BRIEF DESCRIPTION OF THE PROGRAM

MAGIC is designed for interactive use on a direct view storage tube terminal. The type of analysis performed and the manipulation and comparison of bubble diagrams is controlled through a series of menus. Any (or all) of the analyses can be performed from the same data, and previously computed solutions may be retrieved for further manipulation or comparison with new solutions. This thesis elaborates each of the analyses in turn and then describes the manipulative facilities. The description is necessarily "linear" but it should be understood that one of the main advantages of a computerised analysis is the ability to quickly and easily move "backwards and forwards" through different analyses and modifications thus gaining the understanding of the structure of the data which will enable the final layout to be designed. Figure 3.1 defines the basic program structure.

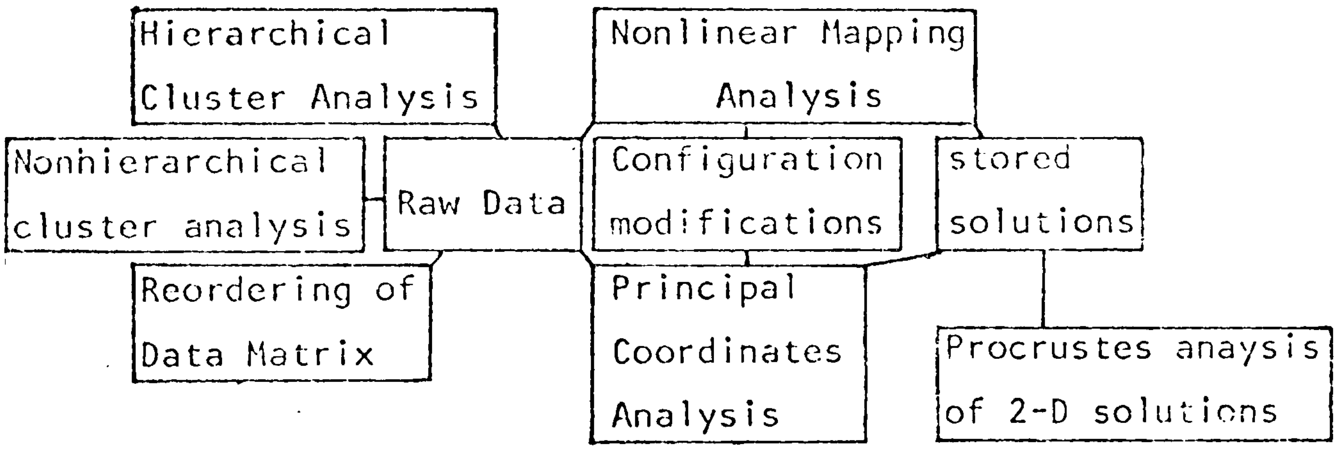

Hierarchical Cluster Analysis
Nonlinear Mapping Analysis
Nonhierarchical cluster analysis
Raw Data
Configuration modifications
stored solutions
Reordering of Data Matrix
Principal Coordinates Analysis
Procrustes anaysis of 2-D solutions

Figure 3.1

Diagrammatic structure of MAGIC

Four main types of analysis are used - hierarchic and nonhierarchic (Euclidean) cluster analysis, a nonlinear mapping ordination and principal coordinates analysis. A fifth type of analysis enables the reordering of the data matrix to cluster highly associated activities. The output from the hierarchic cluster analysis is displayed in the form of a tree-diagram or dendrogram, whilst the nonhierarchical clustering, principal coordinates analysis and nonlinear mapping are all arranged to produce bubble diagrams.

3.3 COMPUTATIONAL ACCURACY

Computational accuracy is an important topic which is often ignored - it should be selfevident that a computational algorithm should not distort the data. Although the old chestnut "garbage in, garbage out" is well known, what is less often realised is that good data may be transformed into garbage by an inefficient or unstable algorithm. When multivariate analysis is being used to explore data sets, as in MAGIC, it is obviously essential that any patterns emerging from the data should be a reflection of the structure of the data rather than the result of badly designed computational methods.

Many computing procedures still in common use were developed for hand calculation or desk calculators. These tend to emphasise ease of computational procedure over accuracy and often require the user to round intermediate results in an intelligent manner. The same algorithm when coded for a digital computer may well produce wildly inaccurate results. Longley (1967), Wampler (1970), and Youngs and Cramer (1971) discuss this point in relation to computer programs for multiple regression analysis (which provides a good example, requiring many summations and inversions of matrices) and their results show that even widely used "packages" do not always employ reliable algorithms.

Pennington (1970) and Dorn and McCracken (1972) develop the numerical analysis problems further. They define three main types of error, arising from inaccuracy in data preparation, errors of machine representation of floating point numbers, and arithmetical errors. These could be generalised as physical, computational and mathematical errors.

Errors in data preparation are almost inevitable and careful checks are made in MAGIC on all input data.

Errors of representation and arithmetic are rather more serious as they cannot be observed. Particular attention has therefore been paid to the machine implementation of the various analytical techniques and the performance of constituent parts carefully checked, both by hand and by reference to standard test problems from Gregory and Karney (1969), Malcolm (1972) and George (1975).

3.4 THE PROGRAM

All data to be used in the analysis should be prepared in advance and stored in a diskfile. The data files should be constructed as follows:

1. Number of rows and columns of data (NR, NC). Maximum NR*NC is about 10000 but depends on the analysis selected. Online data checks ensure program limits are not exceeded.

2. The type of data - distance matrix (1) or otherwise (0).

3. The numbers of different variables of each type, in the order continuous, multistate (if any), binary (if any).

4. Job title - maximum of 60 characters. This is

printed as a heading on all graphical output.

5. The area requirements associated with each activity. If no areas are available, or are not appropriate to the analysis being undertaken write 0 (zero).

6. The names of the activities. If no names or descriptors are required write NONE.

7. The form of the association or activity data matrix.

1 if full matrix

2 if upper triangular matrix (including diagonal)

3 if lower triangular matrix (including diagonal)

8. The data in the form specified in 7.

If the data is an ordinary dissimilarity matrix the interpoint distances should be entered. That is small numbers imply a requirement to be close together. If the analysis is to be carried out on observed data (say number of trips between rooms) where a large number implies a requirement to be close together then this data should be modified in some suitable way (say (nmax + 1) - n) for entry into the data file.

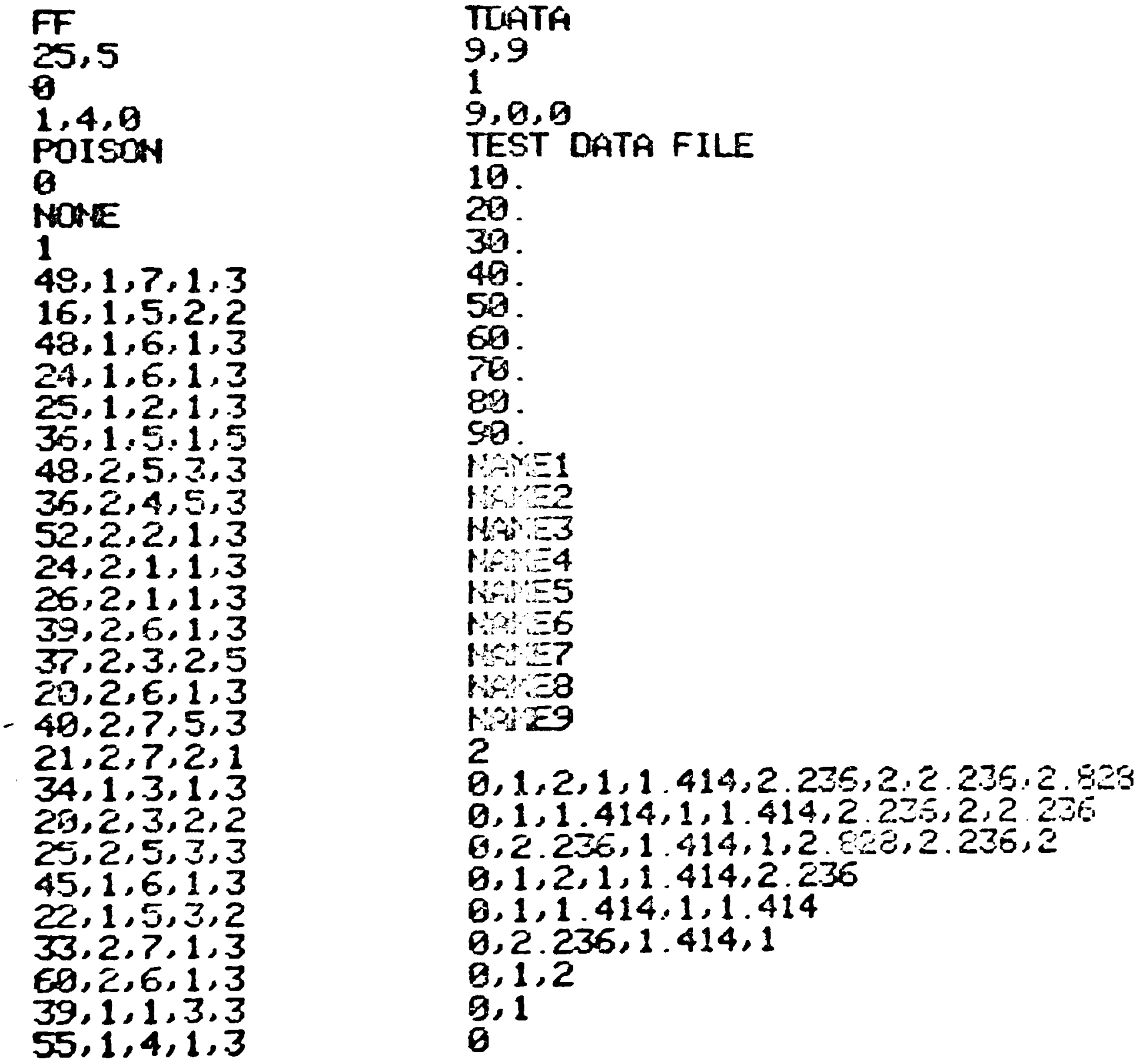
If mixed data is being used it should be ordered by variable type as follows:

1 continuous (quantitative variables)

2 multistate variables (if any)

3 binary variables (if any)

Figure 3.2 shows examples of data files



```
FF
25,5
0
1,4,0
POISON
0
NONE
1
49,1,7,1,3
16,1,5,2,2
48,1,6,1,3
24,1,6,1,3
25,1,2,1,3
36,1,5,1,5
48,2,5,3,3
36,2,4,5,3
52,2,2,1,3
24,2,1,1,3
26,2,1,1,3
39,2,6,1,3
37,2,3,2,5
20,2,6,1,3
40,2,7,5,3
21,2,7,2,1
34,1,3,1,3
20,2,3,2,2
25,2,5,3,3
45,1,6,1,3
22,1,5,3,2
33,2,7,1,3
60,2,6,1,3
39,1,1,3,3
55,1,4,1,3
```

```
TDATA
9,9
1
9,0,0
TEST DATA FILE
10.
20.
30.
40.
50.
60.
70.
80.
90.
NAME1
NAME2
NAME3
NAME4
NAME5
NAME6
NAME7
NAME8
NAME9
2
0,1,2,1,1.414,2.236,2,2.236,2.828
0,1,1.414,1,1.414,2.236,2,2.236
0,2.236,1.414,1,2.828,2.236,2
0,1,2,1,1.414,2.236
0,1,1.414,1,1.414
0,2.236,1.414,1
0,1,2
0,1
0
```

Figure 3.2

Data Files. File FF (on the left) is used in the following examples. Unlike TDATA it does not contain details of areas or activity names.

The program is run interactively. The first display is a title page. To continue press the return key. A menu showing the available analysis routines is then displayed. The menu consists of 8 items as follows:

```
INPUT
EDIT
CLUS
HCLUS
NLMAP
PCOORD
ROWCOL
FINISH
```

INPUT enables the specification of the prepared data file, and should be selected before any analysis is attempted.

EDIT enables selections to be made from the input data file for seperate analysis, and also enables rows and columns to be interchanged.

CLUS enters the Euclidean cluster analysis section.

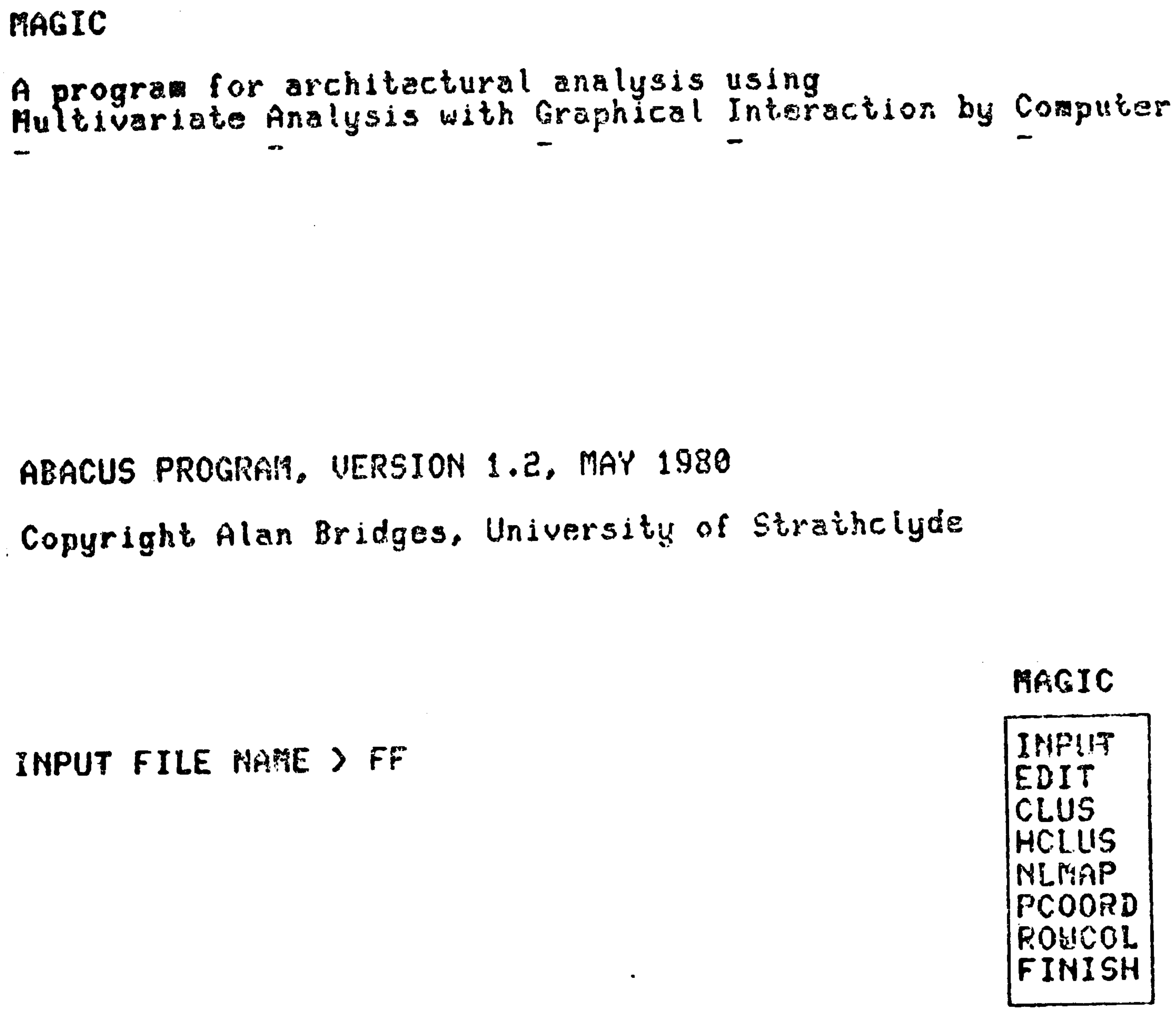
HCLUS enters the hierarchical cluster analysis section.

NLMAP enters the nonlinear mapping section.

PCOORD enters the principal coordinates mapping section.

ROWCOL enters the row/column clustering by reordering section.

FINISH provides for an orderly exit from the program.



```
MAGIC

A program for architectural analysis using
Multivariate Analysis with Graphical Interaction by Computer




ABACUS PROGRAM, VERSION 1.2, MAY 1980
Copyright Alan Bridges, University of Strathclyde


INPUT FILE NAME > FF
```

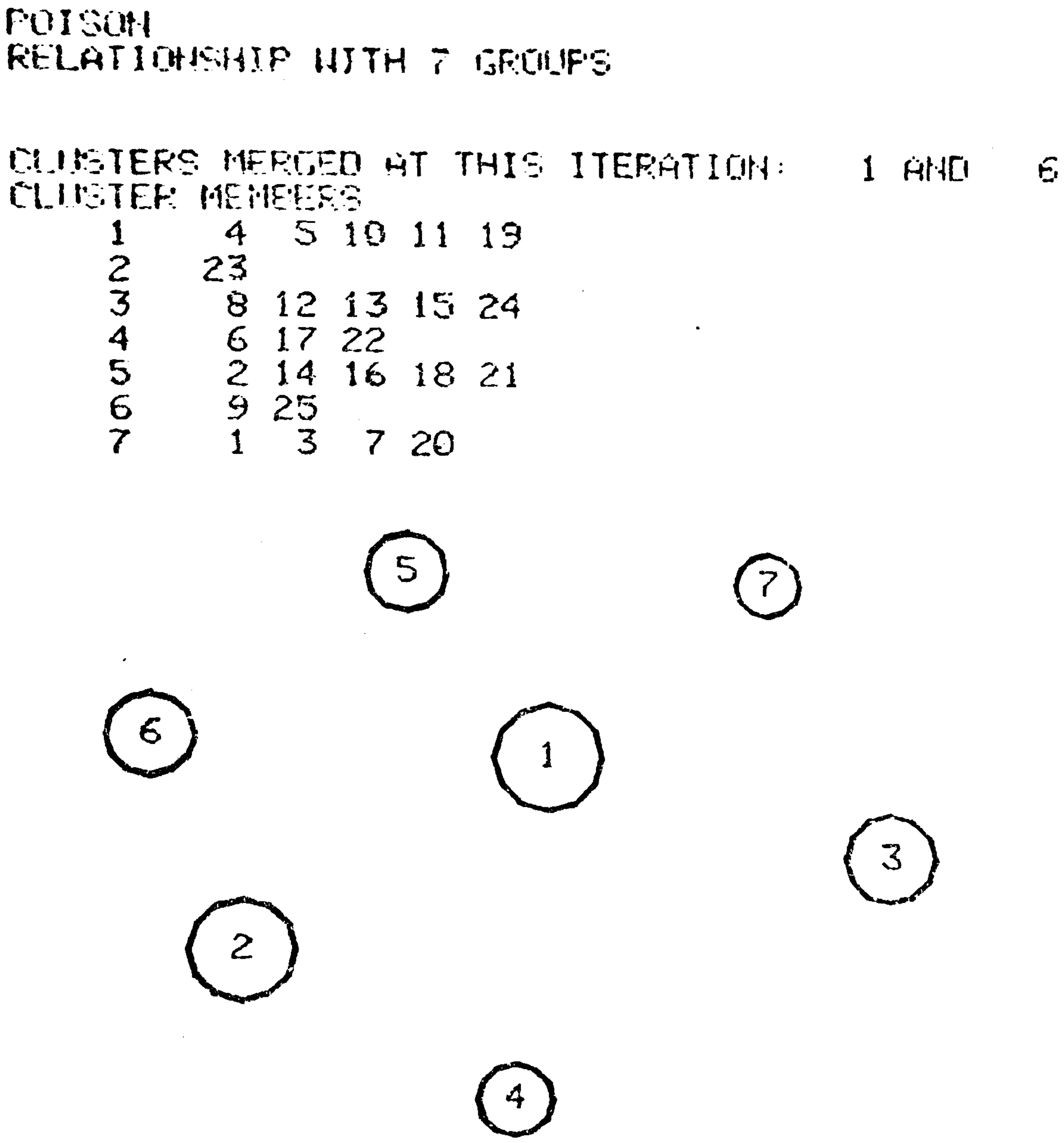
Figure 3.3

Title page and master menu

3.4.1 CLUS (Euclidean Cluster Analysis)

On entering this section the user is requested to specify an initial and terminal number of clusters required. The screen is then automatically erased and the first bubble diagram display is drawn. This either shows the data segregated into the requested number of clusters, or into a smaller number of clusters which represent the maximum number of discrete clusters identifiable in the data. Cluster membership is shown in tabular form and the bubble diagram is derived from a nonlinear mapping of the matrix of cluster centre to centre distances. The size of each bubble is scaled according to the average point to centre distance of that particular cluster.

Pressing the return key restarts the program and the next display shows the clustering with (n-1) clusters (n being the number of clusters shown in the previous display). The number of clusters is successively reduced until the requested terminal number of clusters is reached. After display of this configuration pressing the return key exits this section of the program and returns to the master menu.



```
POISON
RELATIONSHIP WITH 7 GROUPS

CLUSTERS MERGED AT THIS ITERATION:    1 AND   6
CLUSTER MEMBERS
      1    4  5 10 11 19
      2   23
      3    8 12 13 15 24
      4    6 17 22
      5    2 14 16 18 21
      6    9 25
      7    1  3  7 20
```

Figure 3.4

Euclidean Cluster Analysis. Output showing clustering into seven groups.

3.4.2 HCLUS (Hierarchical Cluster Analysis)

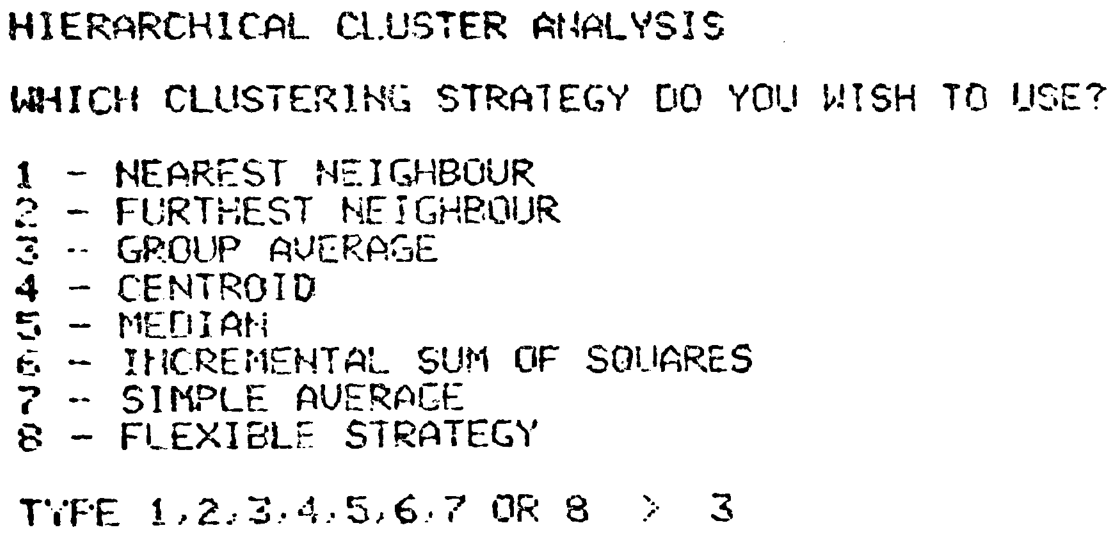
On entering this section a table of available clustering strategies is displayed as follows:

1. NEAREST NEIGHBOUR
2. FURTHEST NEIGHBOUR
3. GROUP AVERAGE
4. CENTROID
5. MEDIAN
6. INCREMENTAL SUM OF SQUARES
7. SIMPLE AVERAGE
8. FLEXIBLE STRATEGY

Typing the appropriate number then selects the required clustering strategy. If the Flexible Strategy is requested then a "Beta coefficient" must be input. Using this coefficient the characteristics of the clustering strategy can be made to range from space-dilating ($\beta = -1$) to space-contracting ($\beta = 1$).

The page is then erased and the pairing sequence displayed. The accuracy of fit measure shown is an adaptation of the cophenetic correlation coefficient.

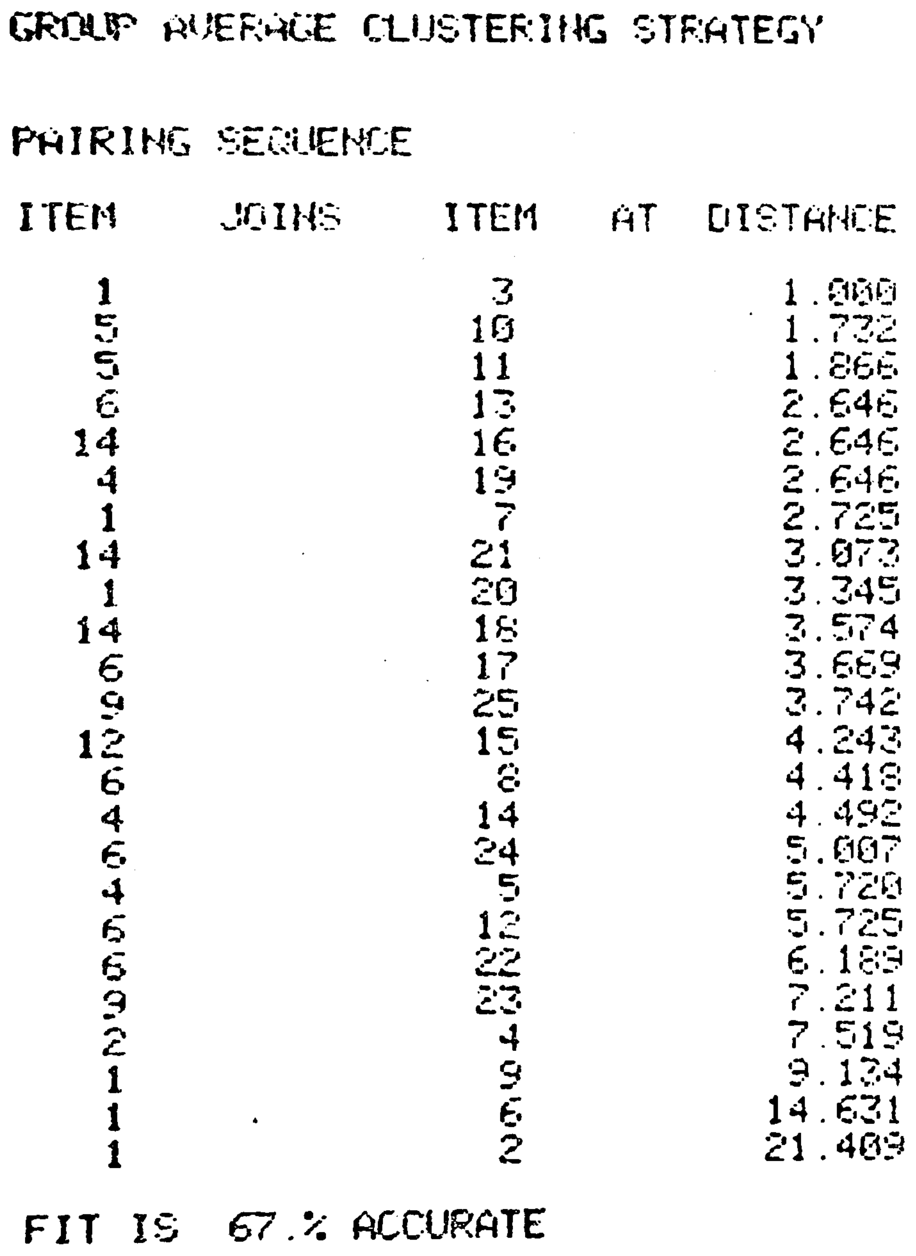
Pressing the return key erases the page and a dendrogram representation of the pairing sequence is drawn. To continue press the return key again. If a further analysis is requested the range of strategies is displayed again. If no alternative clusterings are required the program returns to the master menu.



```
HIERARCHICAL CLUSTER ANALYSIS

WHICH CLUSTERING STRATEGY DO YOU WISH TO USE?

1 - NEAREST NEIGHBOUR
2 - FURTHEST NEIGHBOUR
3 - GROUP AVERAGE
4 - CENTROID
5 - MEDIAN
6 - INCREMENTAL SUM OF SQUARES
7 - SIMPLE AVERAGE
8 - FLEXIBLE STRATEGY

TYPE 1,2,3,4,5,6,7 OR 8  >  3
```

Figure 3.5

Hierarchical Cluster Analysis options



```
GROUP AVERAGE CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT   DISTANCE

  1                  3            1.000
  5                 10            1.732
  5                 11            1.866
  6                 13            2.646
 14                 16            2.646
  4                 19            2.646
  1                  7            2.725
 14                 21            3.073
  1                 20            3.345
 14                 18            3.574
  6                 17            3.669
  9                 25            3.742
 12                 15            4.243
  6                  8            4.418
  4                 14            4.492
  6                 24            5.007
  4                  5            5.720
  6                 12            5.725
  6                 22            6.189
  9                 23            7.211
  2                  4            7.519
  1                  9            9.134
  1                  6           14.631
  1                  2           21.409

FIT IS  67.% ACCURATE
```

Figure 3.6

Hierarchical Cluster Analysis - Pairing Sequence

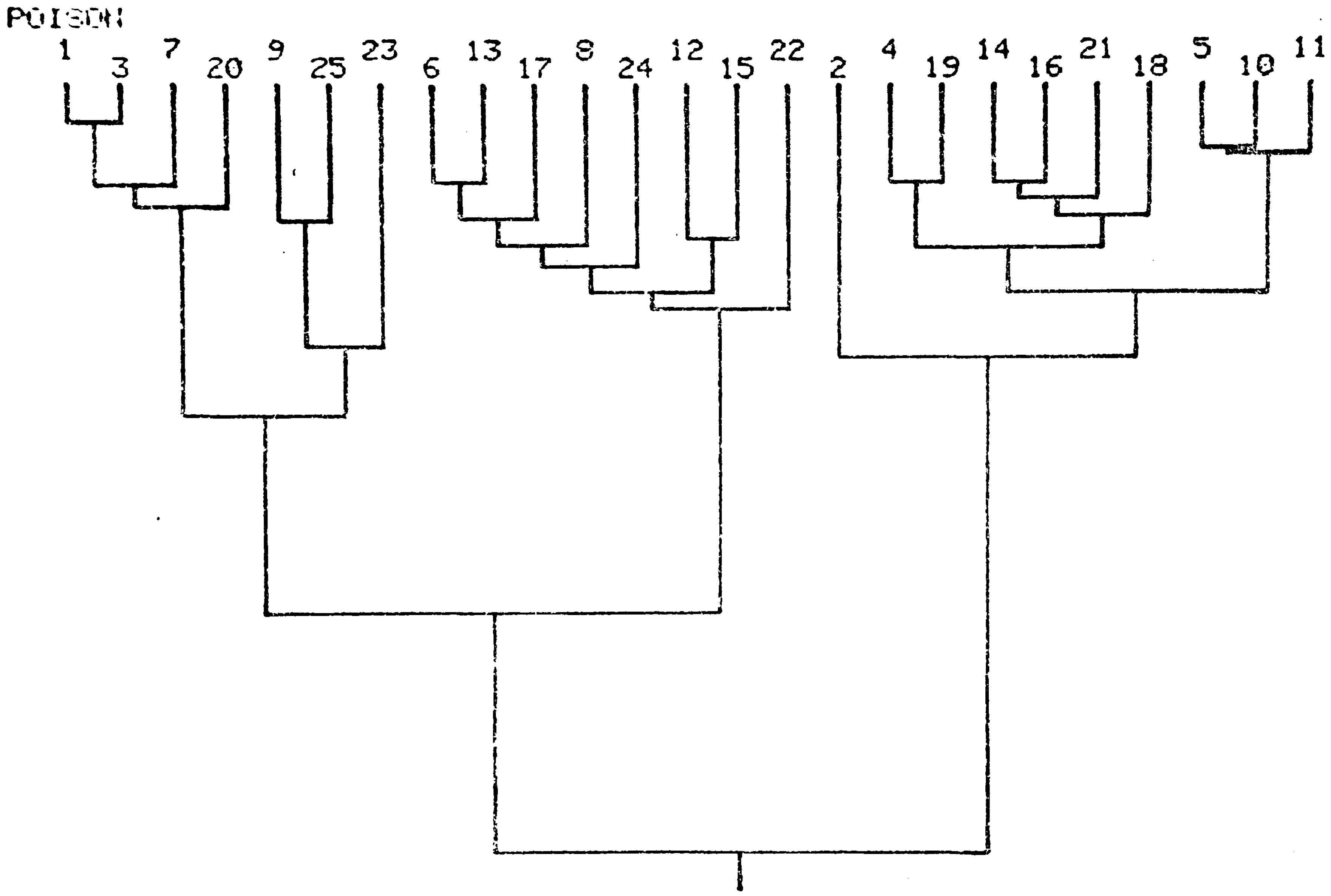


Figure 3.7

Hierarchical Cluster Analysis - Dendrogram

3.4.3 NLMAP (Nonlinear Mapping Analysis)

This section is controlled by its own menu as follows:

DRAW

MOVE

LRFLIP

TBFLIP

ROTATE

ASCALE

COMPAR

NAMES

SAVE

UNSAVE

INICON

/DRAW

EXIT

BDRAW draws the bubble diagram representation. Bubble size is relative to area (if specified in the data file).

MOVE enables bubbles to be moved interactively using the crosswire cursor. After selecting MOVE the cursor is displayed. To move a bubble first point to the bubble and press any key, then to the location of the new bubble centre and press any key.

LRFLIP flips the display from left to right

TBFLIP flips the display from top to bottom

ROTATE requests an angle of rotation, erases the display and redraws at the required orientation.

ASCALE alters the scale of the display.

COMPAR enables the comparison of the current bubble diagram with a previously stored diagram (or the comparison of two stored diagrams).

NAMES displays bubble names.

SAVE writes the current bubble configuration to file for future reference.

UNSAVE retrieves a previously stored configuration.

INICON performs a nonlinear mapping with the currently displayed configuration as the initial configuration of the calculation. In conjunction with BMOVE this command can check against solutions being found in local rather than global minima.

/DRAW erases and redraws the current display. Used for tidying up after BMOVE.

EXIT returns to the master menu.

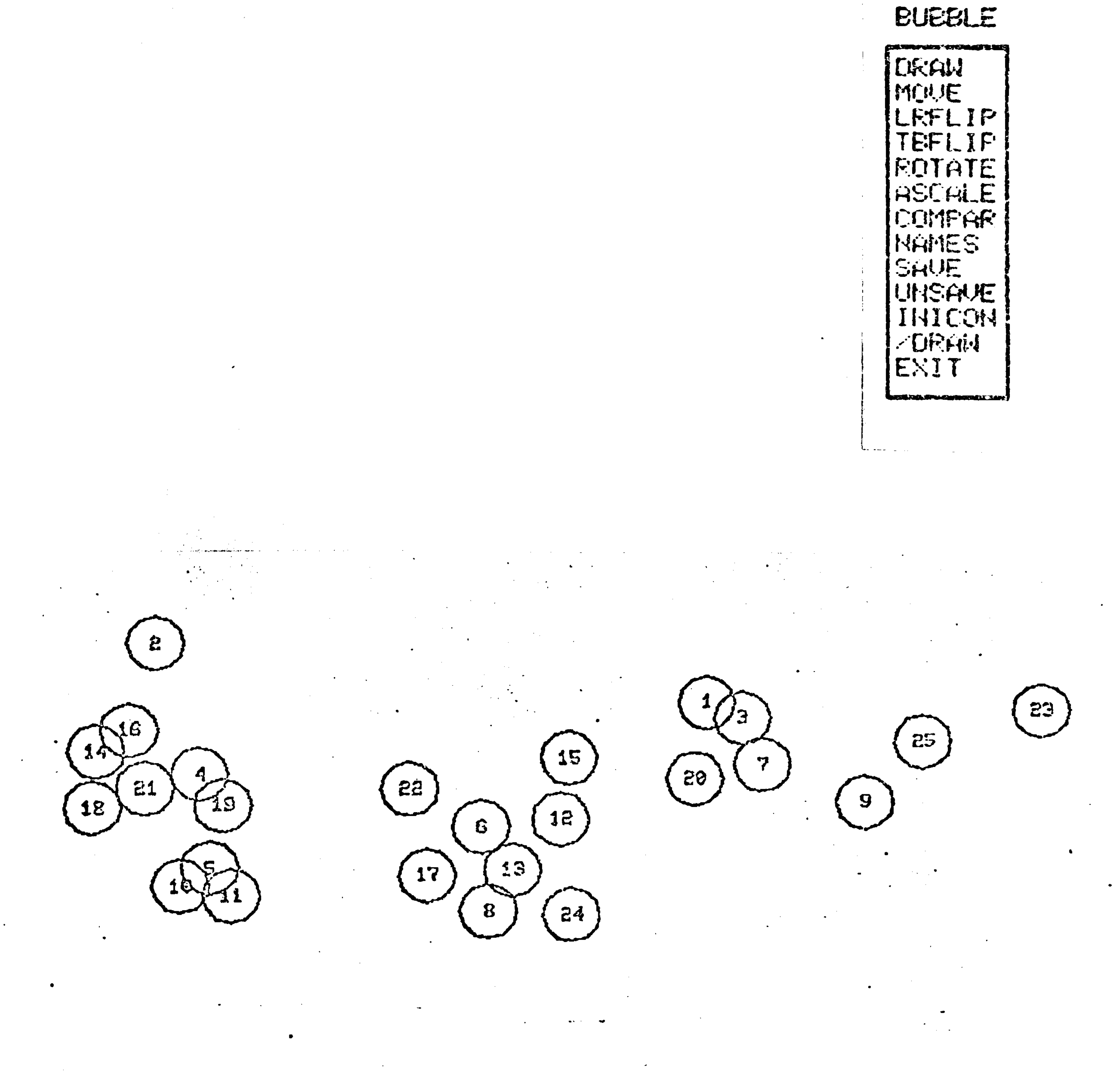


Figure 3.8

Nonlinear Mapping Analysis

```
CONFIGURATION FITTING
MATRIX A WILL BE FITTED TO MATRIX B
MATRIX A IS T2
INPUT FILE FOR MATRIX B > T1
```

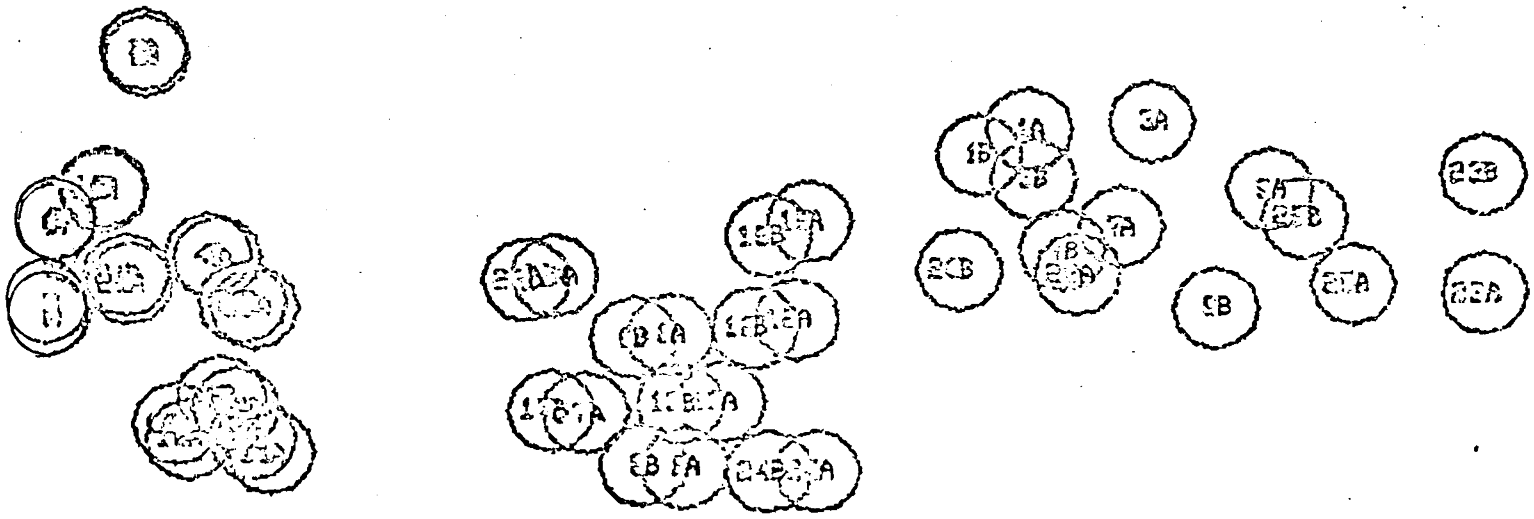


Figure 3.9

Procrustes Comparison

3.4.4 PCOORD (Principal Coordinates Mapping Analysis)

This analysis allows output of the results in two- or three-dimensional mappings. The two dimensional display is comparable with the nonlinear mapping. The three dimensional display is presented in the form of a plan and two elevations.

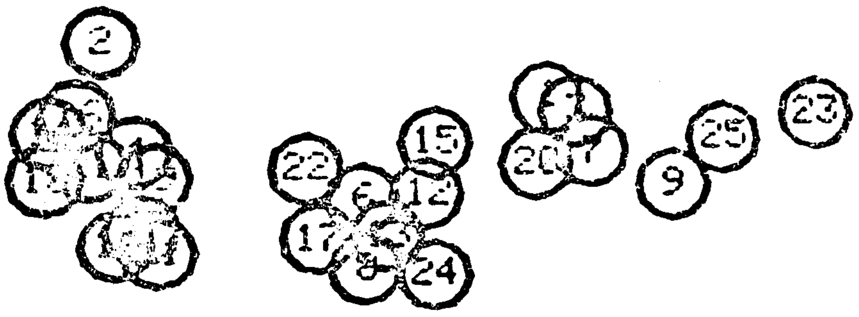


Figure 3.10

2-D Principal Coordinates Mapping

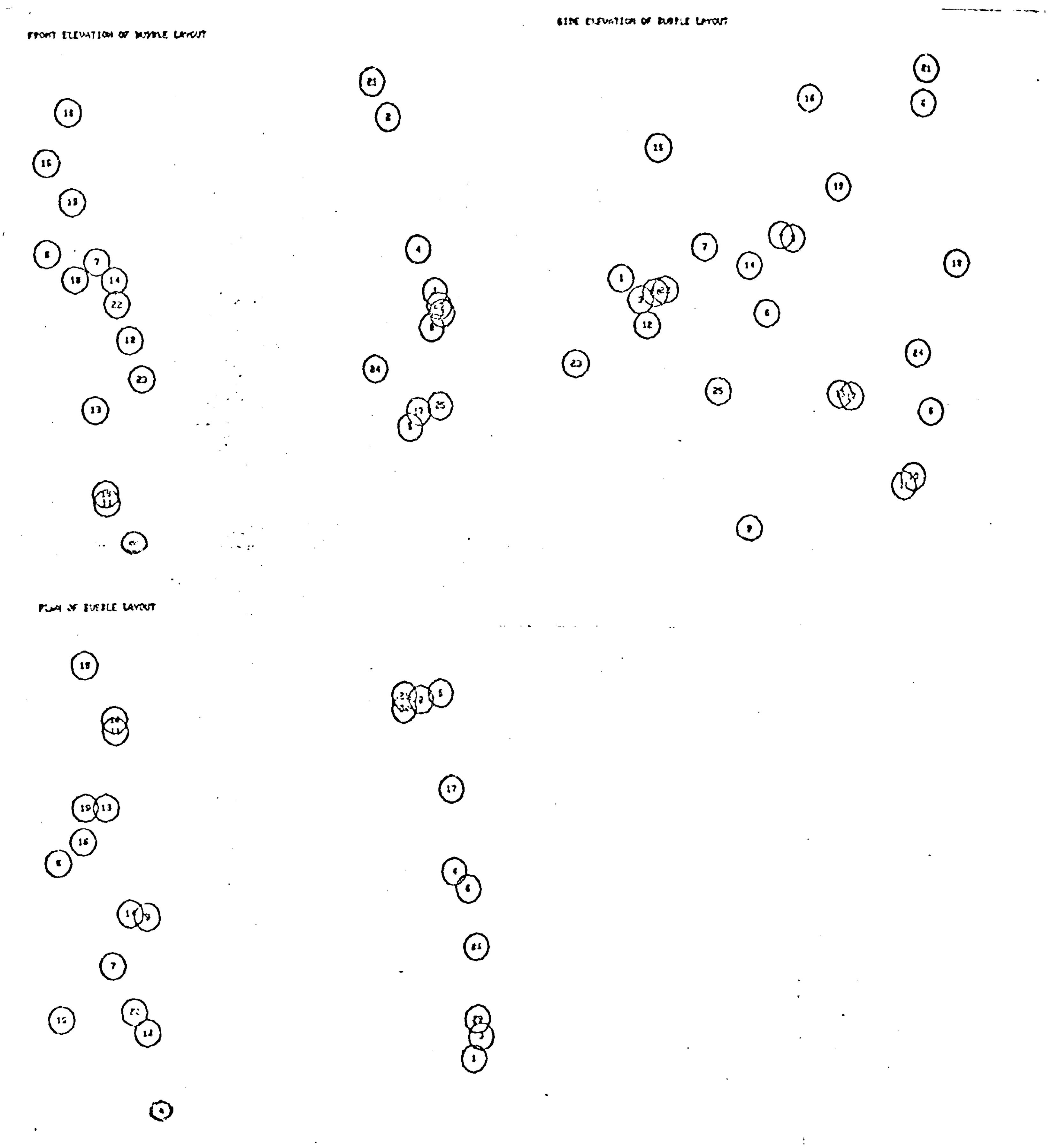
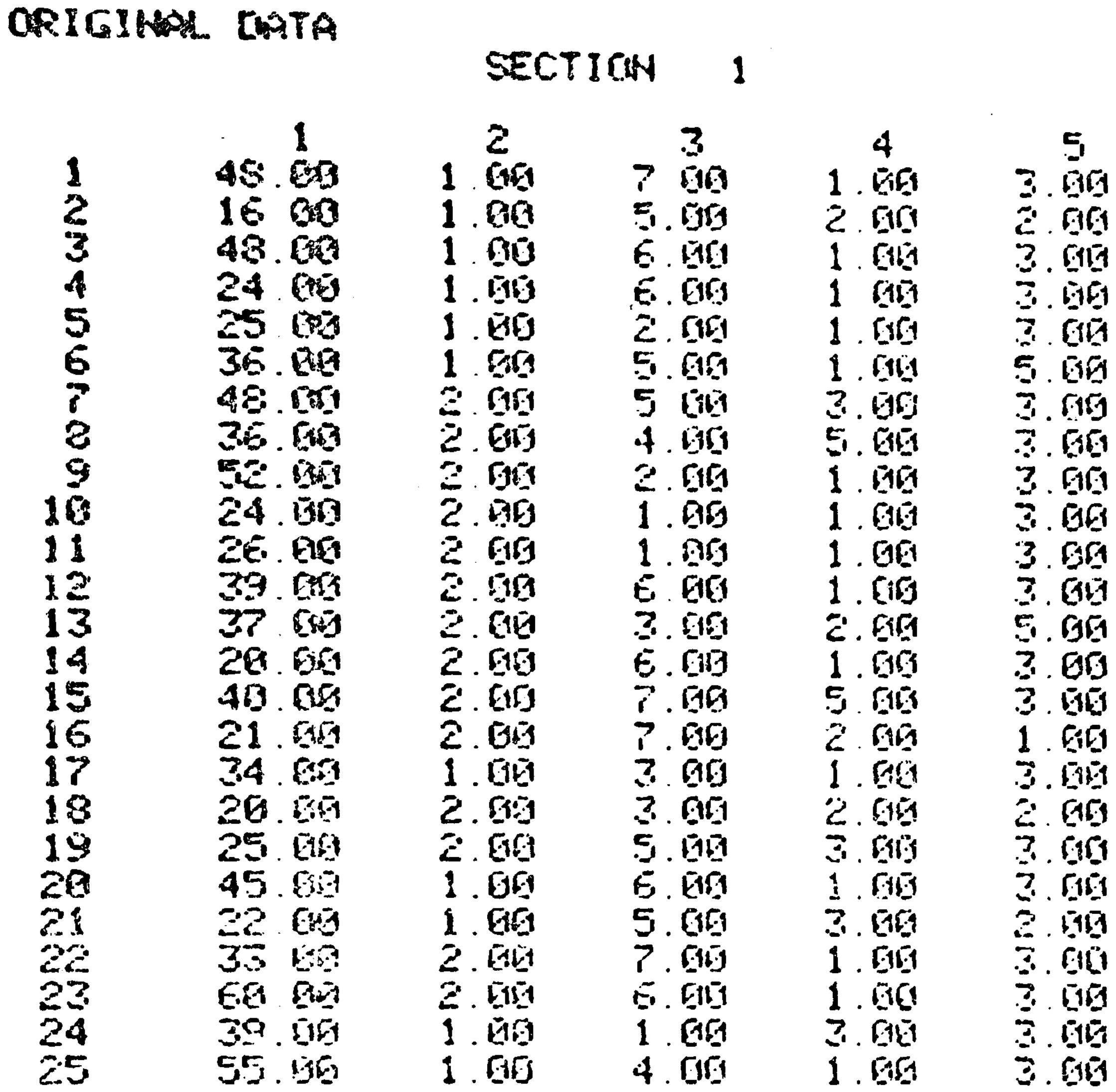


Figure 3.11

3-D Principal Coordinates Mapping

3.4.5 ROWCOL (Reordering Of Rows And Columns)

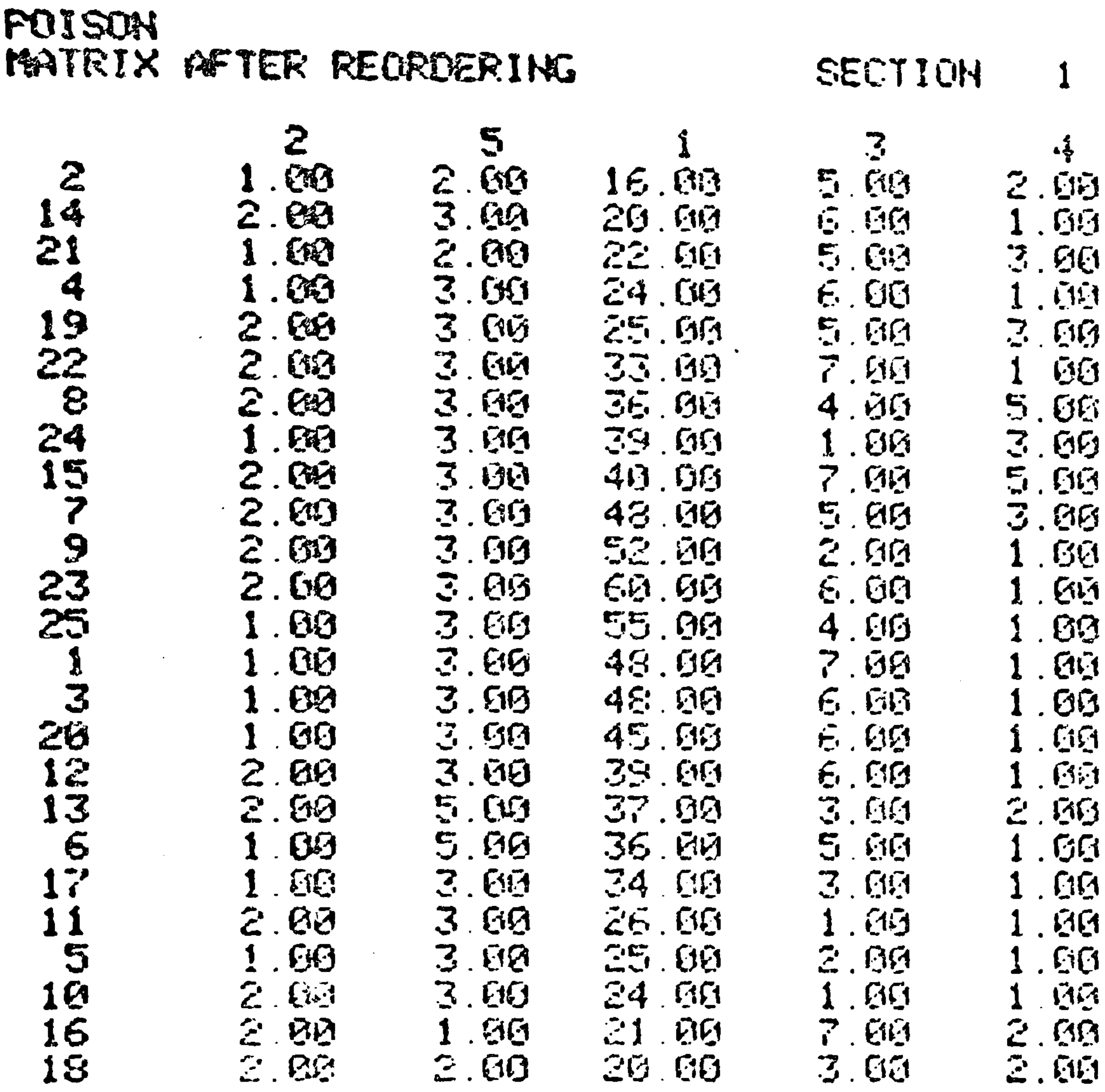
Reorders rows and columns of the data matrix to cluster large numbers around the main diagonal.



```
ORIGINAL DATA
                         SECTION    1

              1          2          3          4          5
  1         48.00      1.00       7.00       1.00       3.00
  2         16.00      1.00       5.00       2.00       2.00
  3         48.00      1.00       6.00       1.00       3.00
  4         24.00      1.00       6.00       1.00       3.00
  5         25.00      1.00       2.00       1.00       3.00
  6         36.00      1.00       5.00       1.00       5.00
  7         48.00      2.00       5.00       3.00       3.00
  8         36.00      2.00       4.00       5.00       3.00
  9         52.00      2.00       2.00       1.00       3.00
 10         24.00      2.00       1.00       1.00       3.00
 11         26.00      2.00       1.00       1.00       3.00
 12         39.00      2.00       6.00       1.00       3.00
 13         37.00      2.00       3.00       2.00       5.00
 14         20.00      2.00       6.00       1.00       3.00
 15         40.00      2.00       7.00       5.00       3.00
 16         21.00      2.00       7.00       2.00       1.00
 17         34.00      1.00       3.00       1.00       3.00
 18         20.00      2.00       3.00       2.00       2.00
 19         25.00      2.00       5.00       3.00       3.00
 20         45.00      1.00       6.00       1.00       3.00
 21         22.00      1.00       5.00       3.00       2.00
 22         33.00      2.00       7.00       1.00       3.00
 23         60.00      2.00       6.00       1.00       3.00
 24         39.00      1.00       1.00       3.00       3.00
 25         55.00      1.00       4.00       1.00       3.00
```

Figure 3.12

Original Data Matrix



```
POISON
MATRIX AFTER REORDERING                      SECTION    1

                2        5        1          3          4
   2         1.00     2.00    16.00       5.00       2.00
  14         2.00     3.00    20.00       6.00       1.00
  21         1.00     2.00    22.00       5.00       3.00
   4         1.00     3.00    24.00       6.00       1.00
  19         2.00     3.00    25.00       5.00       3.00
  22         2.00     3.00    33.00       7.00       1.00
   8         2.00     3.00    36.00       4.00       5.00
  24         1.00     3.00    39.00       1.00       3.00
  15         2.00     3.00    40.00       7.00       5.00
   7         2.00     3.00    48.00       5.00       3.00
   9         2.00     3.00    52.00       2.00       1.00
  23         2.00     3.00    60.00       6.00       1.00
  25         1.00     3.00    55.00       4.00       1.00
   1         1.00     3.00    48.00       7.00       1.00
   3         1.00     3.00    48.00       6.00       1.00
  20         1.00     3.00    45.00       6.00       1.00
  12         2.00     3.00    39.00       6.00       1.00
  13         2.00     5.00    37.00       3.00       2.00
   6         1.00     5.00    36.00       5.00       1.00
  17         1.00     3.00    34.00       3.00       1.00
  11         2.00     3.00    26.00       1.00       1.00
   5         1.00     3.00    25.00       2.00       1.00
  10         2.00     3.00    24.00       1.00       1.00
  16         2.00     1.00    21.00       7.00       2.00
  18         2.00     2.00    20.00       3.00       2.00
```

Figure 3.13

Reordered Data Matrix

CHAPTER 4

CLASSIFICATION

4.1 INTRODUCTION

Classification is essentially the identification of groups of similar activities from the set of activities being studied. Two approaches to classification are possible - the identification of groupings, termed classification proper, and the allocation of activities to existing groups, termed discrimination. As MAGIC is used almost exclusively in an exploratory data analysis form discrimination techniques are not of relevance here.

Classification may be further subdivided into different classificatory procedures, which may include the simplification of data by ordination. Clustering methods tend to emphasise discontinuities, whereas ordination methods display the continuity of the data. Prior to ordination the activities being considered are assigned to positions in a multidimensional space defined by their properties or some measure of their dissimilarity. Efforts are then made to express the relationships between the activities in fewer dimensions than those originally considered. Ordination is, however, here considered seperately from classification and dealt with in detail in chapter 7.

4.2 TYPES OF CLASSIFICATION

Many classification methods exist, and a "classification of classifications" is shown in figure 4.1.

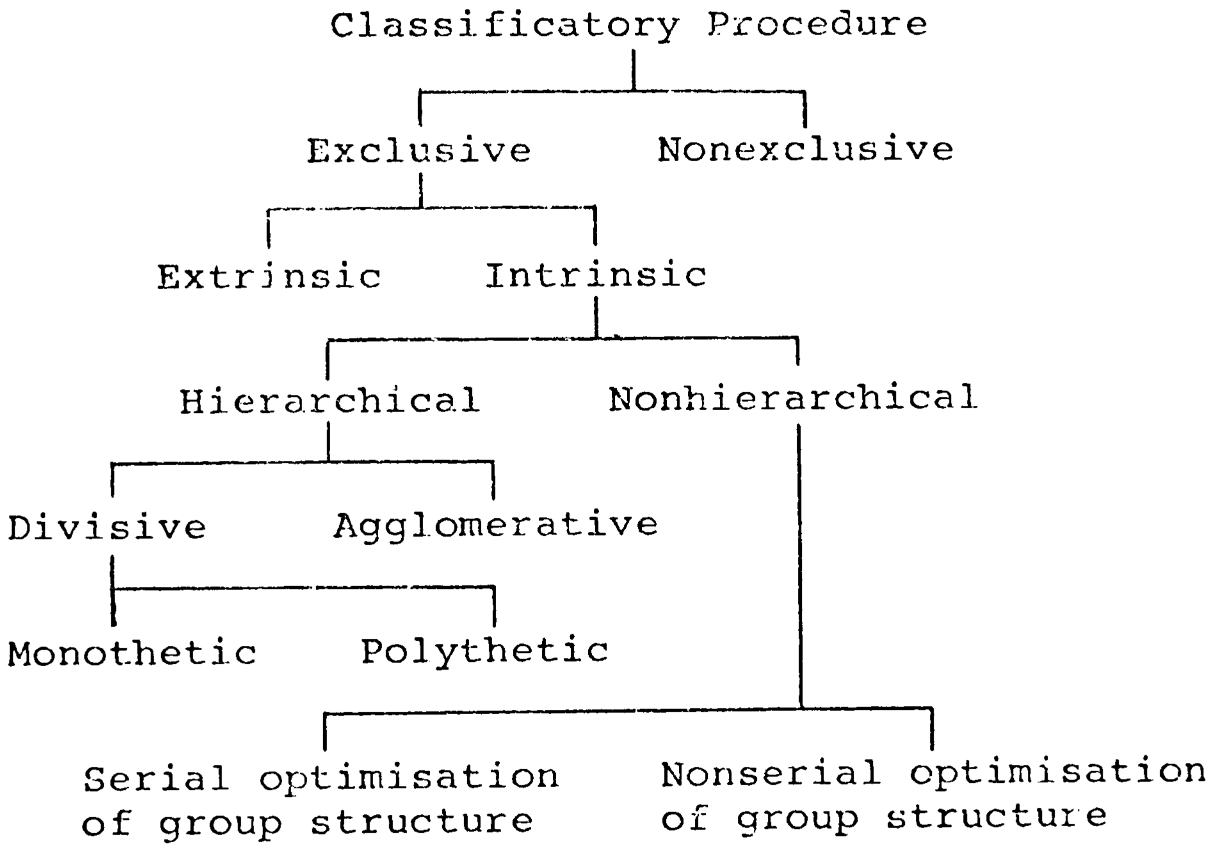


Figure 4.1

Relationship between classificatory procedures

The first differentiating feature is whether the method is exclusive or nonexclusive in its treatment of individual activities. In an exclusive (or nonoverlapping) classification a given element can occur in one and only one subset; in the nonexclusive (or overlapping) case the same element may occur in more than one subset. Nonexclusive classification methods are of use in library catalogues and information retrieval systems (a book may appear under several different subject headings) or medical statistics (a single patient may suffer from more than one disease), but in attempting to simplify architectural data the exclusive methods are preferred. The purpose of the classification sections in MAGIC is to seperate the activities. The ordination techniques look at the continuities and overlaps in the data, and so nonexclusive classifications will not be considered further.

The exclusive classifications may themselves be technically divided into extrinsic and intrinsic methods. Formally, intrinsic classifications are used to derive groups solely from their attributes. Extrinsic methods attempt to form clusterings on (n-1) attributes to "explain" the nth attribute. MAGIC only considers intrinsic classifications.

Intrinsic classifications may be hierarchical or nonhierarchical. In a nonhierarchical classification groups are selected such that each is individually as homogeneous as possible. In the hierarchical case groups are considered in pairs, as possible candidates for fusion; and the criterion for fusion is that the decrease in homogeneity on fusion shall be as small as possible. This is usually formally expressed by saying the nonhierarchical classification optimises the internal properties of subsets; a hierarchical classification optimises a route between individuals and the complete population. No such route between groups and their constituent individuals (enabling examination of the group infrastructure), or between groups and the complete population is provided by nonhierarchical clustering. However, there are several applications in which homogeneity of groups is of prime importance, and the nonhierarchical strategy, as well as the more developed (computationally) hierarchical techniques, is included in MAGIC. Hierarchical, nonoverlapping classification produces groups, or clusters, whose relationships to one another are readily expressed in two-dimensions, generally in the form of a dendrogram.

The clusters arise as a consequence of the methodology adopted to establish the hierarchy and do not necessarily exhibit the same homogeneity. In contrast nonhierarchical methods can produce clusters of defined heterogeneity but do not link them together in any systematic framework. The nonhierarchical techniques are relatively undeveloped and MAGIC introduces a major advance in, firstly, cycling through a number of iterations with successively fewer clusters, and secondly, mapping the cluster relationships into a two-dimensional display. Both of these techniques introduce some method of interpreting relationships in a systematic framework.

Hierarchical methods may be further divided into agglomerative or divisive techniques. In an agglomerative classification the individuals are progressively fused into subsets of increasing size until the entire population is in a single set; in a divisive classification the whole population of elements is progressively subdivided until an acceptable degree of subdivision is attained. Agglomerative techniques are computationally much the more efficient and the hierarchical strategies used in MAGIC are all agglomerative.

Finally, hierarchical techniques may be monothetic or polythetic. In a monothetic classification clustering is effected by reference to a single attribute of maximal information-content. In the polythetic case all attributes are of equal importance. Agglomerative strategies are always polythetic.

4.3 TYPES OF DATA

The standard data structure used in MAGIC assumes a number of activities (in statistical terminology "operational taxonomic units") on each of which a number of variables (properties or characteristics) is measured. The variables may, in principle, take values in any space, but in practice there are three types of variables of relevance to architectural data, and these may be classified according to the nature of their underlying scale:

(i) Binary - the taking of one of two contrasting states, such as the presence or absence of a particular characteristic.

(ii) Multistate - determined by an ordered classification in a hierarchy of contrasting forms which encompass the total variation in the range of entities under study. An example of this "ordinal" scale would be the grouping of activities according to whether they required full, partial or no blackout facilities.

(iii) Continuous - measures on a continuous scale, as with attributes such as temperature, distance, etc.

A given set of data may be mixed (contain variables of different types), it may be heterogeneous (variables of the same type but of different scales, such as temperature and distance), or it may be homogeneous (variables measured on the same scale, such as a simple distance matrix). There are a number of techniques for transforming variables of one type to another, or converting all variables to a standard scale, but all these methods rely on measures of similarity or distance, for, in order to cluster variables it is necessary to have some numerical similarity measurements to characterise the relationships among the variables. The conventional approach to this requirement is to compute a measure of association for every pairwise combination of variables; in a problem with n variables this results in n/2(n-1) different pairs. The next section considers the range of different measures of association among variables.

4.4 MEASURES OF SIMILARITY AND DIFFERENCE

Similarity and difference are mutually dependent concepts and, in much of the classical statistical literature, the former term applies to both. The majority of clustering techniques begin with the calculation of a matrix of similarities or differences between activities, and, therefore, consideration is needed of the possible ways of defining these quantities.

A wide variety of interentity similarity measures have been proposed but relatively few are in current use. The restriction in number has resulted from several causes. Many of the neglected indices are mere variants of others and have similar properties; others are highly specialised; and others display unfavourable mathematical qualities.

Some of the measures discussed below estimate dissimilarity rather than similarity, but since the two are complimentary concepts this need not cause any confusion. The reason for stressing dissimilarity in certain situations is that such measures are readily envisaged as "distances apart".

4.4.1 Coefficients Of Similarity

Similarity coefficients have a long history and, in the older literature, were usually known as association or correlation coefficients. A similarity coefficient measures the relationship between two entities, given the values of a set of variates common to both. With most of the coefficients values range from zero (no similarity) to unity (complete similarity).

A great number of similarity coefficients are known, and the most common have been listed and defined by Goodman and Kruskal (1954,1959), Sokal and Sneath (1963), and Sneath and Sokal (1973). Many of the coefficients were developed to accommodate particular forms of data, as for example, those restricted to binary data. Others allow for particular distributions of the properties measured, or minimise the influence of large or small values in the data. In some instances each of these considerations may influence the choice of index, but it is emphasised that each stress a particular property of the data and that all indices are not interchangeable. Indeed they do not all necessarily yield similar results when the entities whose similarities they measure are clustered. The various main measures applying to each data type are briefly discussed below.

4.4.2 Similarity Measures Applying To Binary Data

To facilitate the comparison of the coefficients for binary data a standard nomenclature will be adopted. Consider a single binary attribute with outcomes of 1 or 0. There are only four outcomes possible when comparing two activities. These are that both activities record the attribute in the first state (1,1) or the alternative state (0,0), or that one activity records one state and the other records the alternative, i.e. (0,1), (1,0). For a number of activities the summated values of each of the four possibilities may be calculated. The values may be summarised in a two-way association table (figure 4.2).

		Activity j			
		1	0		
	1	(1,1) a	(1,0) b	a+b	
Activity i	0	(0,1) c	(0,0) d	c+d	
		a+c	b+d	n	(n=a+b+c+d)

Figure 4.2

Association table for binary data

Here the letters a,b,c,d refer to the summated number of attributes. That is, a represents the number of attributes in one state (1,1) shared by both activities; b is the number of attributes for which the joint score is (1,0), the number possessed by the first activity but not the second; c the number possessed by the second but not the first (0,1); and d the number possessed by both activities in the alternative state (0,0). The sum, n = a+b+c+d , is the total number of attributes for which the entities have been compared.

The status of d in figure 4.2 is ambiguous. In most circumstances it would seem ridiculous to regard two activities as similar largely on the basis of them both lacking something. In certain other circumstances it would seem improper to neglect conjoint absences when estimating similarity. In order to resolve these difficulties similarity coefficients with and without the inclusion of d have been designed and each group is considered below. In the similarity measure finally included in MAGIC it is possible to include or discount d, although the default built into the program discounts it.

Table 4.3 provides a summary of the various measures along with names traditionally associated with them. Every mechanically derived combination is included in the table even though five possibilities appear to be worthless. The fourteen measures are discussed individually below. Except where noted otherwise the range assumed by each measure is (0 to 1).

(a) Equal weighting of matches, mismatches

0-0 matches in denominator	0-0 matches in numerator	
	Excluded	Included
Included	1 Russell and Rao $\frac{a}{a+b+c+d} = \frac{a}{n}$	2 Simple matching $\frac{a+d}{a+b+c+d} = \frac{a+d}{n}$
Excluded	3 Jaccard $\frac{a}{a+b+c}$	4 Nonsense $\frac{a+d}{a+b+c}$

(b) Double weight for matched pairs

0-0 matches in denominator	0-0 matches in numerator	
	Excluded	Included
Included	5 Not recommended $\frac{2a}{2(a+d)+b+c}$	6 Sokal and Sneath $\frac{2(a+d)}{2(a+d)+b+c}$
Excluded	7 Dice $\frac{2a}{2a+b+c}$	8 Nonsense $\frac{2(a+d)}{2a+b+c}$

Figure 4.3 (i)

Matching coefficients

(c) Double weight for unmatched pairs

0-0 matches in denominator		0-0 matches in numerator		
		Excluded		Included
Included	9	Not recommended $\frac{a}{a+d+2(b+c)}$	10	Rogers-Tanimoto $\frac{a+d}{a+d+2(b+c)}$
Excluded	11	Sokal and Sneath $\frac{a}{a+2(b+c)}$	12	Nonsense $\frac{a+d}{a+2(b+c)}$

(d) Matched pairs excluded from denominator

		0-0 matches in numerator		
		Excluded		Included
	13	Kulczynski $\frac{a}{b+c}$	14	Unnamed $\frac{a+d}{b+c}$

Figure 4.3 (ii)

Matching coefficients (continued)

Coefficient 1. The value of this measure is the probability that a randomly chosen data unit will score 1 on both variables. It excludes 0-0 matches as irrelevant in counting the number of times the two variables match (the numerator) but does count 0-0 matches in determining the number of possibilities for a match (the denominator).

Coefficient 2. The value of this measure is the probability of a randomly chosen data unit achieving the same score on both variables. The 0-0 matches are given full weight.

Coefficient 3. The value of this measure is the conditional probability that a randomly chosen data unit will score 1 on both variables, given that data units with 0-0 matches are discarded first. The 0-0 matches are treated as being totally irrelevant.

Coefficients 4, 8 and 12. These measures treat the 0-0 matches as relevant in the numerator but exclude such matches in the denominator. Since the numerator usually can be viewed as the number of relevant possibilities fulfilled, it is nonsense to include in the numerator that which is specifically excluded from the denominator.

Coefficients 5 and 9. These two measures are analogous to coefficient 1 since they exclude 0-0 measures in the numerator whilst including them in the denominator. They have not appeared in the statistical literature and no obvious interpretation of them appears possible. However they do not have such obvious faults as coefficients 4,8, and 12 which might prompt their rejection.

Coefficient 6. Sokal and Sneath (1963) include this measure in their list without attribution. It may be viewed as an extension of coefficient 2 such that matched pairs are given double weight. The double weighting seems to preclude any possibility for a probabilistic interpretation.

Coefficient 7. This measure excludes 0-0 matches entirely whilst double weighting 1-1 matches. It may be viewed as an extension of coefficient 3, though the probabilistic interpretation is lost. Hall (1969, p 322) offers an alternative interpretation:

> However, for 0,1 mismatches the zero is just as trivial as in the 0,0 case. Mismatches should then lie about midway along the scale of significance between the 0,0 and 1,1 cases respectively. The number of mismatches in the coefficient should by this reasoning be multiplied by 0.5.

Clearing the 0.5 fraction then results in double weight for the 1-1 matches.

Coefficient 10. In the context of association among variables, this coefficient is best viewed as an extension of measure number 2 based on double weighting of unmatched pairs.

Coefficient 13. This measure is the ratio of matches to mismatches with 0-0 matches excluded. The range of this coefficient is 0 to ∞.

Coefficient 14. This measure is the ratio of matches to mismatches including 0-0 matches.

Coefficients 3, 7 and 11 are all monotonic to each other. To illustrate this suppose there are two tables denoted by 1 and 2 and that measure 7 gives the result

$$\frac{2a_1}{2a_1+b_1+c_1} \geqslant \frac{2a_2}{2a_2+b_2+c_2}$$

Since the table entries are all nonnegative, the fractions may be cleared to give

$$4a_1a_2+2a_1(b_2+c_2) \geqslant 4a_1a_2+2a_2(b_1+c_1)$$

Subtracting $2a_1a_2$ from both sides and dividing by 2 gives

$$a_1a_2+a_1(b_2+c_2) \geqslant a_1a_2+a_2(b_1+c_1)$$

which implies

$$\frac{a_1}{a_1+b_1+c_1} \geqslant \frac{a_2}{a_2+b_2+c_2}$$

or monotonicity with coefficient 3. Coefficients 2, 6 and 10 may similarly be shown to be monotonic to each other. This result is important because when using monotonically invariant clustering techniques (such as nearest neighbour) measures 3, 7 and 11 are equivalent to each other; measures 2, 6, and 10 are likewise equivalent.

Among the matching measures, numbers 1, 2 and 3 possess reasonably useful probabilistic interpretations. There are several additional measures with probabilistic foundations.

The quantity a/(a+b) is the conditional probability that a randomly chosen data unit scores 1 on variable B given that it scored 1 on variable A. Likewise the quantity a/(a+c) is the conditional probability of scoring a 1 on variable A given that a 1 was scored on variable B. Assuming variable A is estimated half the time and variable B the other half, the symmetric measure:

$$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

is obtained. This is the conditional probability of scoring a 1 on one variable given a score of 1 on the other. Sokal and Sneath (1963, p130) attribute this measure to Kulczynski.

The final 2x2 measure is

$$\frac{(a+d) - (b+c)}{a+b+c+d}$$

which is the probability that a randomly chosen data unit will score the same on both variables minus the probability it will score differently on the two variables. Since b+c = n-(a+d) this measure may also be written as

$$\frac{2(a+d)}{a+b+c+d} - 1$$

which is related monotonically to measures 2, 6 and 10 of the matching coefficients. Sokal and Sneath (1963) attribute this measure to Hamann. Its range is -1 to +1.

4.4.3 Similarity Measures Applying To Multistate Variables

Multistate variables may be effectively transformed into a series of binary variables (see section 4.4.2). The techniques applicable to binary data then all apply to multistate data. A further range of measures based on probability theory are also possible.

It is possible to draw up a contingency table for multistate data (figure 4.4)

		Activity j						
		Class						
		1	2	3	.	.	q	Totals
Activity i	1	n_{11}	n_{12}	n_{13}	.	.	n_{1q}	Σn_{1x}
	2	n_{21}	n_{22}	n_{23}	.	.	n_{2q}	Σn_{2x}
	3	n_{31}	n_{32}	n_{33}	.	.	n_{3q}	Σn_{3x}
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	p	n_{p1}	n_{p2}	n_{p3}	.	.	n_{pq}	Σn_{px}
Totals		Σn_{x1}	Σn_{x2}	Σn_{x3}	.	.	Σn_{xq}	

Figure 4.4

General form of contingency table for Multistate data

An n_{ij} entry in the table is the number of activities falling in the ith class of activity i, and the jth class of activity j. If all entries and marginal totals are divided by the total number of data units the table entries become frequencies (f_{ij}). It is then possible to apply the chi-square statistic, comparing the observed value in cell ij (O_{ij}) with the expected value under an hypothesis of independence $e_{ij} = \frac{n_{ix}n_{xj}}{n_{xx}}$.

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} (o_{ij} - e_{ij}) \Big/ e_{ij}$$

This is a traditional measure of association, but is of dubious value. The range of χ^2 increases without bound as the number of data units increases. A partial remedy is found in

$$\phi^2 = \chi^2 / n_{xx}$$

which is known as the mean-square contingency. However this quantity is itself dependent on the size of the table. In an attempt to norm ϕ^2 to the conventional range of 0 to 1 a number of measures have been suggested.

Sokal and Sneath (1963) give one example using the geometric mean of (p-1) and (q-1) as a norming factor to give the measure

$$T = \left[\frac{\chi^2/n_{xx}}{[(p-1)(q-1)]^{1/2}}\right]^{1/2}$$

and a further possibility is the use of the maximum value of ϕ^2 as a norming factor to give

$$C = \left[\frac{\chi^2/n_{xx}}{\min[(p-1),(q-1)]}\right]^{1/2}$$

Pearson (1926) suggested another measure based on ϕ^2

$$P = \left(\frac{\phi^2}{1+\phi^2}\right)^{1/2}$$

This measure is known as the coefficient of contingency.

None of these measures are really of use as measures of similarity. Goodman and Kruskal (1954, p740) pinpoint the major problem:

> One difficulty with the use of traditional

> measures, or any of the measures that are not given operational interpretation is that it is difficult to compare meaningfully their values for two cross classifications.

In cluster analysis meaningful comparisons among all pairwise combinations of variables are essential.

4.4.4 Similarity Measures Applying To Continuous Variables

The traditional measure of similarity most commonly used for continuous data is the product-moment correlation coefficient (r). A simple symbolic expression of this coefficient is:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(\Sigma(x-\bar{x})^2\ \Sigma(y-\bar{y})^2)^{1/2}}$$

where n is the number of activities, $\bar{x}$, $\bar{y}$, the mean values of the attributes in the activities, and x and y are the individual measurements of a given pair of attributes.

Despite the relatively widespread use of r as a similarity measure (cf Sokal and Michener 1967, Boyce 1969, Strauss et al 1973) a number of problems cast some doubt on its true value. In statistics the correlation coefficient is used to give a measure of the linear relationship between a pair of variables. However classification studies are carried out on a set of objects described by various variables or on a set of variables describing various objects. If the correlation coefficient is used to compare two objects it seems difficult in general to give any interpretation to a term like $\bar{x}$ which involves summing over the variables describing a single object. Furthermore it is not obvious whether two objects approximately satisfying such a linear relationship should necessarily be regarded as very similar to one another. Eades (1965) provides statistical evidence of its indeterminancy as well as its theoretical problems.

The better measure is the use of Euclidean distance as a dissimilarity measure. In essence this is the distance between two activities whose positions are determined with respect to their coordinates, these being defined by reference to a set of Cartesian axes. It is a dissimilarity measure which can be applied to both binary and continuous data. With respect to any given attribute the Euclidean distance (D) between two activities is $|x_1-x_2|$ where x_1 is the score for one activity and x_2 that for the other. For n attributes

$$D = (\sum(x_1-x_2)^2)^{1/2}$$

where x_1 and x_2 are successively the scores for the n attributes.

To ensure that attribute scores are additive it is common practice to use D^2 rather than D as the measure of dissimilarity. In certain cases however D^2 on binary data is not fully metric in that it may fail to satisfy the "triangal inequality" (metrics are discussed further in 4.5.1). A further point to be observed involves the scale of the axes of x_1 and x_2 in figure 4.5.

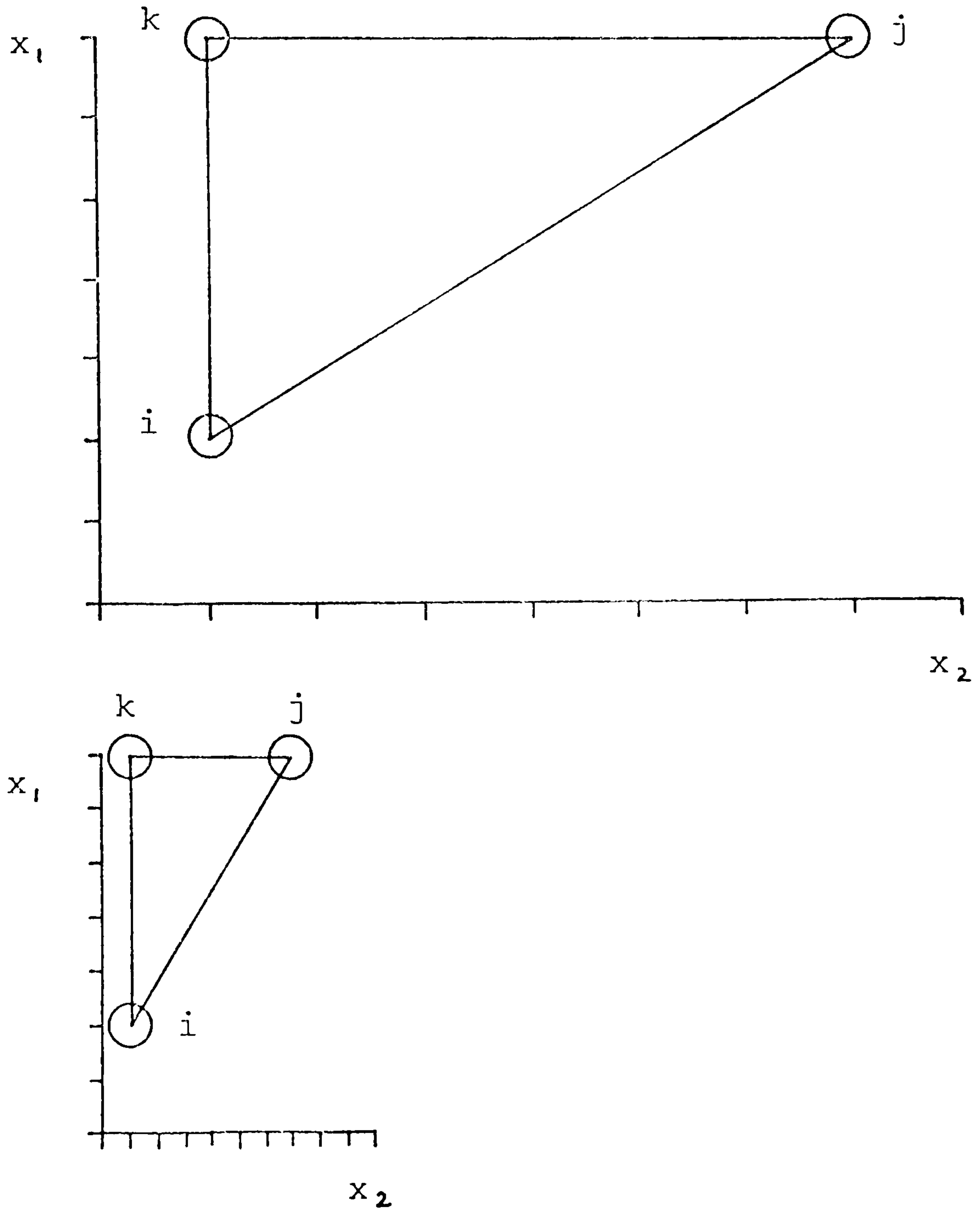


Figure 4.5

Euclidean distance coefficient - effect of scale

The scale will obviously affect the distance between points, and, since scales of measurement are quite often arbitary some standardisation process is often adopted. The convention most commonly used is to give each variable equal weight by transforming observed values so that each variable has zero mean and unit variance:

$$z_{ij} = (x_{ij} - \bar{x}_j)/s_j$$

where z_{ij} is the standard score, equivalent of x_{ij}, the observed score of activity i on variable j, $\bar{x}_j$ is the mean value of observations on variable j, and s_j the standard deviation of variable j.

4.5 STRATEGIES FOR MIXED VARIABLE DATA SETS

It is rare in real-life situations to have attributes all of the same type; it is therefore imperative that some means of combining data with different attributes be available. Indeed, one of the major criticisms of much work in the field of early stage design analysis is the reliance on a single measure of cost or distance as the sole criterion of planning efficiency. The previous discussion includes no provision for measuring association between variables of different types, much less the more difficult problem of obtaining a consistent measure across all pairwise combinations of variables in a mixed data set. A variety of difficulties surround this problem, but there are a number of workable strategies for dealing with mixed variable types.

4.5.1 Partitioning Of Variables

Perhaps the most obvious approach is to partition the variables into types and confine the analysis to the dominant type. The question of which type is "dominant" is a matter of judgement and may depend on such factors as the number of variables of each type, the variables considered most important to the analysis, relevant theory, and so on. In one way or another many statistical analyses are restricted to avoid the problems of heterogeneous data sets. Often the problem is formulated at the outset in terms of only one variable type. Up until the publication of Sokal and Sneath's classic text in 1963 the majority of classificatory strategies were designed to operate on a single type of data, usually binary.

A logical extension of this approach is to partition the variables into types and perform seperate independent analyses for each type. Gower (1971) has developed a technique applicable to binary, multistate and continuous scales, and this is discussed in some detail in section 4.5.3.

4.5.2 Conversion Of Variables

Another possible solution is to transform a set of mixed variables into a new set of variables, all of a single type: but which variable type should be chosen? From a practical point of view, this choice may be determined by which variable type is most numerous in the data set and the relative effort required for each kind of conversion. However, the conversion of all data to binary variables is the most generally useful.

Conversion to binary variables permits use of a wide array of association measures, many of which have probabilistic interpretations. Also, the use of binary variables may enable substantial compression of storage and increased computational efficiency. The problem is how to dichotomise all the variables that are not already in binary form. For multistate variables the problem is a special case of interval to nominal conversion. For continuous variables it is a problem of fixing a division point.

For example, consider a multistate variable with four categories; A, B, C, and D. If the analysis is to be carried out in terms of binary variables, there are seven alternative dichotomies:

1. (A) (B,C,D)
2. (B) (A,C,D)
3. (C) (A,B,D)
4. (D) (A,B,C)
5. (A,B) (C,D)
6. (A,C) (B,D)
7. (A,D) (B,C)

In effect, a single variable is given a multidimensional representation. This, of course, causes considerable growth in the size of the problem.

The conversion of continuous data to binary form is even less satisfactory. Any distribution may be arbitrarily divided into two sections, thereby being converted to a binary attribute. The disadvantage of this conversion may be seen by considering a normally distributed variable with the mean taken as the dividing line. In this case two entities differing only slightly from one another but placed on either side of the mean become equally dissimilar to a pair drawn from the extremes of the range. Considered another way, all entities on either side of the mean acquire identical binary scores.

An alternative strategy for dealing with mixed variable data sets is to use a set of measures which are compatible with each other and collectively cover every pairwise combination of variables. This is extremely restrictive in practice.

4.5.3 Gower's General Coefficient Of Similarity

The problems of handling mixed data have been particularly studied by Goodall (1966), Lance and Williams (1967), Gower (1967), and Burr (1968). Gower's work, developed in later publications (Gower 1971) is of particular interest and is the method adopted in MAGIC for dealing with mixed data.

To obtain his coefficient of similarity Gower defines similarity between two activities i and j as the average score taken over all the possible comparisons:

$$S_{ij} = \sum s_{ijk} \, d_{ijk} \,/\, \sum d_{ijk}$$

Where s_{ijk} is the score on variable k for activities i and j, and d_{ijk} equals 1 when variable k can be compared for i and j, and 0 otherwise. When all comparisons can be made $\Sigma d_{ijk} = n$, the total number of variables. The scores, s_{ijk}, are assigned as follows:

For binary variables the presence of that measure is denoted by + and its absence by -. Four different combinations may occur for two activities and the score and validity assigned to each combination is shown in figure 4.6.

	Values of variable k			
activity i	+	+	-	-
activity j	+	-	+	-
s_{ijk}	1	0	0	0
d_{ijk}	1	1	1	0

Figure 4.6

Scores and validity of binary variable comparisons

For multistate variables s_{ijk} is set to 1 if the two activities i and j agree in the kth variate and $s_{ijk} = 0$ if they differ.

For continuous variables with values x_1, x_2, . . . , x_n, on variate k for the total sample of n activities we set

$$s_{ijk} = 1-|x_i - x_j| /R_k$$

where R_k is the range of the variate k and may be the total range in the population or the range in the sample. When $x_i = x_j$ then $s_{ijk} = 1$, and when x_i and x_j are at opposite ends of their range, s_{ijk} is a minimum (0 when R_k is determined from the sample). With intermediate values s_{ijk} is a positive fraction.

Thus S_G ranges between 0 and 1; a value of 1 meaning the two activities differ in no variables, whereas a value of 0 means they differ maximally over all the measures.

A further important characteristic of this similarity measure is in the representation of the data as a set of points in space. With n activities the n x n matrix, S, can be formed whose element, s_{ij}, is the similarity (as defined above) between activities i and j. A convenient representation of the n activities in Euclidean space can be obtained by taking the distance between the ith and jth activities as proportional to $(2(1-s_{ij}))^{½}$. The coordinates of points with these distances are the elements of the latent vectors of S scaled so that their sums of squares equal the latent roots. Thus to obtain a real Euclidean representation it is sufficient for S to be positive semi-definite. Gower (1971) presents a proof that S is always positive semi-definite. This important characteristic is crucial to the operation of the ordination techniques discussed in Chapter 7.

4.6 MEASURES OF ASSOCIATION BETWEEN ACTIVITIES

In a simple problem with only two variables it is possible to plot the activities in two dimensions (as in figure 4.7).

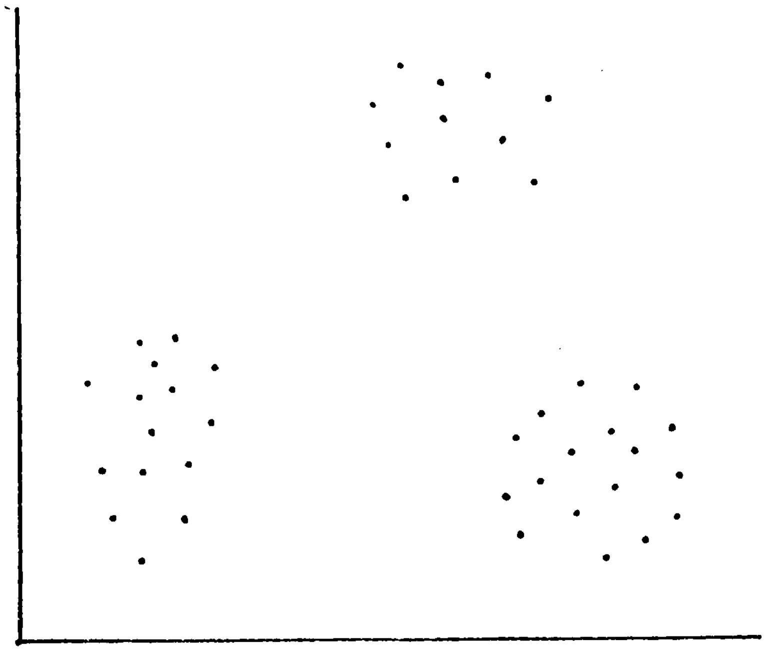


Figure 4.7

Two-dimensional clustering

The distances between points can be assessed visually and clusters identified by inspection. Visual assessment of distances is, however, impossible in spaces of more than three dimensions and must give way to computational methods.

4.6.1 Metric Measures For Continuous Variables

The most mathematically sophisticated of the distance functions are those called metrics. This class of function is of general mathematical interest and consequently has received considerable study. This discussion will present only those results most directly applicable in cluster analysis.

Let E be a symbolic representation for a measurement space and let X, Y, and Z be any three points in E. Then a distance function D is a metric if and only if it satisfies the following conditions:

$D(X,Y) = 0$ if and only if $X=Y$

$D(X,Y) \geqslant 0$ for all X and Y in E

$D(X,Y) = D(Y,X)$ for all X and Y in E

$D(X,Y) \leqslant D(X,Z)+D(Y,Z)$ for all X,Y, and Z in E.

The first property implies that X is zero distance from itself and that any two points zero distance apart must be identical. The second property prohibits negative distances. The third property imposes symmetry by requiring the distance from X to Y to be the same as the distance from Y to X. The fourth property is known as the triangle inequality and it requires that the length of one side of a triangle be no longer than the sum of the lengths of the other two sides. These properties are in accordance with intuitive notions because the popular conception of distance is the Euclidean distance of elementary geometry, itself a metric.

It may be verified quite easily that the sum of two metrics is also a metric. However, the product of two metrics (in particular the square of a metric) does not necessarily satisfy the triangle inequality and so may not be a metric. Any positive multiple of a metric is a metric. If D is a metric and w is any positive number, then

$$D' = \frac{D}{w+D}$$

is also a metric. A function which satisfies the first three conditions of a metric but not the triangle inequality is known as a semimetric. A metric which additionally satisfies

$$D(X,Y) \leqslant \max\big(D(X,Z),D(Y,Z)\big) \quad \text{for all } X,Y,Z \text{ in } E$$

is called an ultrametric (Johnson, 1967). This latter property is considerably stronger than the triangle inequality.

4.6.2 The Minkovski Metric And Special Cases

Let x_{ij} be the score achieved by the jth activity on the ith variable and let the vector of scores for the jth activity be $X_j = (x_{1j}, \ldots .x_{nj})$. Then the Minkovski metric between activities j and k is

$$D_p(X_j, X_k) = \left(\sum |x_{ij} - x_{ik}|^p\right)^{1/p}$$

where $p \geqslant 1$. By choosing various values of p many different metric distance functions can be obtained. The so-called "city block" or L_1 metric is obtained by taking p=1:

$$D_1(X_j, X_k) = \sum |x_{ij} - x_{ik}|$$

The familiar Euclidean distance or L_2 metric is obtained by taking p=2:

$$D_2(X_j, X_k) = \left(\sum (x_{ij} - x_{ik})^2\right)^{1/2}$$

The Chebychev metric is obtained as the limit of $D_p(X_j, X_k)$ as p increases without bound and so sometimes is called the L_∞ (L-infinity) metric:

$$D_\infty(X_j, X_k) = \max |x_{ij} - x_{ik}|$$

Of all possible metrics most attention is given to the Euclidean or L_2 metric. The L_1 metric occasionally is encountered and metrics based on other values of p hardly ever are of more than theoretical interest.

4.7 SUMMARY

The main types of classification have been formally identified. Most cluster analysis methods require a measure of similarity to be defined for every pairwise combination of the activities to be clustered. The types of data encountered in architectural data analysis and the appropriate measures of similarity have been defined. The problems arising from mixed data types have been discussed and methods of coping with the problem proposed. The software implementation in MAGIC allows for mixed data sets by utilising Gower's General Coefficient of Similarity; it is also possible to operate with simple adjacency matrices which are interpreted as distance matrices. The various ways of measuring distance in n-dimensional space were defined. All the techniques incorporated in MAGIC use the Minkovski L_2 metric (Euclidean distance).

CHAPTER 5

HIERARCHICAL CLUSTER ANALYSIS

5.1 INTRODUCTION

The measures of association discussed in chapter 4 may be used to construct a similarity matrix describing all pairwise relationships among the entities in the data set. The methods of cluster analysis operate on this similarity matrix to produce the clusters of activities. In the implementation of these methods in MAGIC the similarity measure used in all cases is Euclidean distance, obtained either directly from a simple distance matrix or by a transformation of Gower's S.

There are two main approaches to cluster analysis: a hierarchical classification or a partitioning method. This chapter discusses hierarchical methods. These may be broadly categorised as seeking the optimal partition into g groups for all values of g between 2 and n (the number of individual activities). If for every g_1, g_2

satisfying $2 \leqslant g_1 < g_2 \leqslant n$, each group in the g_2-partition is wholly contained within a single group in the g_1-partition, the set of partitions is said to be hierarchically nested. A hierarchically nested set of partitions can be represented by a tree diagram, or dendrogram, such as the one shown in figure 5.8. Each of the n branch ends represents a single activity. Each position up the tree at which branches join has an associated numerical value, d, d_{ij} being the level at which the ith and jth branches join, and is the lowest level at which the ith and jth activities belong to the same group. The smaller the value of d, the more similar the ith and jth activities are regarded as being, and the higher up the tree they are seen to join. Sectioning a dendrogram at any level yields a partition of the data set.

The general strategy underlying agglomerative polythetic clustering on a data matrix may be represented as follows.

(1) Consider each of the n activities as a cluster consisting of just one entity. Let these clusters be numbered 1 to n.

(2) Search the similarity matrix for the most similar pair of clusters. Let these be labelled p and q.

(3) Merge p and q and calculate their associated similarity s_{pq}. Label the product of the merger q.

(4) Reduce the number of clusters by one (because of the merger in (3)), and update the similarity matrix to show the revised similarities between cluster q and all other existing clusters. Delete the row and column of S pertaining to cluster p_{l}.

(5) Repeat steps 2,3, and 4 (n-1) times.

The different methods vary in the procedure for defining the most similar pair at step (2), and the measurements used in updating the similarity matrix at step (4).

This general strategy is easily conceptualised as a geometric model. Consider each of the n activities as a point in space; combine the closest pair, p and q, into a single group. The distances of all other points to this group replace their distances to p and q individually. Repeat the process until all points have been incorporated into a single group. As the majority of strategies have readily conceived geometrical interpretations there are advantages if the measures used are metrics, although this is not essential.

There are eight main clustering strategies available, some of which have been known for several years and acquired a series of alternative names. The strategies implemented in MAGIC are Nearest Neighbour, Furthest Neighbour, Group Average, Centroid, Median, Incremental Sum of Squares, Simple Average, and Flexible Strategy. These are each discussed in turn below.

Each strategy exhibits particular properties which affect the relationships between the clusters formed. Lance and Williams (1967) describe the main characteristic as "space distortion". Considering the geometric model of points in multidimensional space certain strategies leave the properties of this space unaltered, but in others the clusters alter the space near to them. Certain strategies operate in effect by erecting boundaries between groups of points, but do not change the relative positions of the points in the original space. Such strategies are said to be space conserving. In other strategies the space around a group appears to stretch as the group grows, so that the group appears to recede from the other points as it grows. Such strategies are said to be space dilating; they cluster intensely, and the groups appear to be more distinct than is really the case. In other strategies the space appears to contract around a group as it grows (space-contracting strategies); inherent clustering is reduced and there may be much "chaining" - the successive addition of single points.

In pragmatic terms the space-contracting strategies (exemplified by the Nearest Neighbour method) are weakly clustering, give chains of activities and are not always of any great conceptual value in exploratory data analysis. The space-dilating strategies, for example Incremental Sum of Squares, are strongly clustering and of considerable conceptual value. Intermediate to these are the space-conserving strategies such as the Group Average method. The Flexible Strategy is unique in that it can be altered from space-contracting to space-dilating; in its usual operation it has become space-dilating.

Another property of certain strategies is nonmonotonicity. For the complete clustering process to be visually represented the activities can be arranged in a convenient order along an abscissa and the successive clusterings shown as a dendrogram. To enable the dendrogram to be drawn the string of dissimilarities associated with the successive clusterings should rise monotonically. In certain cases the Median and Centroid methods become nonmonotonic (when, in the combinatorial equation (see below) $\alpha_i+\alpha_j+\beta<1$).

Finally, if, given the initial interactivity dissimilarity matrix, all subsequent individual/group and group/group measures can be calculated from this alone by a recursive process, the system is said to be combinatorial. Lance and Williams (1967) have shown that all (i,j) measures in common use can be encompassed within a single linear combinatorial model. Given two groups (i) and (j) with n_i and n_j elements respectively and intergroup dissimilarity d_{ij}, if we assume that d_{ij} is the smallest measure remaining in the system, then (i) and (j) fuse to form a new group (k) with $n_k=(n_i+n_j)$ elements. Consider a third group (h) with n_h elements. Before the fusion the values of d_{hi}, d_{hj}, d_{ij}, n_h, n_i, and n_j are all known (see figure 5.1).

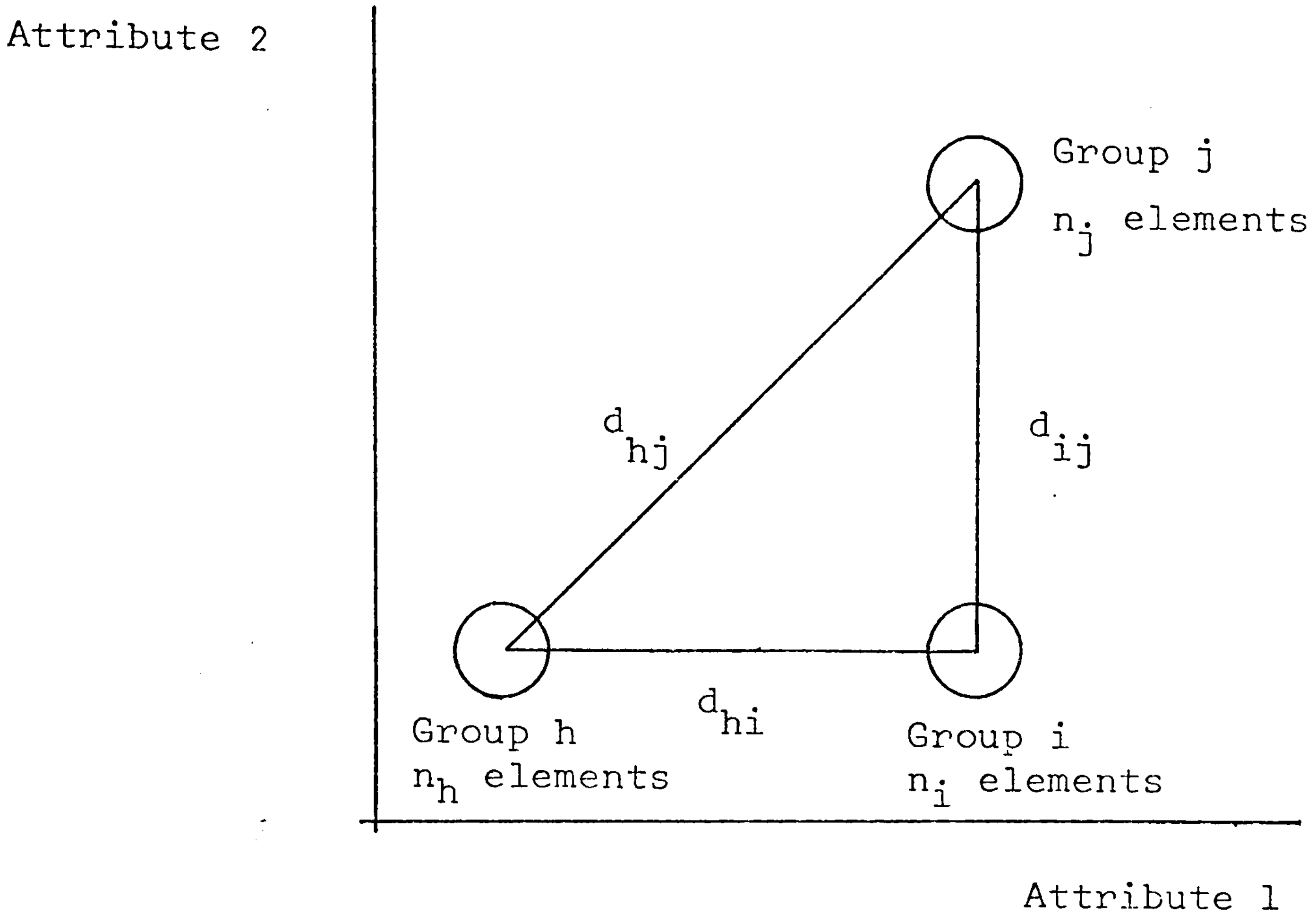


Figure 5.1

Elements of Lance and Williams combinatorial equation

We may then set

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

where the parameters α_i, α_j, β and γ determine the nature of the strategy. In a few cases these parameters may be actual numbers; in most, however, they are simple algebraic functions of some or all of n_i, n_j, n_k, and n_h. The actual values or expressions are given below in connection with the detailed discussion of the individual strategies and summarised in figure 5.2. Nomenclature is a problem as the various strategies are given different names by different authors. Synonyms are therefore given and the name adopted here is generally the most widely accepted.

Examples of each of the clustering strategies are given using standard data from Sneath and Sokal (1973), summarised in figure 5.3.

Strategy	α_i	α_j	β	γ	Mono-tonic	Spatial Effect
N.N.	0.5	0.5	0	-0.5	yes	contract
F.N.	0.5	0.5	0	0.5	yes	dilate
G.A.	n_i/n_k	n_j/n_k	0	0	yes	conserve
C.	n_i/n_k	n_j/n_k	$-\alpha_i\alpha_j$	0	no	conserve
M.	0.5	0.5	-0.25	0	no	conserve
I.S.S.	$\frac{(n_h+n_i)}{(n_h+n_k)}$	$\frac{(n_h+n_j)}{(n_h+n_k)}$	$\frac{-n_h}{(n_h+n_k)}$	0	yes	dilate
S.A.	0.5	0.5	0	0	yes	dilate
F.(i)	0.625	0.625	-0.25	0	yes	dilate

(i) Flexible Strategy
convention to set β =-0.25

Combinatorial equation $d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma(d_{hi} - d_{hj})$

n_i = no. of elements in group i
n_j = no. of elements in group j
$n_k = n_i + n_j$

Figure 5.2

Values of various strategies in combinatorial equation

The coordinates of 16 activities in a two-space defined by axes X_1 and X_2 are given in the first two rows of the table. Euclidean distances between the pairs of activities are shown in the lower triangular matrix, their squares in the upper triangular matrix.

	a	b	c	d	e	f	g	h
X_1	0	0	1	2	3	2	2	1
X_2	4	3	5	4	3	2	1	0
a	x	1	2	4	10	8	13	17
b	1.000	x	5	5	9	5	8	10
c	1.414	2.236	x	2	8	10	17	25
d	2.000	2.236	1.414	x	2	4	9	17
e	3.162	3.000	2.828	1.414	x	2	5	13
f	2.828	2.236	3.162	2.000	1.414	x	1	5
g	3.606	2.828	4.123	3.000	2.236	1.000	x	2
h	4.123	3.162	5.000	4.123	3.606	2.236	1.414	x
i	5.099	5.385	4.000	3.162	2.828	4.243	5.000	6.403
j	6.083	6.325	5.000	4.123	3.606	5.000	5.657	7.071
k	7.280	7.616	6.083	5.385	5.000	6.403	7.071	8.485
l	5.099	5.000	4.472	3.162	2.000	3.162	3.606	5.000
m	7.071	7.000	6.325	5.099	4.000	5.099	5.385	6.708
n	6.325	6.083	5.831	4.472	3.162	4.000	4.123	5.385
o	6.708	6.325	6.403	5.000	3.606	4.123	4.000	5.099
p	8.544	8.246	8.062	6.708	5.385	6.083	6.000	7.071

Figure 5.3a (i)

Test data from Sneath and Sokal (1973)

The coordinates of 16 activities in a two-space defined by axes X_1 and X_2 are given in the first two rows of the table. Euclidean distances between the pairs of activities are shown in the lower triangular matrix, their squares in the upper triangular matrix.

	i	j	k	l	m	n	o	p
X_1	5	7	7	5	7	6	6	8
X_2	5	5	6	3	3	2	1	1
a	26	37	53	26	50	40	45	73
b	29	40	58	25	49	37	40	68
c	16	25	37	20	40	34	41	65
d	10	17	29	10	26	20	25	45
e	8	13	25	4	16	10	13	29
f	18	25	41	10	26	16	17	37
g	25	32	50	13	29	17	16	36
h	41	50	72	25	45	29	26	50
i	x	1	5	4	8	10	17	25
j	1.000	x	2	5	5	9	16	20
k	2.236	1.414	x	13	9	17	26	26
l	2.000	2.236	3.606	x	4	2	5	13
m	2.828	2.236	3.000	2.000	x	2	5	5
n	3.162	3.000	4.123	1.414	1.414	x	1	5
o	4.123	4.000	5.099	2.236	2.236	1.000	x	4
p	5.000	4.472	5.099	3.606	2.236	2.236	2.000	x

Figure 5.3a (ii)

Test data from Sneath and Sokal (1973)

```
16,2                    No. of activities, variables
1                       Type of data
2,0,0                   No. of each variable type
SNEATH & SOKAL DATA     Heading
0                       No areas
A                       Names (refer to 5.3a)
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
1                       Full matrix data
0,4                     Data (from 5.3a)
0,3
1,5
2,4
3,3
2,2
2,1
1,0
5,5
6,5
7,6
5,3
7,3
6,2
6,1
8,1
```

Figure 5.3b

MAGIC data file for Sneath and Sokal data

5.2 NEAREST NEIGHBOUR CLUSTERING STRATEGY

This technique, introduced by Florek et al (1951) and Sneath (1957), is also known as the Single Linkage Method (Sneath and Sokal 1973) and the Minimum Method (Johnson 1967), and is the oldest of the conventional strategies. The distance between two groups is defined as the distance between those two individuals (one in each group) which are the nearest. The parameters are

$$\alpha_i = \alpha_j = 0.5; \quad \beta = 0; \quad \gamma = -0.5, \text{ giving}$$

$$d_{hk} = 0.5d_{hi} + 0.5d_{hj} - 0.5|d_{hi} - d_{hj}|$$

$$= 0.5(d_{hi} + d_{hj} - |d_{hi} - d_{hj}|)$$

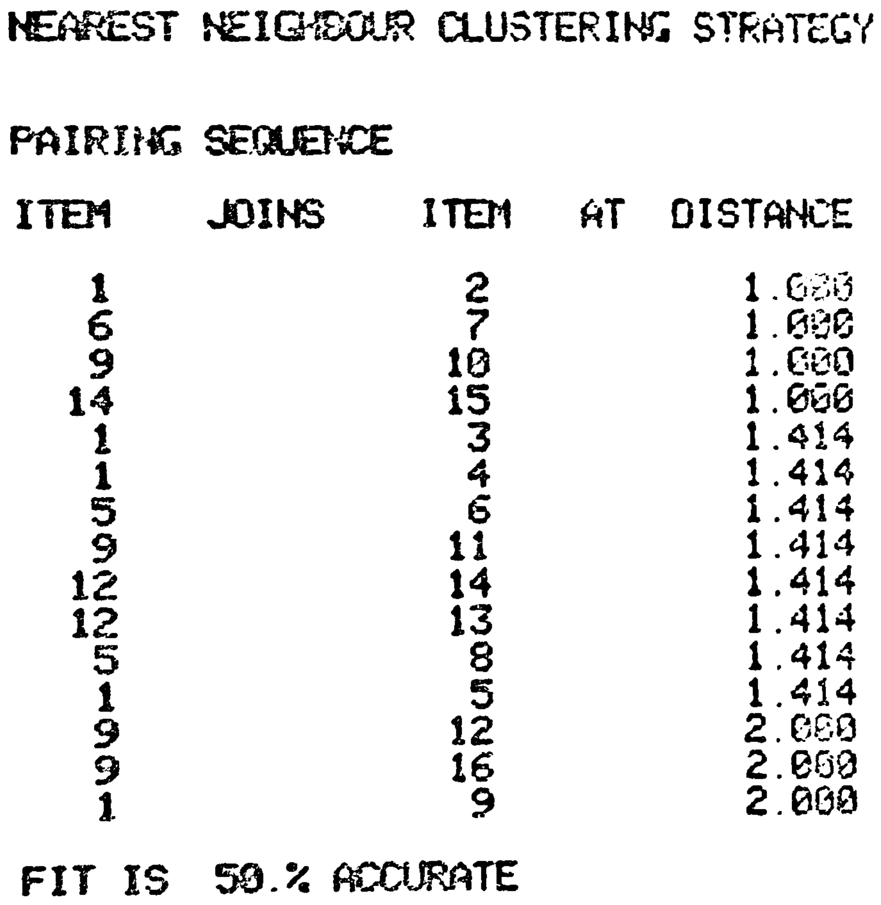
It is a monotonic, intensely space-contracting strategy, with a number of theoretical mathematical and computational advantages (Rohlf 1973, Jardine and Sibson 1968, 1971, Sibson 1973).

The technique does not delineate poorly seperated clusters, tending to produce long serpentine clusters. This property, termed "chaining", is often criticised because elements at opposite ends of a chained group may be markedly dissimilar. In a comparison of strategies Pritchard and Anderson (1971) described this as the least useful technique because of the tendency to chain. In most utilitarian aspects, therefore, this strategy may be regarded as obsolete, but it has received support on mathematical grounds from Jardine and Sibson (1968). As the cluster updating process involves choosing only the minimum (or, in the case of Furthest Neighbour, the maximum), single-link clustering is invariant to any transformation which leaves the ordering of similarities unchanged, that is, any monotonic transformation. Jardine and Sibson develop further criteria they believe should apply to classificatory strategies which virtually confine one to the use of Nearest Neighbour (discussed briefly in section 5.9.3). A controversy arose between a "Cambridge School" and an "Australian School" over this and related issues (Williams et al 1971, Sibson 1971, Jardine and Sibson 1971), but, in the end the criterion of the validity of application of particular methods must come down to "how well does it work?".

5.2.1 Example Of Nearest Neighbour Clustering

We first find the mutually most similar pairs, which turn out to be (1,2), (6,7), (9,10) and (14,15), all at a distance of 1.0 from each other (see the pairing sequence, figure 5.4, and figure 5.3). The geometric result of this is shown in figure 5.5 where these pairs are connected with solid lines. New candidates for fusion (either between themselves or to established groups) are now examined. At this next level of grouping (1.414) several activities join cluster 1. The pairs are (1,3), (1,4), (1,5), (1,6), (1,8), (9,11), (12,14), and (12,13). Finally (9,12), (9,16) and (1,9) are joined. Note how in the geometric representation the clusters are strung out in what is the characteristic single linkage fusion fashion. In figures 5.6 and 5.7 the links existing prior to that stage are shown in light line and the new links in bold line.

Referring to the dendrogram (figure 5.8) we can see that Nearest Neighbour clustering has revealed three levels of clustering. The most closely related activities are (1,2), (6,7), (9,10), and (14,15), with 3,4,5,8,11,12 and 13 remaining unattached at that level. The next highest level is represented by (1,2,3,4,5,6,7,8), (9,10,11), (12,13,14,15), and 16; whilst at the final level all the activities come together as a single entity. The representation of clusters by dendrograms is discussed further in Chapter 10.



```
NEAREST NEIGHBOUR CLUSTERING STRATEGY

PAIRING SEQUENCE
ITEM    JOINS    ITEM    AT  DISTANCE

   1               2           1.000
   6               7           1.000
   9              10           1.000
  14              15           1.000
   1               3           1.414
   1               4           1.414
   5               6           1.414
   9              11           1.414
  12              14           1.414
  12              13           1.414
   5               8           1.414
   1               5           1.414
   9              12           2.000
   9              16           2.000
   1               9           2.000

FIT IS  50.% ACCURATE
```

Figure 5.4

Nearest Neighbour strategy - pairing sequence

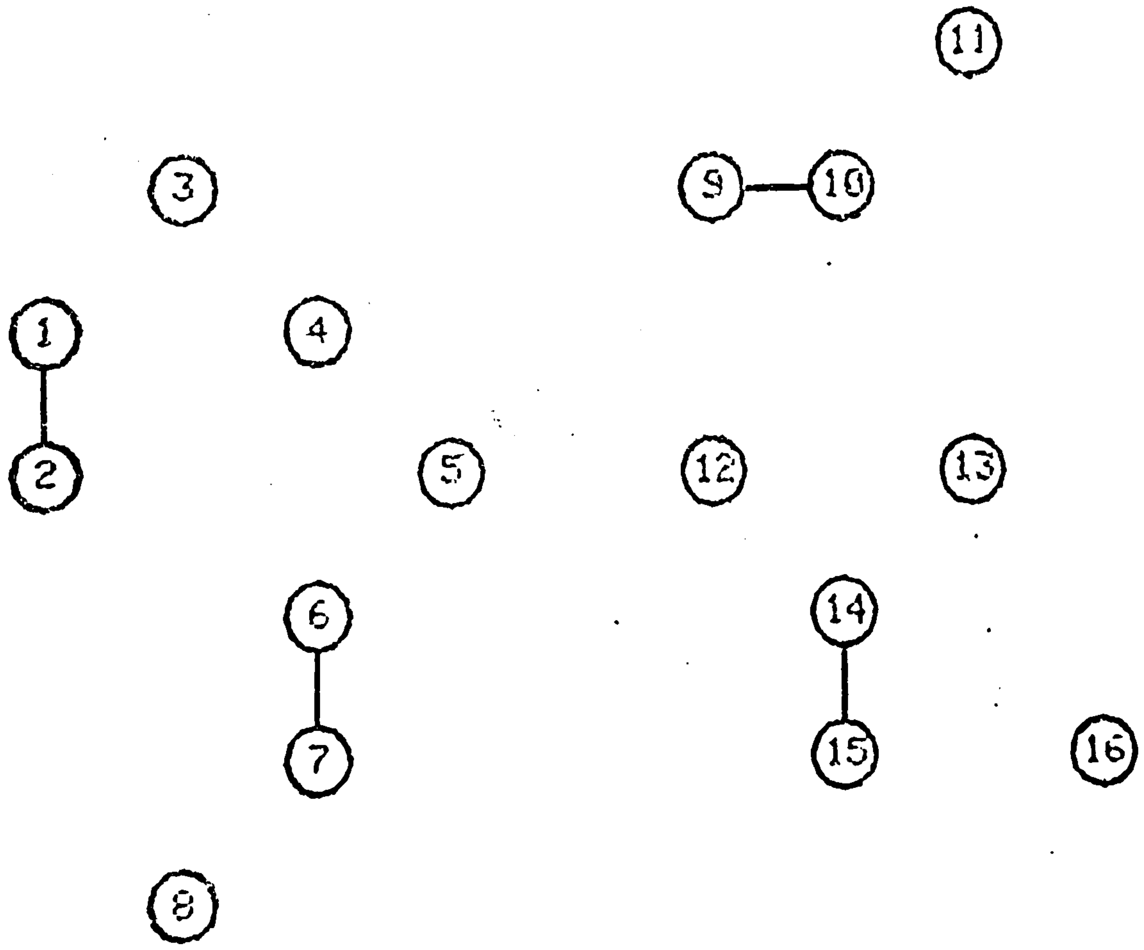


Figure 5.5

Nearest Neighbour strategy - first step

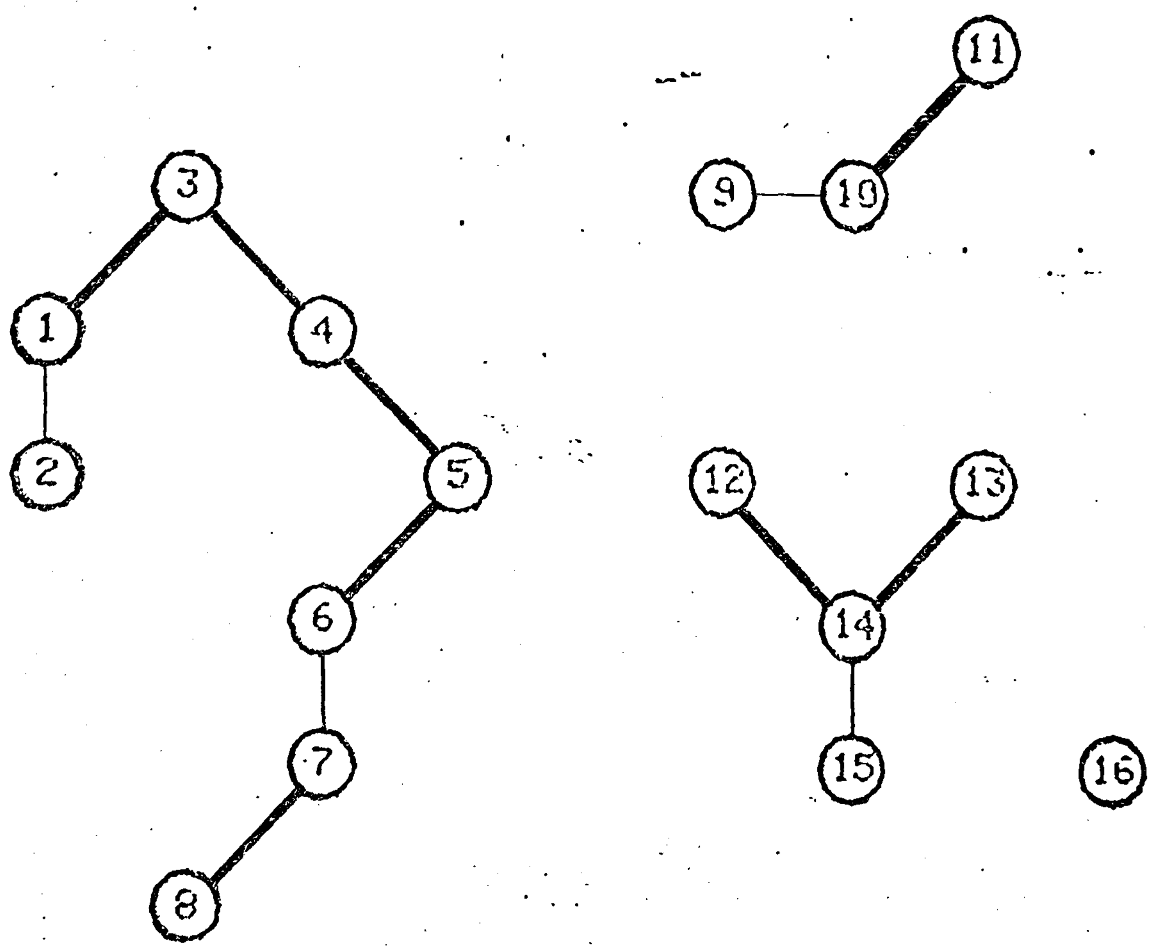


Figure 5.6

Nearest Neighbour strategy - second step

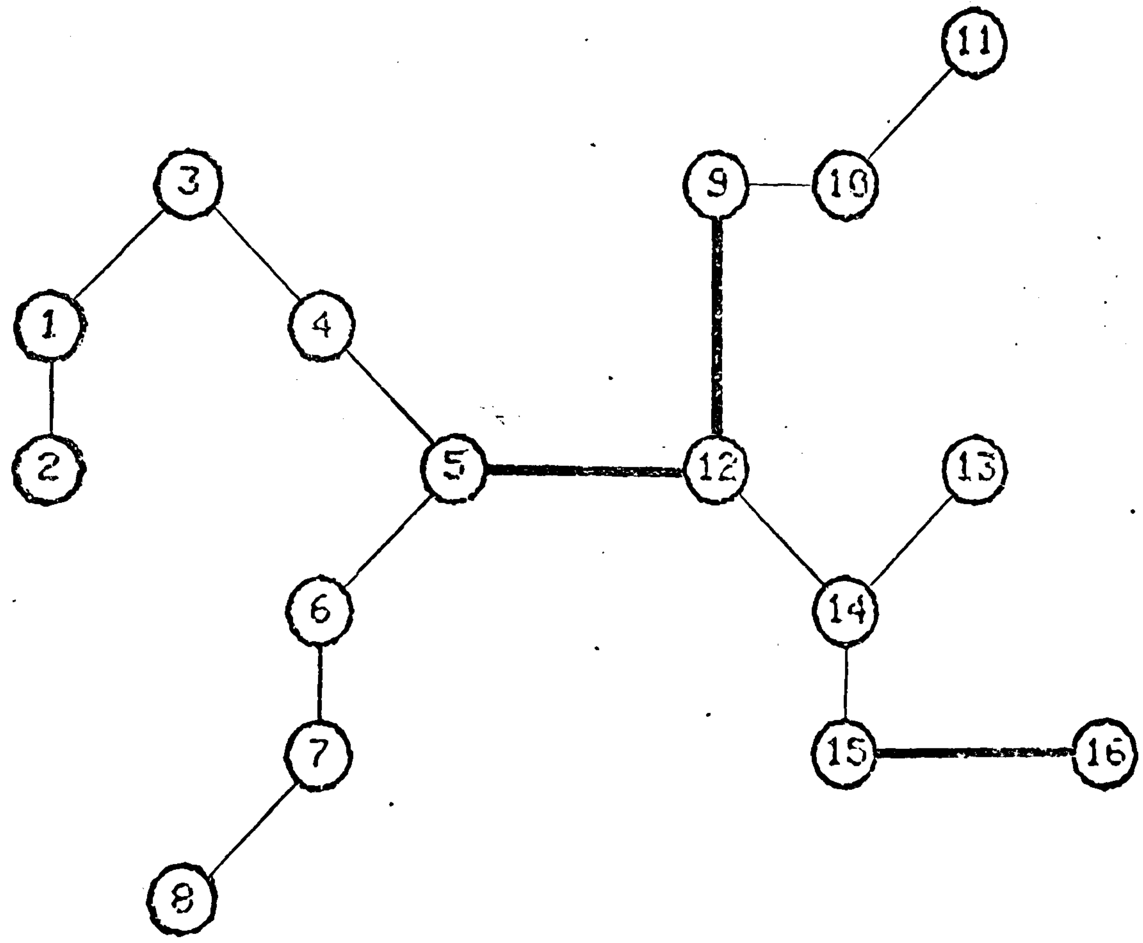


Figure 5.7

Nearest Neighbour strategy - third step

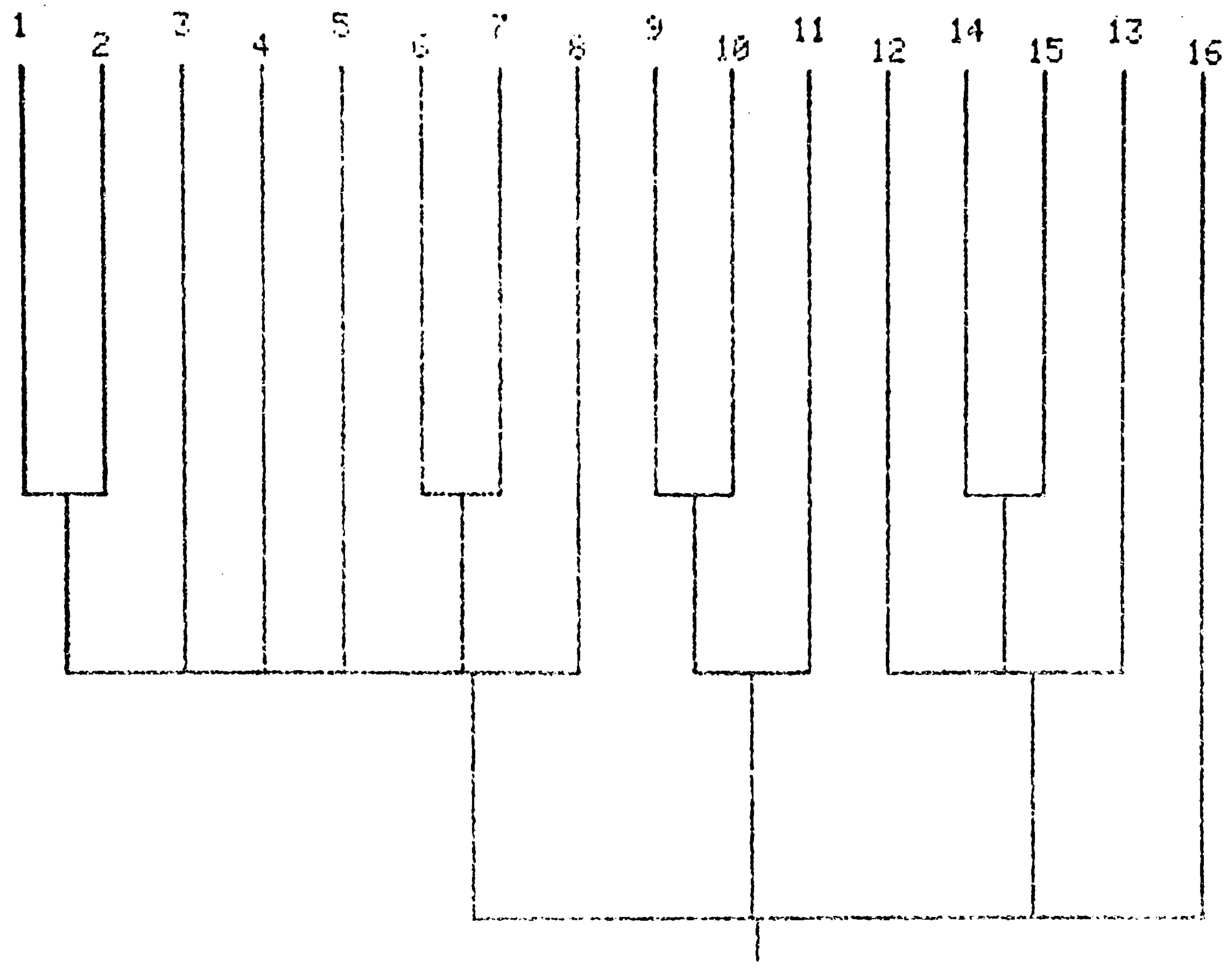


Figure 5.8

Nearest Neighbour strategy - dendrogram

5.3 FURTHEST NEIGHBOUR CLUSTERING STRATEGY

This technique was originated by Sorensen (1948), and is also known as the Complete Linkage Method (Sokal and Sneath 1973), and the Maximum Method (Johnson 1967). Its current name was established by Lance and Williams (1967).

The distance between two groups is defined as the distance between their two most remote individuals, and linkages made on the basis of the closest of these distances. It is a monotonic, intensely space-dilating strategy which has been largely superceded by the Flexible Strategy. The parameters are

$$\alpha_i = \alpha_j = 0.5, \ \beta = 0, \ \gamma = 0.5$$

giving

$$d_{hk} = 0.5d_{hi} + 0.5d_{hj} + 0.5|d_{hi} - d_{hj}|$$

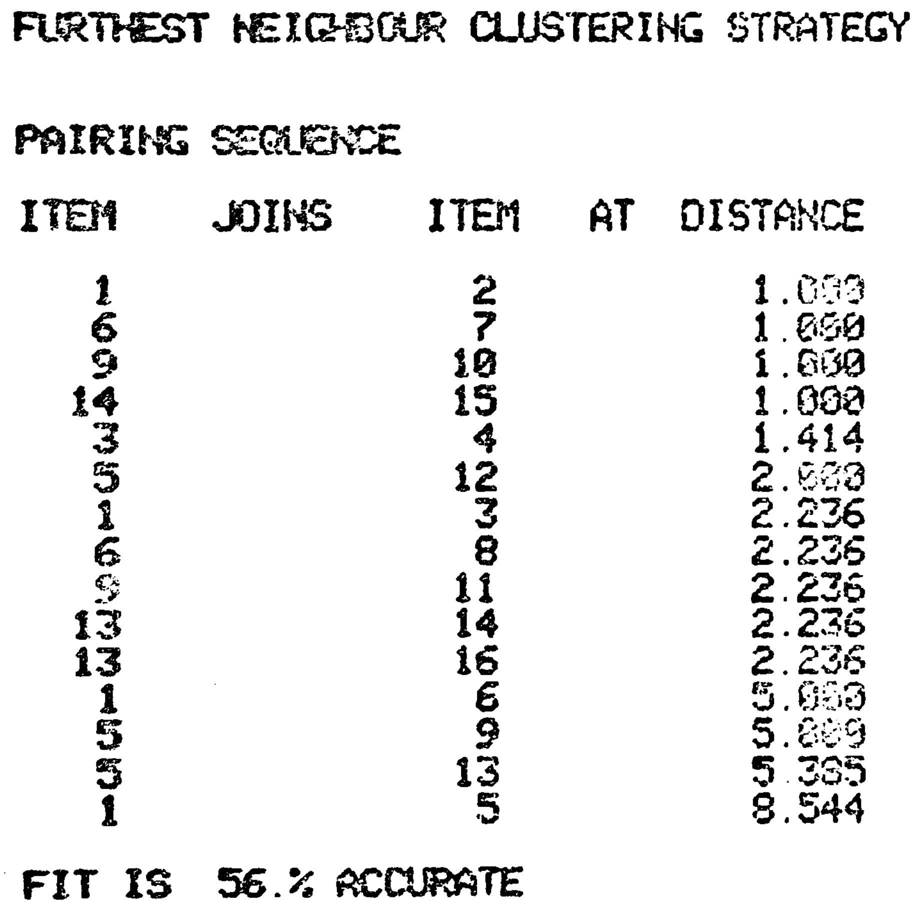
$$= 0.5 (d_{hi} + d_{hj} + |d_{hi} - d_{hj}|)$$

5.3.1 Example Of Furthest Neighbour Clustering

The method commences in the same manner as the Nearest Neighbour technique (figure 5.10). For an activity to now join an existing cluster the distance criterion is, in this strategy, now taken, not to the nearest element of that cluster, but to the furthest. When two clusters join the similarity is that existing between the farthest member pair, one from each cluster. This method thus leads to a number of tight, discrete clusters that join each other only with difficulty and at relatively low similarity levels.

Inspection of the clusters generated (see particularly figure 5.13) shows their induced compactness in comparison with the loose, strung-out clusters of the Nearest Neighbour strategy (figure 5.6). The data is more structured, showing more clusters and more levels than the Nearest Neighbour strategy.

While the most highly connected activities are (1,2), (6,7), (9,10) and (14,15) as before, the next levels produce (3,4), then (5,12) so that even at the fourth level there are five distinct groups - (1,2,3,4), (5,12), (6,7,8), (9,10,11) and (13,14,15,16). These five reduce to four in the next step and then three and two. The final fusion is then at a very much lower level. The dendrogram showing this structure is drawn in figure 5.17.



```
FURTHEST NEIGHBOUR CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM    JOINS    ITEM   AT  DISTANCE

  1                2          1.000
  6                7          1.000
  9               10          1.000
 14               15          1.000
  3                4          1.414
  5               12          2.000
  1                3          2.236
  6                8          2.236
  9               11          2.236
 13               14          2.236
 13               16          2.236
  1                6          5.000
  5                9          5.099
  5               13          5.385
  1                5          8.544

FIT IS  56.% ACCURATE
```

Figure 5.9

Furthest Neighbour strategy - pairing sequence

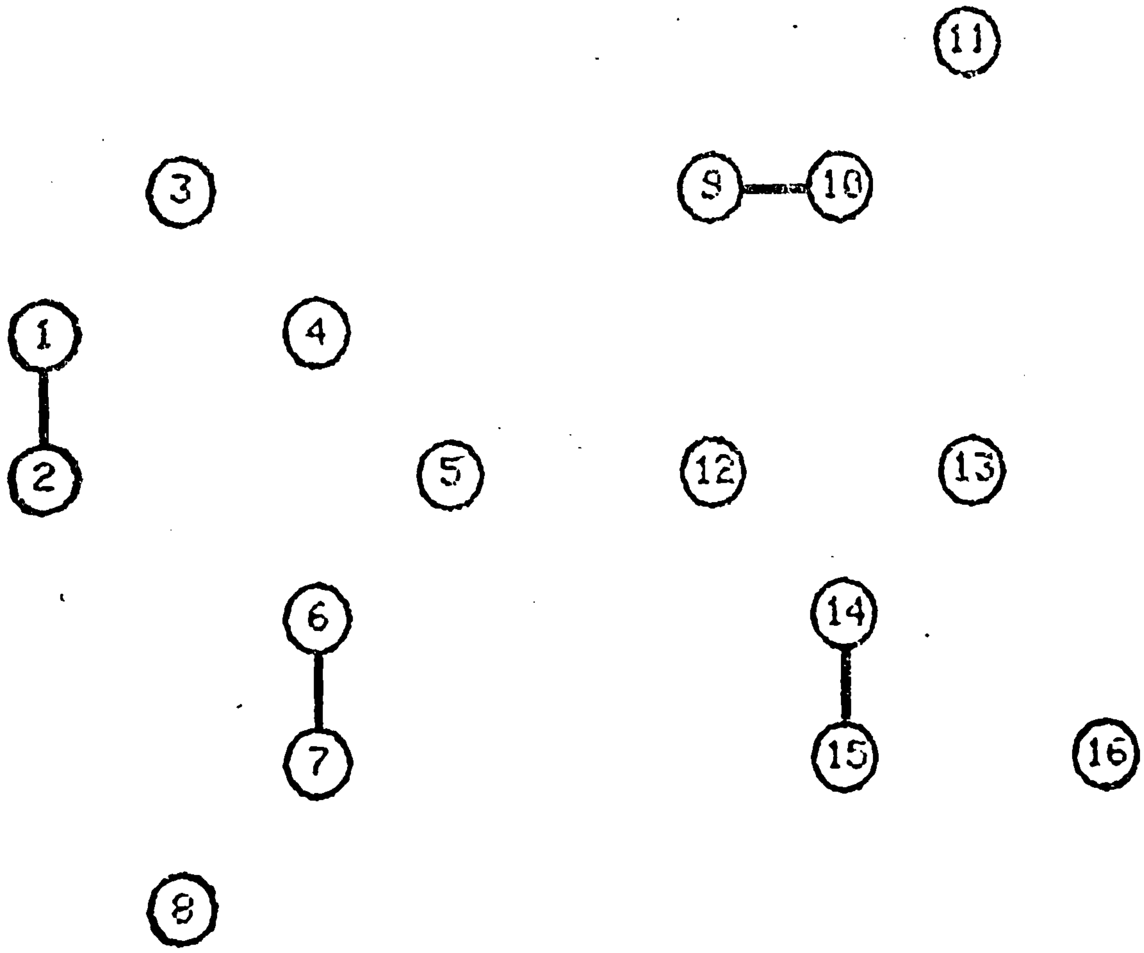


Figure 5.10

Furthest Neighbour strategy - first step

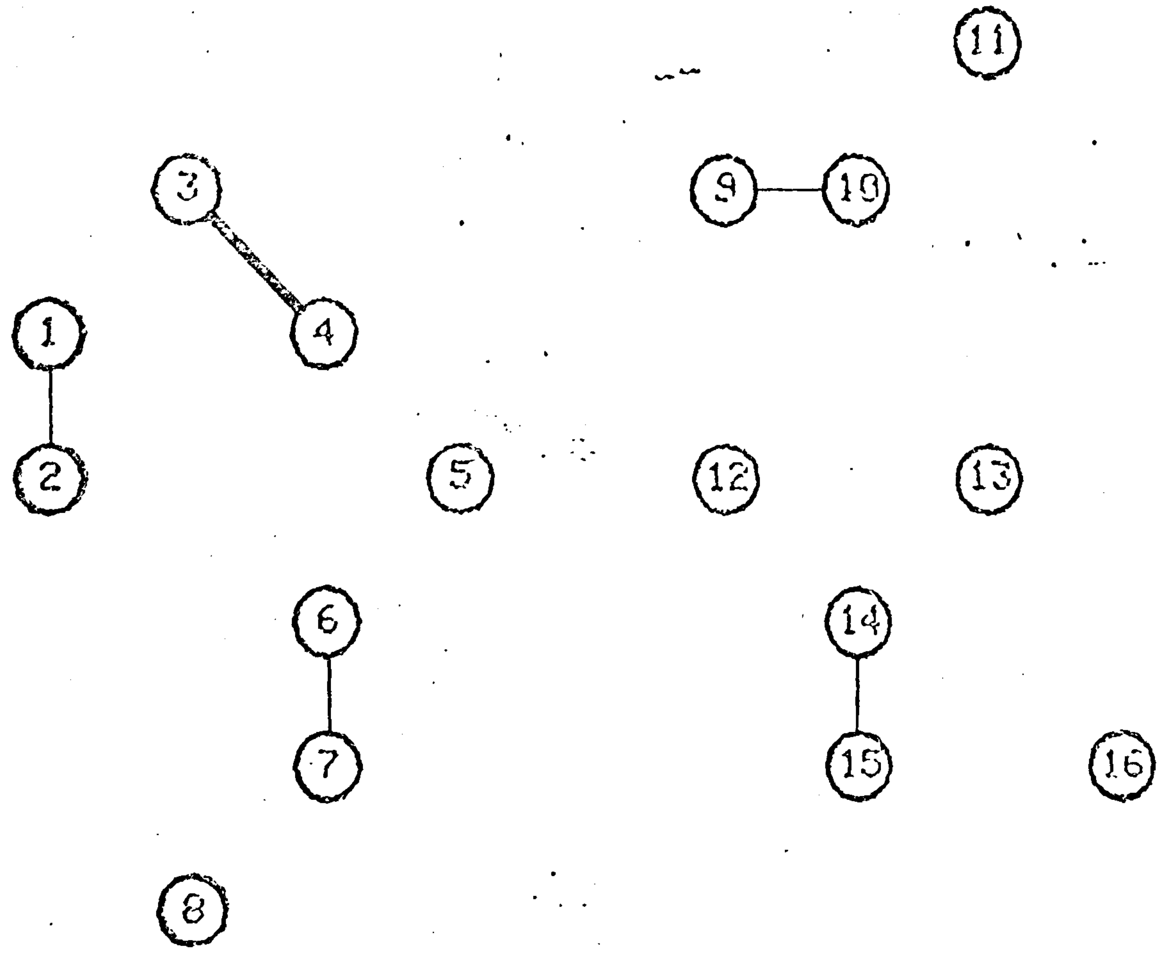


Figure 5.11

Furthest Neighbour strategy - second step

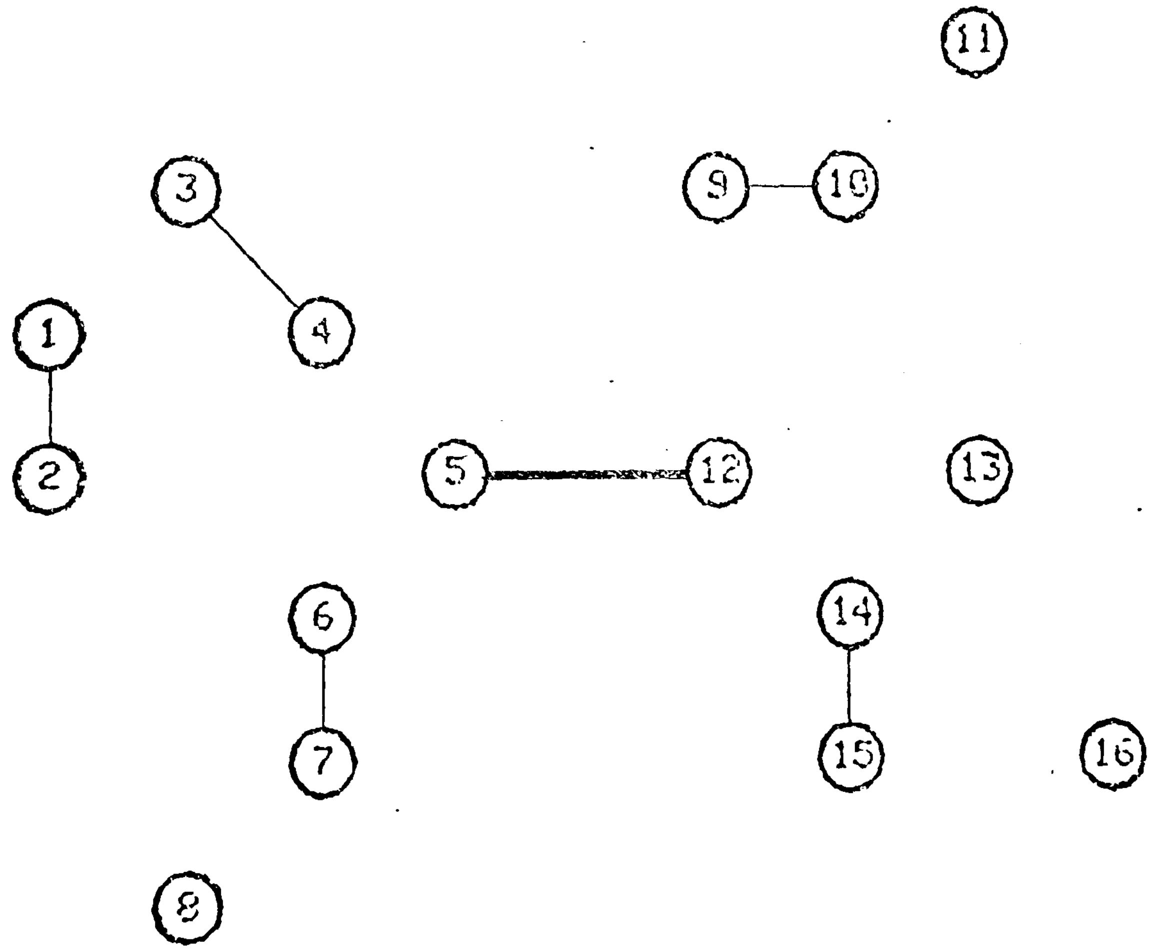


Figure 5.12

Furthest Neighbour strategy - third step

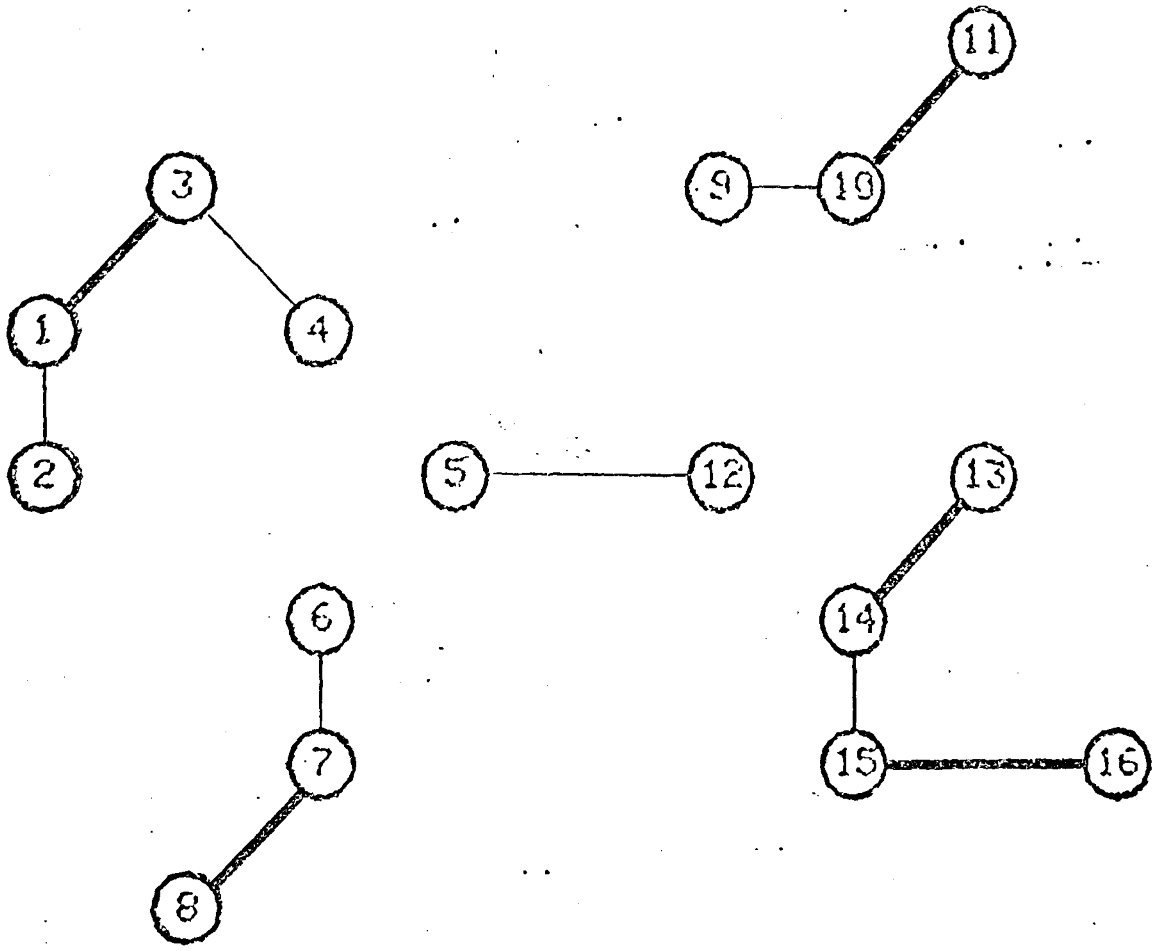


Figure 5.13

Furthest Neighbour strategy - fourth step

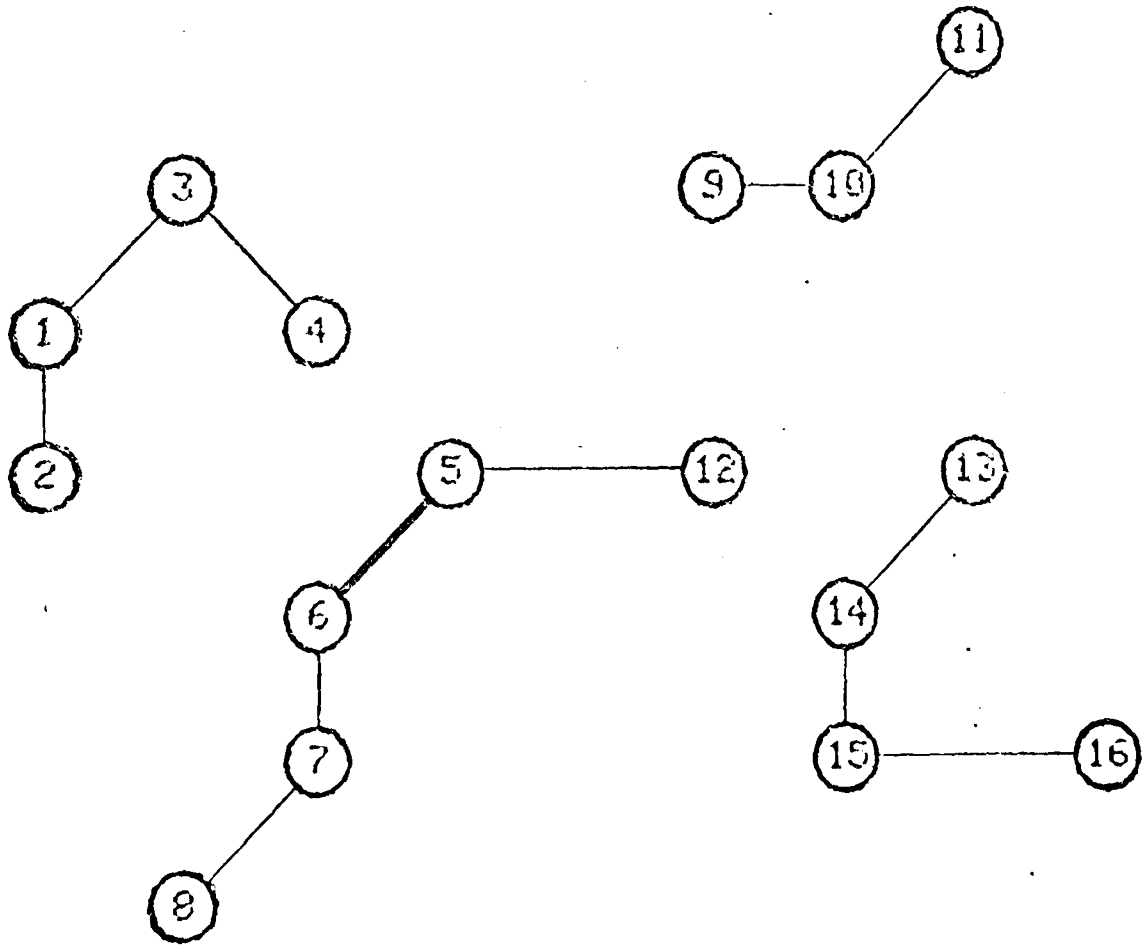


Figure 5.14

Furthest Neighbour strategy - fifth step

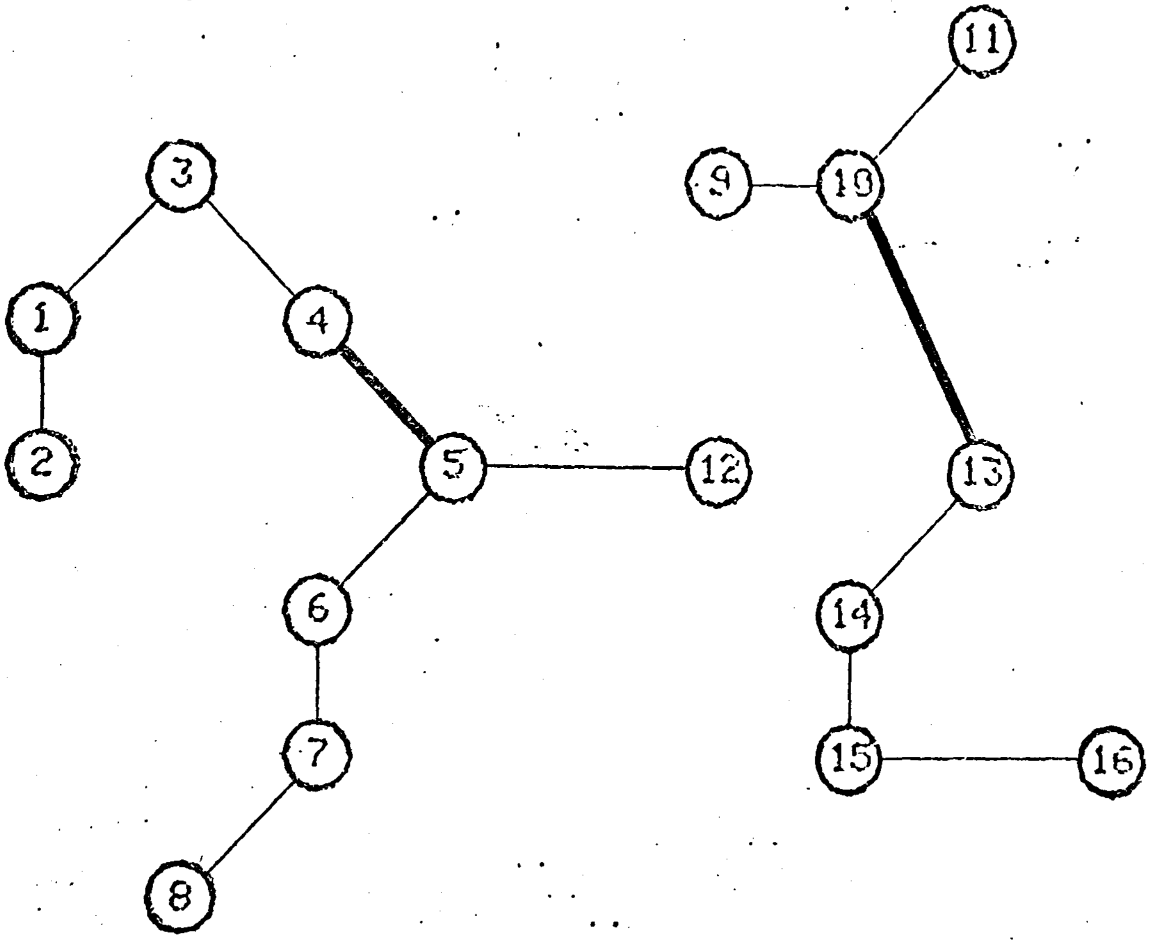


Figure 5.15

Furthest Neighbour strategy - sixth step

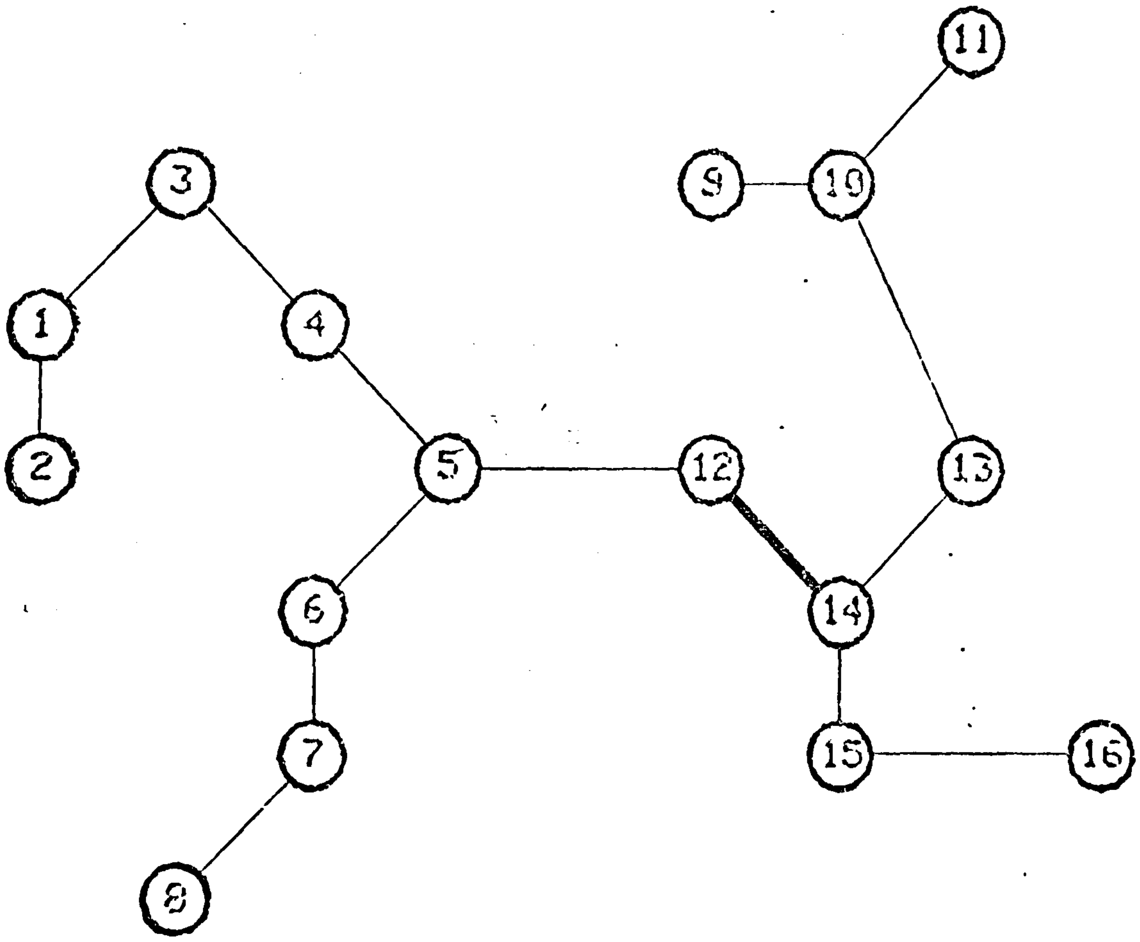


Figure 5.16

Furthest Neighbour strategy - seventh step

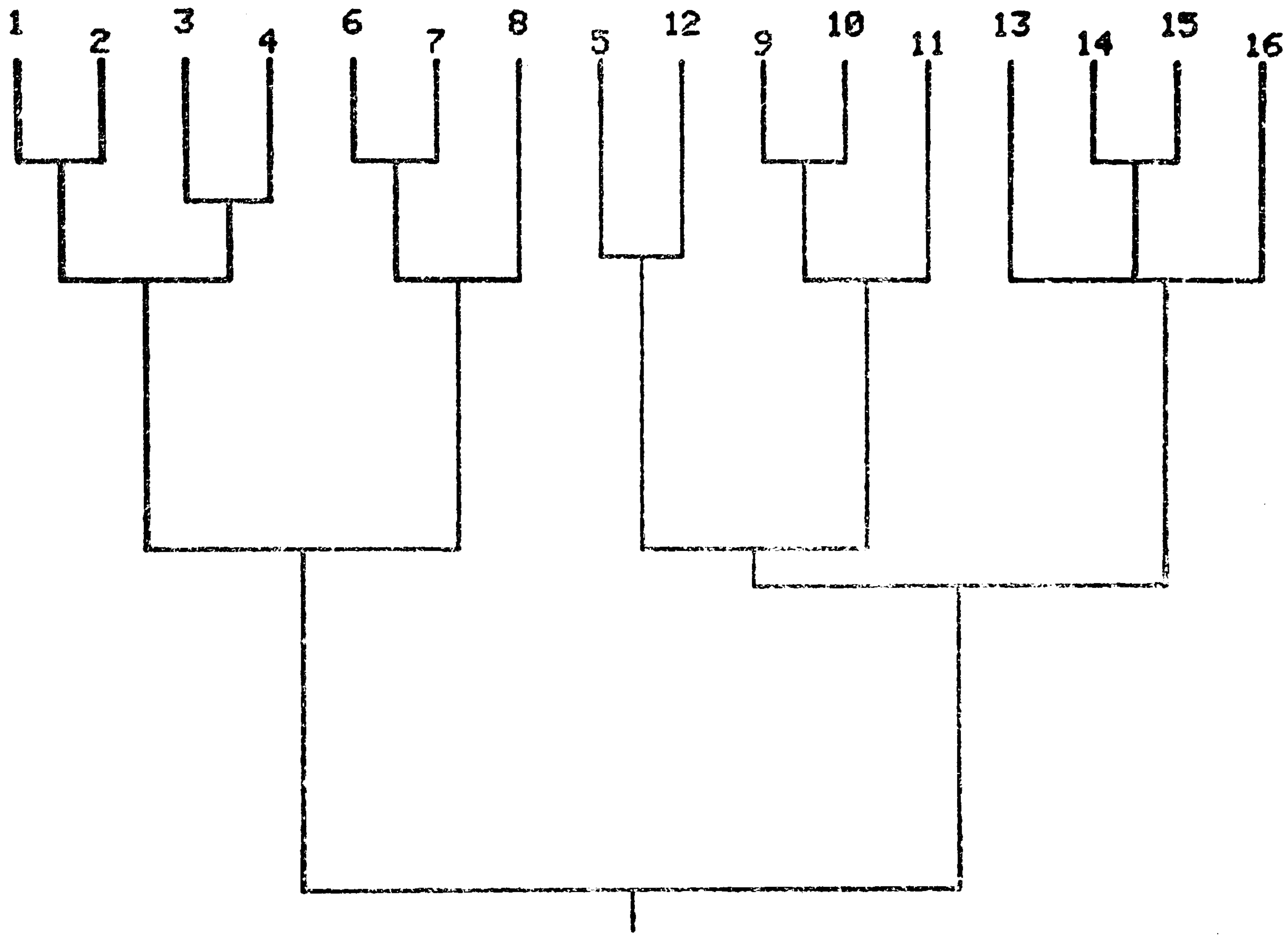


Figure 5.17

Furthest Neighbour strategy - dendrogram

5.4 GROUP AVERAGE CLUSTERING STRATEGY

This technique is also described as the Unweighted Pairgroup Method Using Arithmetic Averages (Sokal and Michener 1958). The Group Average name was established by Lance and Williams (1967).

If there are m_1 individuals in one group and m_2 in another, the distance between them is defined as the arithmetic mean of all $m_1 m_2$ interindividual distances. Fusion is between the two groups with the shortest mean distance. The parameters are

$$\alpha_i = n_i/n_k; \ \alpha_j = n_j/n_k; \ \beta = \gamma = 0$$

$$d_{hk} = (n_i d_{hi} + n_j d_{hj})/n_k$$

where n_i = number in group i
n_j = number in group j
$n_k = n_i + n_j$

It is monotonic and substantially space conserving.

This method is less rigorously space conserving than the Centroid method, but, having no marked tendencies to contraction or dilation may be regarded as a conserving strategy. Group Average clustering is a generally satisfactory technique giving moderately distinct clusters, with the advantages of being monotonic, little prone to misclassification, and with little group size dependence. It may therefore be usefully employed both as a general "work-horse" technique and also to check for misclassifications resulting from the application of more intensely clustering strategies.

5.4.1 Example Of Group Average Clustering

The strategy computes the average similarity of a point relative to an extant cluster, weighting each element in that cluster equally. Fusion is then made with the cluster giving the shortest mean distance. To show this point the first steps of the clustering of the example data are worked manually.

The initial clustering step is the same as in the previous cases: (1,2), (6,7), (9,10), (14,15). The new distance between (1,2) and 3 can be computed by simply averaging d_{13} and d_{23}, i.e.

$$d_{(12)3} = 0.5(1.414+2.236)$$
$$= 1.825$$

Distances involving two new clusters, such as $d_{(12)(67)}$ are computed as

$$0.25(d_{16}+d_{17}+d_{26}+d_{27})$$

i.e. $d_{(12)(67)} = 0.25(2.828+3.605+2.236+2.828)$

$$= 2.875$$

Distances between elements that did not join any cluster are transcribed unchanged from the original matrix; for example d_{34}=1.414.

The complete distance matrix after the first clustering is shown in figure 5.18.

	(1,2) AB	3 C	4 D	5 E	(6,7) FG	8 H
(1,2) AB	x					
3 C	1.825	x				
4 D	2.118	1.414	x			
5 E	3.081	2.828	1.414	x		
(6,7) FG	2.875	3.643	2.500	1.825	x	
8 H	3.643	5.000	4.123	3.605	1.825	x
(9,10) IJ	5.723	4.500	3.643	3.217	4.975	6.737
11 K	7.448	6.083	5.385	5.000	6.737	8.485
12 L	5.050	4.472	3.162	2.000	3.384	5.000
13 M	7.036	6.325	5.099	4.000	5.242	6.708
(14,15) NO	6.360	6.117	4.736	3.384	4.062	5.242
16 P	8.395	8.062	6.708	5.385	6.041	7.071

	(9,10) IJ	11 K	12 L	13 M	(14,15) NO	16 P
(9,10) IJ	x					
11 K	1.825	x				
12 L	2.118	3.606	x			
13 M	2.532	3.000	2.000	x		
(14,15) NO	3.571	4.611	1.825	1.825	x	
16 P	4.736	5.099	3.606	2.236	2.118	x

Figure 5.18

Distance Matrix After First Level of Clustering

This matrix is then examined for the most similar (least dissimilar) pairs in the same manner as the original matrix. This results in a fusion between 3 and 4. Elements 4 and 5 are also at the same distance, but cannot join because of the prior clustering of 4 with 3. The next fusions are, successively, 5 and (6,7), (9,10) and 11, 12 and (14,15). Here the "unweighted" aspect of this method first comes into play, for, so far the earlier clusters would have been the same in the weighted method. Thus, for example,

$$\begin{aligned} d_{5(12,14,15)} &= 0.33(d_{5,12} + d_{5,14} + d_{5,15}) \\ &= 0.33(2.0+3.162+3.605) \\ &= 2.923 \end{aligned}$$

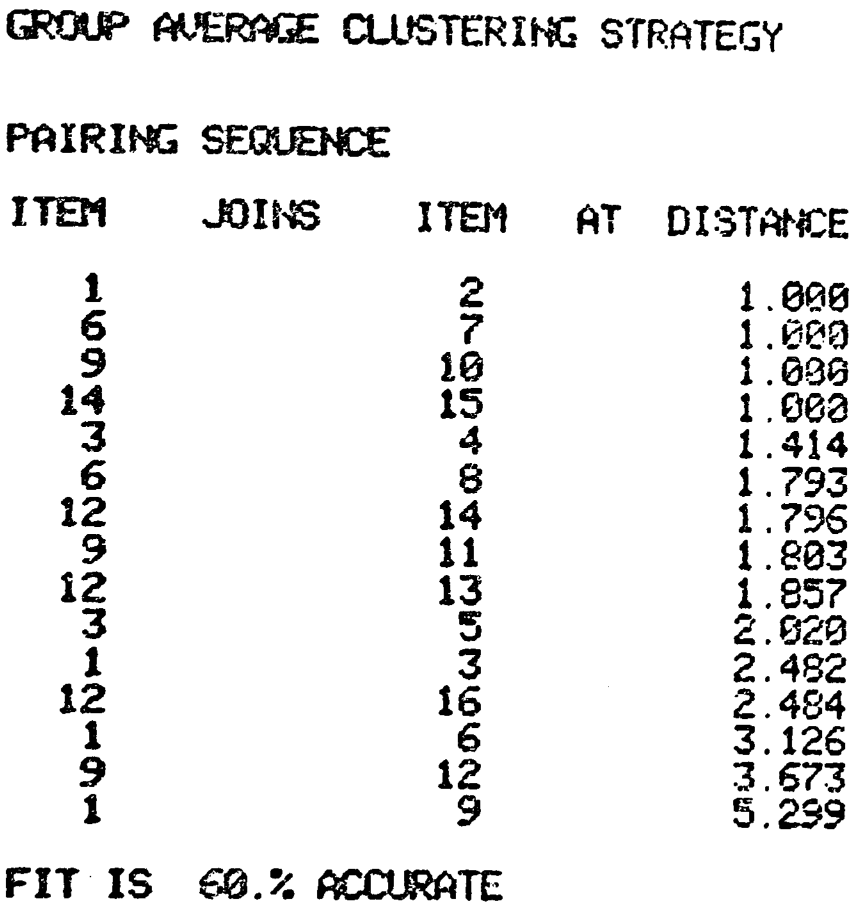
This is the noncombinatorial approach using the original data from figure 5.3, and is described here as it follows more closely the conceptual geometric method. To obtain the same result using Lance and Williams combinatorial equation (i.e. the method used in MAGIC) the formula

$$d_{hk} = (n_i d_{hi} + n_j d_{hj}) / n_k$$

applied to $d_{5(12,14,15)}$ becomes

$$\begin{aligned} d_{5(12,14,15)} &= \frac{((n_{12} \times d_{5(12)}) + (n_{(14,15)} \times d_{5(14,15)}))}{n_{12} + n_{(14,15)}} \\ &= \frac{1}{1+2}((1 \times 2) + (2 \times 3.3839)) \\ &= 2.923 \end{aligned}$$

which is as before, but here the distances are derived from figure 5.18. Continuing through the successive iterations we eventually obtain the dendrogram shown in figure 5.27.



```
GROUP AVERAGE CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM    JOINS    ITEM   AT  DISTANCE

  1                2          1.000
  6                7          1.000
  9               10          1.000
 14               15          1.000
  3                4          1.414
  6                8          1.793
 12               14          1.796
  9               11          1.803
 12               13          1.857
  3                5          2.020
  1                3          2.482
 12               16          2.484
  1                6          3.126
  9               12          3.673
  1                9          5.299

FIT IS  60.% ACCURATE
```

Figure 5.19

Group Average strategy - pairing sequence

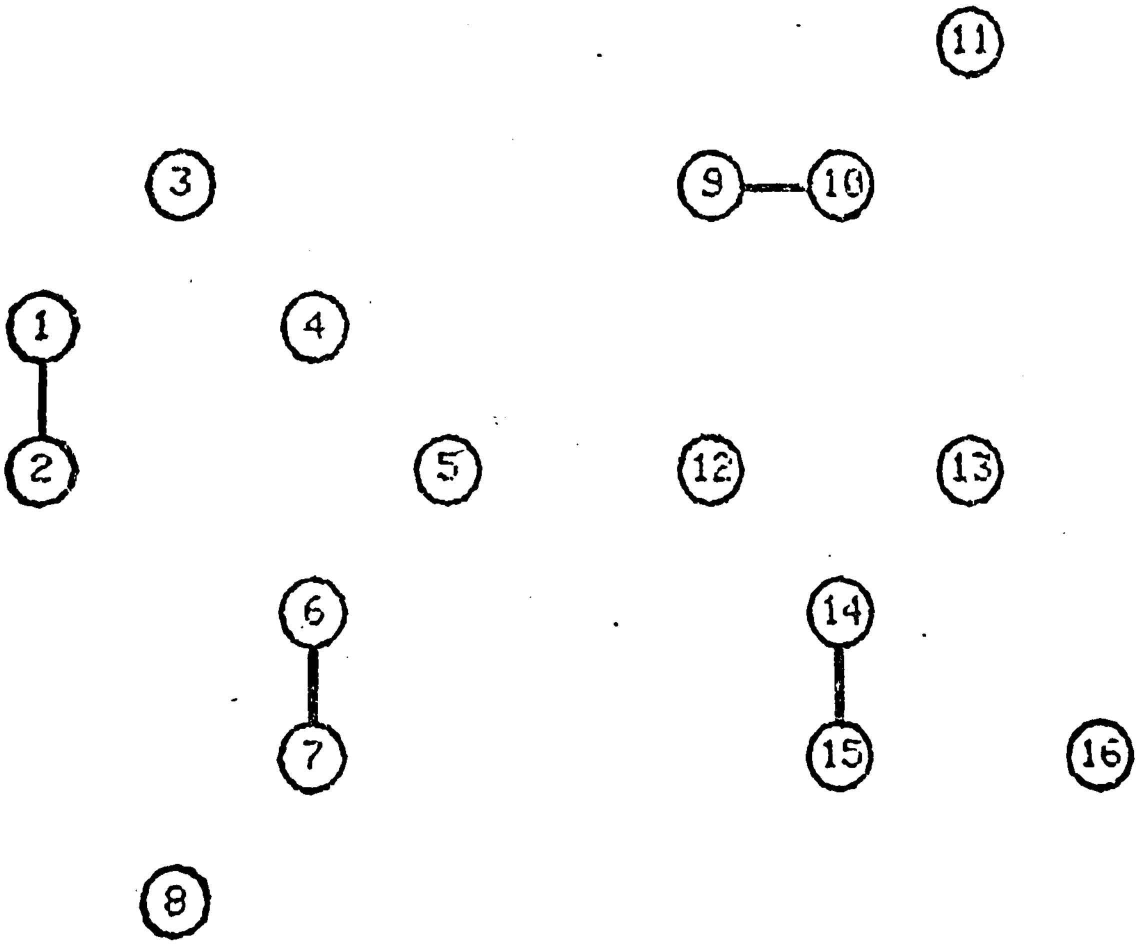


Figure 5.20

Group Average strategy - first step

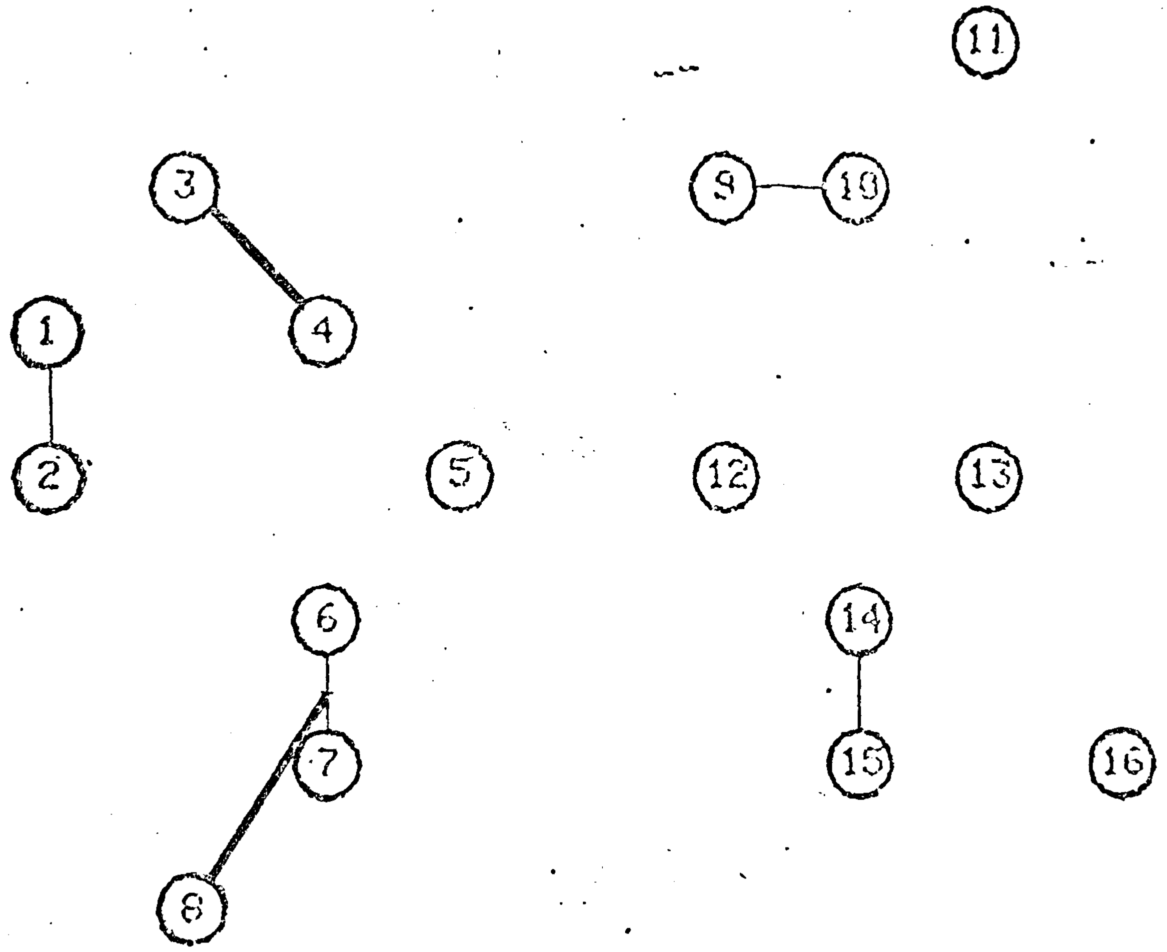


Figure 5.21

Group Average strategy - second and third steps

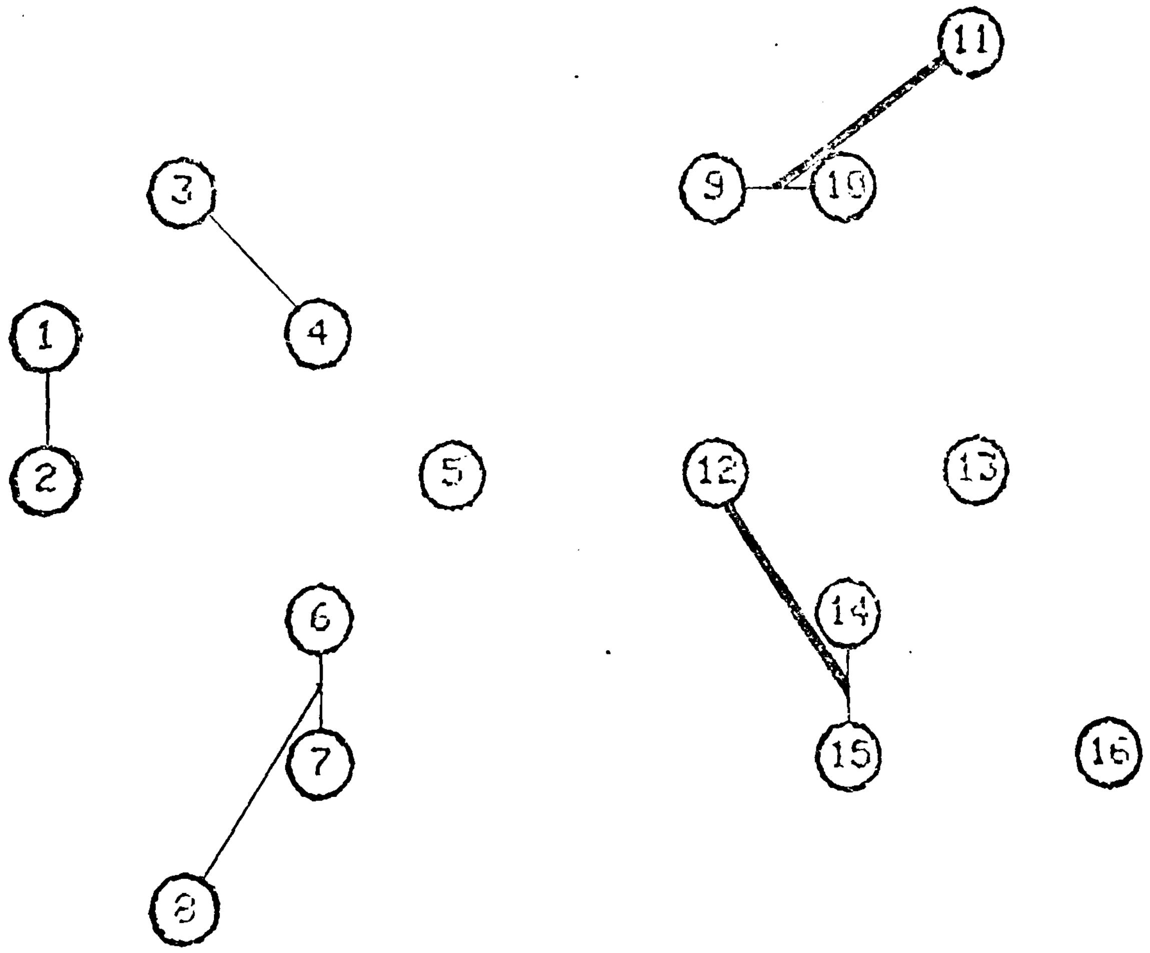


Figure 5.22

Group Average strategy - fourth and fifth steps

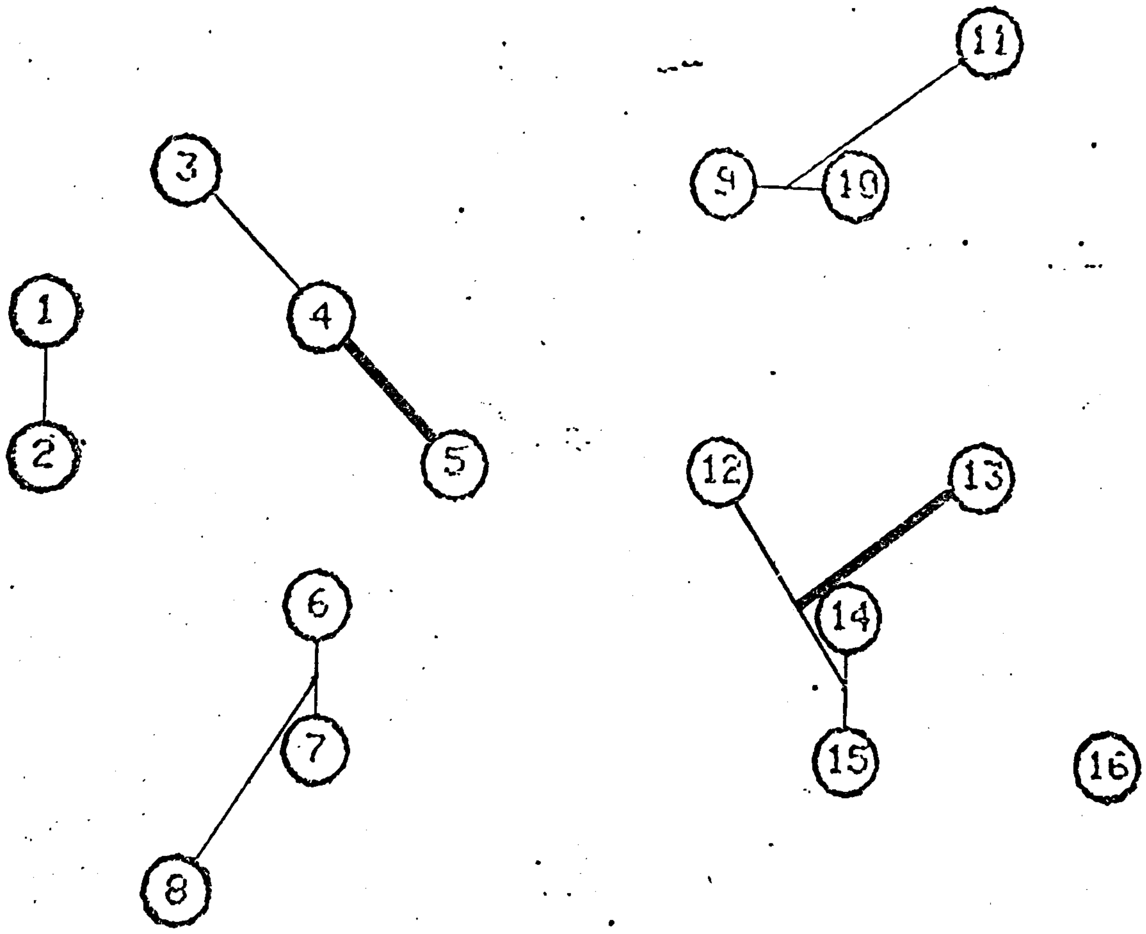


Figure 5.23

Group Average strategy - sixth and seventh steps

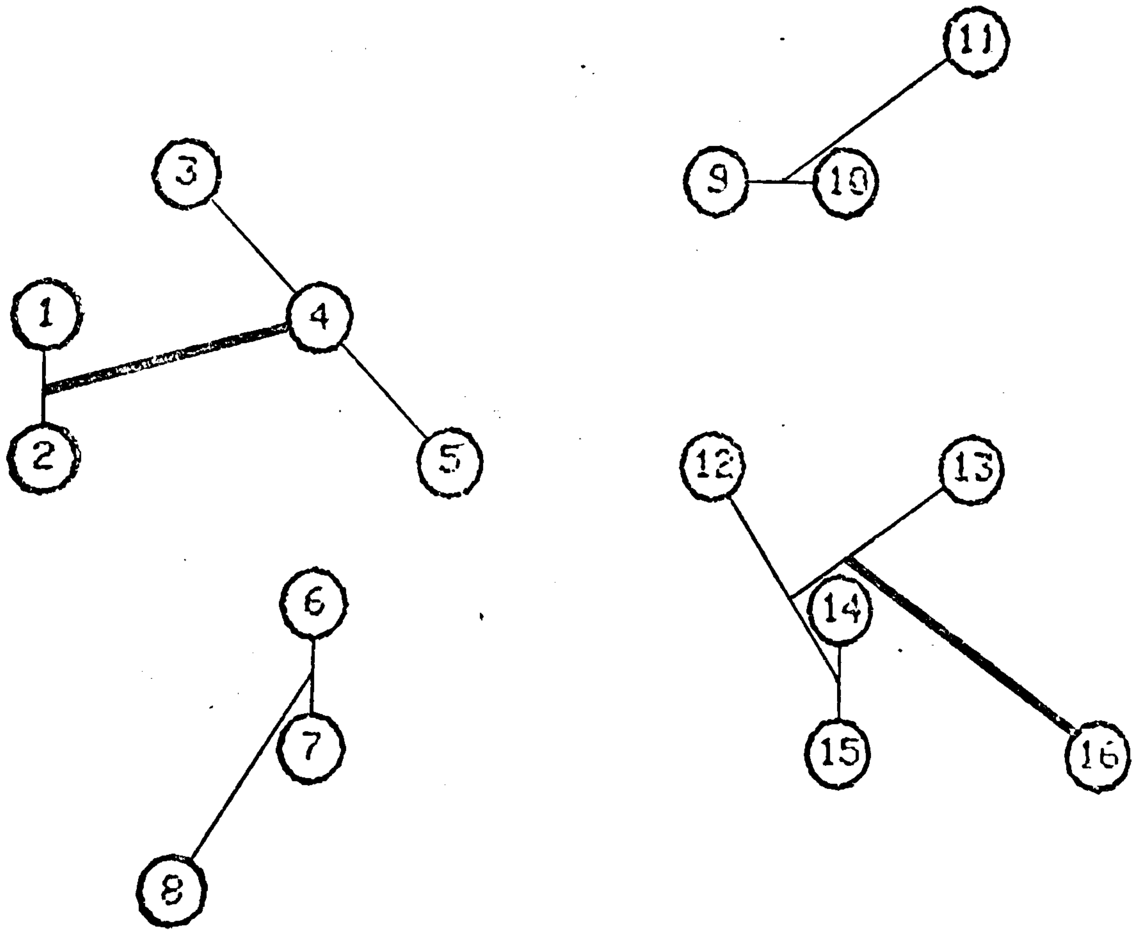


Figure 5.24

Group Average strategy - cigth and ninth steps

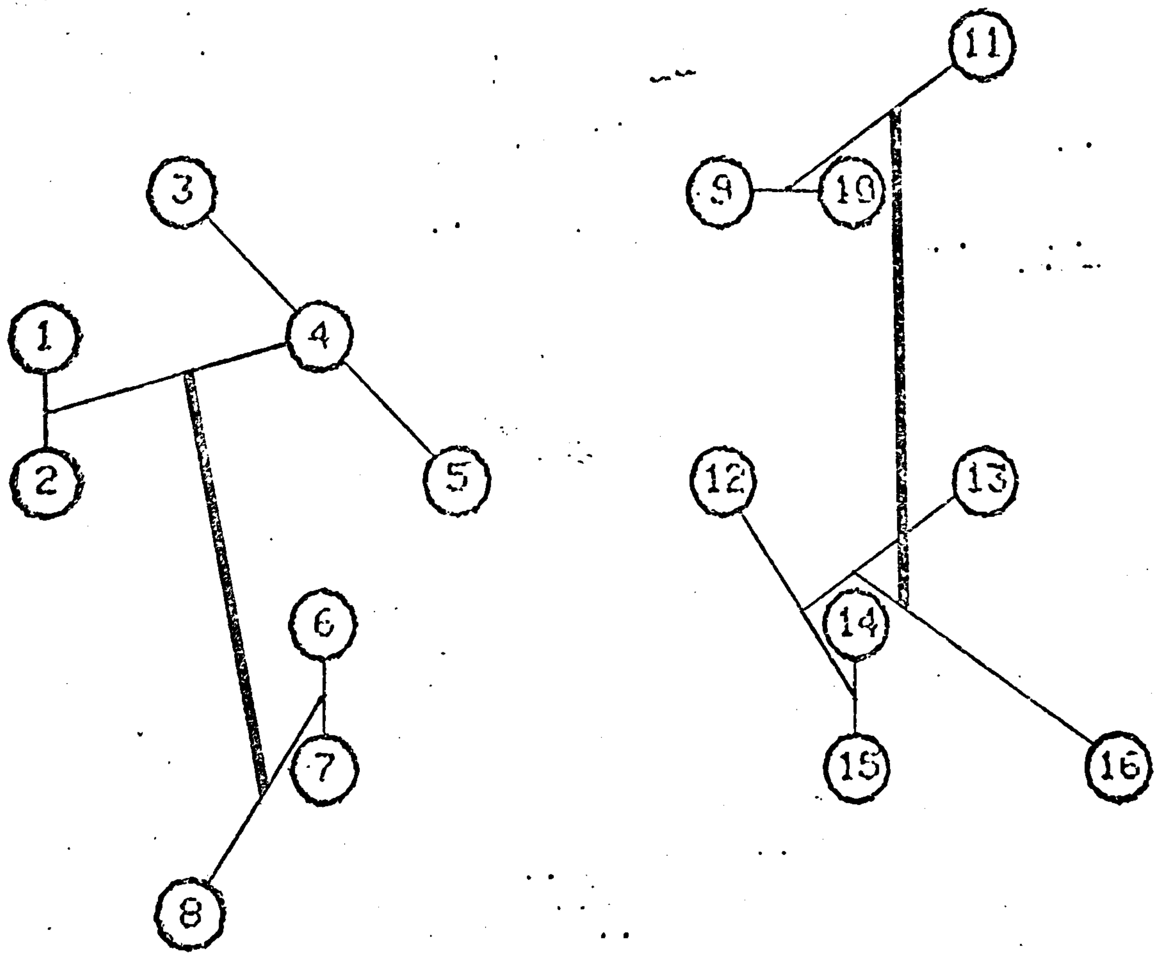


Figure 5.25

Group Average strategy - tenth and eleventh steps

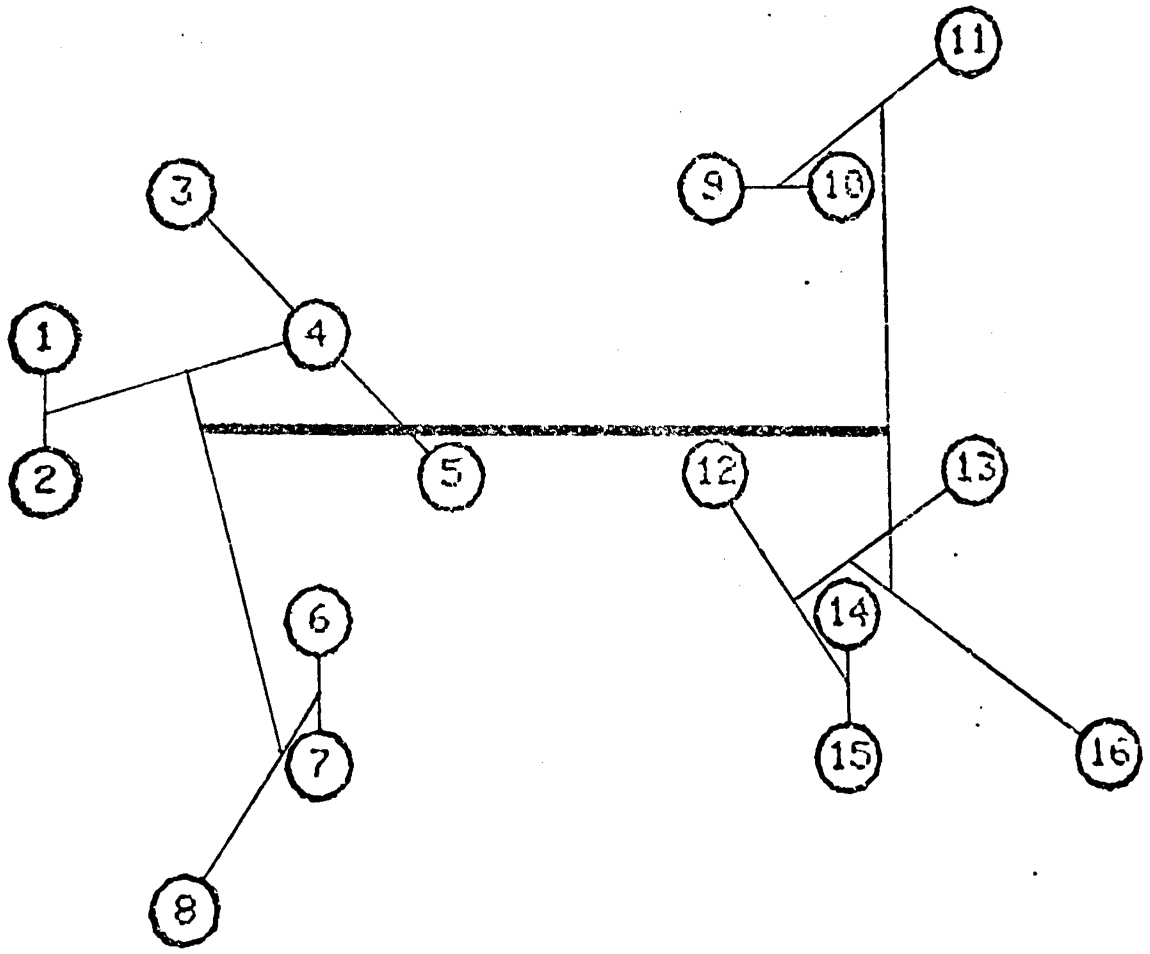


Figure 5.26

Group Average strategy - twelvth step

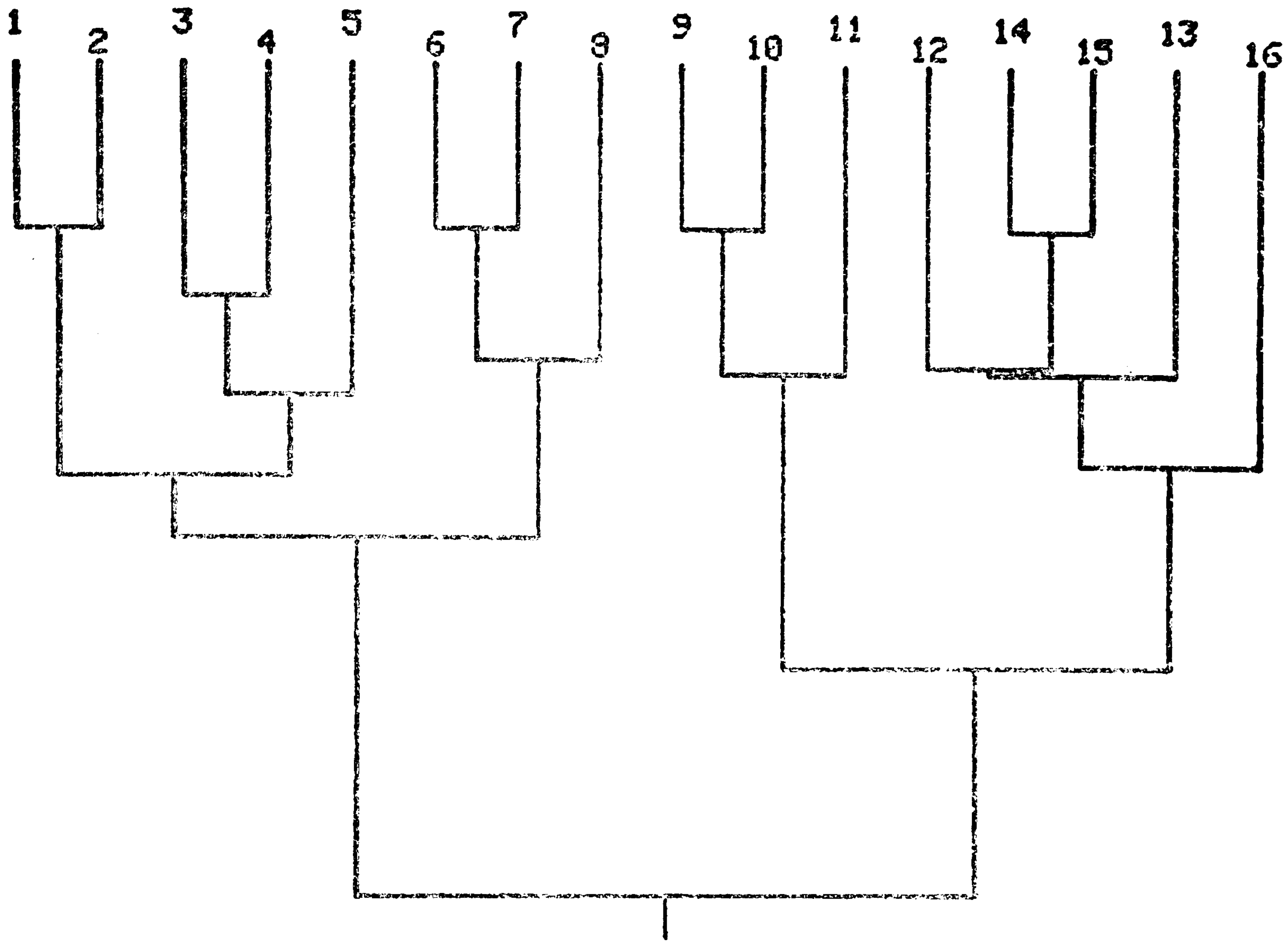


Figure 5.27

Group Average strategy - dendrogram

5.5 CENTROID CLUSTERING STRATEGY

Again the name was established by Lance and Williams (1967). The technique was previously known as the Unweighted Pairgroup Centroid Method (Sokal and Michener 1958, and King 1966,1967).

In a Euclidean model, the distance between two groups is defined as the distance between their centroids. It is combinatorial only when d^2 is used. The parameters are

$$\alpha_i = n_i/n_k;\ \alpha_j = n_j/n_k;\ \beta = -\alpha_i\alpha_j;\ \gamma = 0$$

$$d_{hk} = \frac{n_i d_{hi}}{n_k} + \frac{n_j d_{hj}}{n_k} - \frac{n_i n_j d_{ij}}{n_k^2}$$

$$= \frac{1}{n_k}\left(n_i d_{hi} + n_j d_{hj} - \frac{n_i n_j d_{ij}}{n_k}\right)$$

It is strictly space-conserving, but nonmonotonic, and reversals are frequent, thus rendering the strategy almost obsolete. It is conceptually attractive in that it computes cluster centroids, distances then being calculated between centroids, but disadvantages of nonmonotonicity outweigh this consideration.

Reversal can be seen in 13 joining (12,14,15) in the dendrogram (figure 5.37). Reversals occur when an element (or cluster) joins an existing cluster, but at a higher level of similarity than that at which the cluster had formed. As this is conceptual nonsense the technique has dropped into disuse.

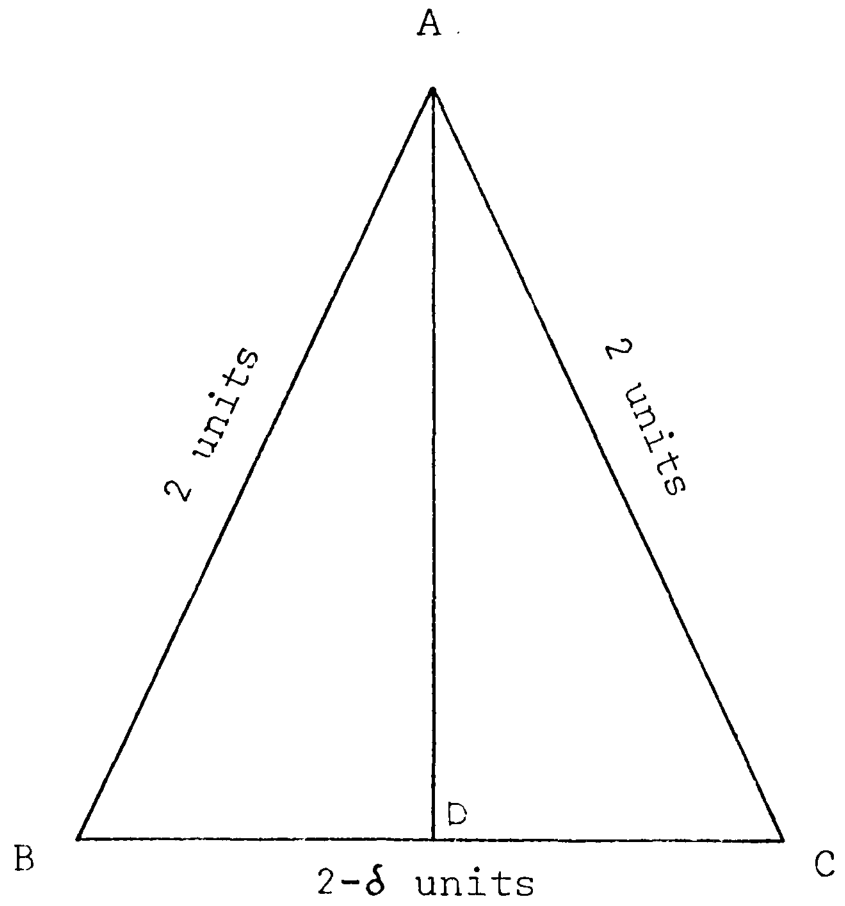
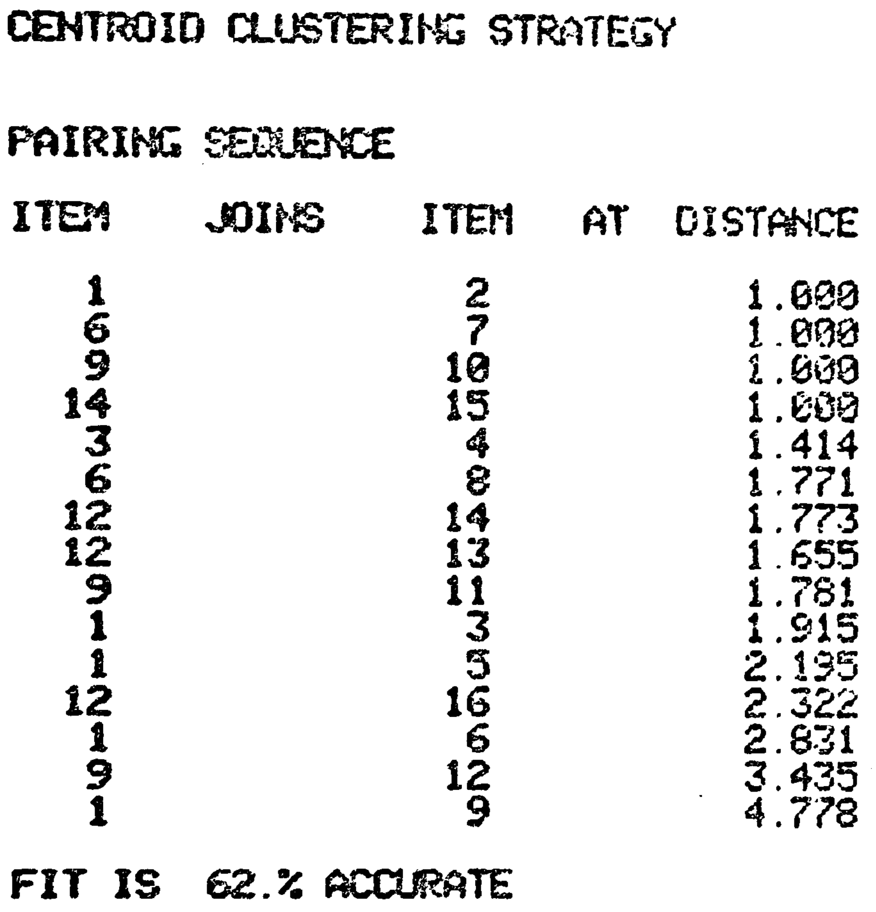


Figure 5.28

Inversion in Centroid clustering

An illustration of the manner in which Centroid clustering may lead to inversions when three almost equally dissimilar entities are clustered is shown in figure 5.28. A, B and C are the entities and D the product of the first fusion. The distance AD is now less than the length of any triangle side.



```
CENTROID CLUSTERING STRATEGY

PAIRING SEQUENCE
ITEM    JOINS    ITEM   AT  DISTANCE

  1                2         1.000
  6                7         1.000
  9               10         1.000
 14               15         1.000
  3                4         1.414
  6                8         1.771
 12               14         1.773
 12               13         1.655
  9               11         1.781
  1                3         1.915
  1                5         2.195
 12               16         2.322
  1                6         2.831
  9               12         3.435
  1                9         4.778

FIT IS  62.% ACCURATE
```

Figure 5.29

Centroid strategy - pairing sequence

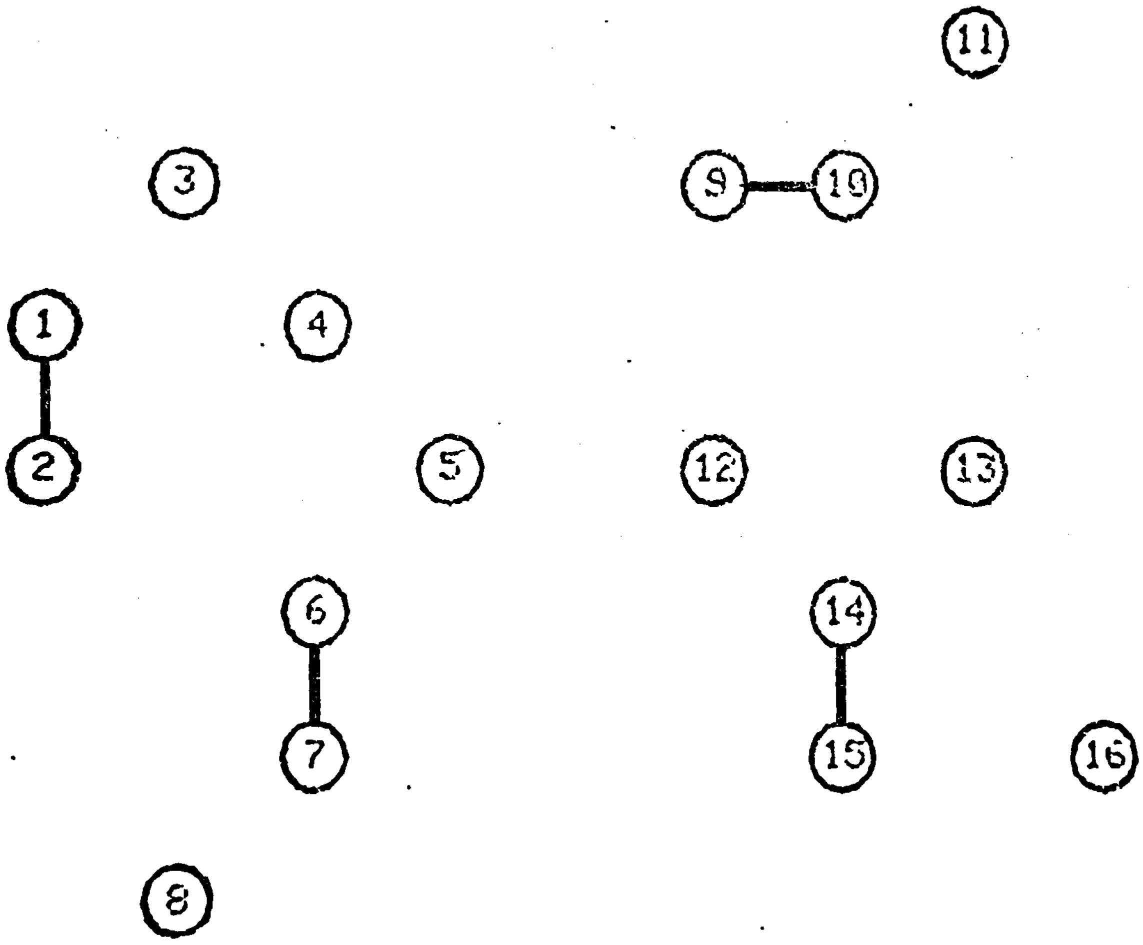


Figure 5.30

Centroid strategy - first step

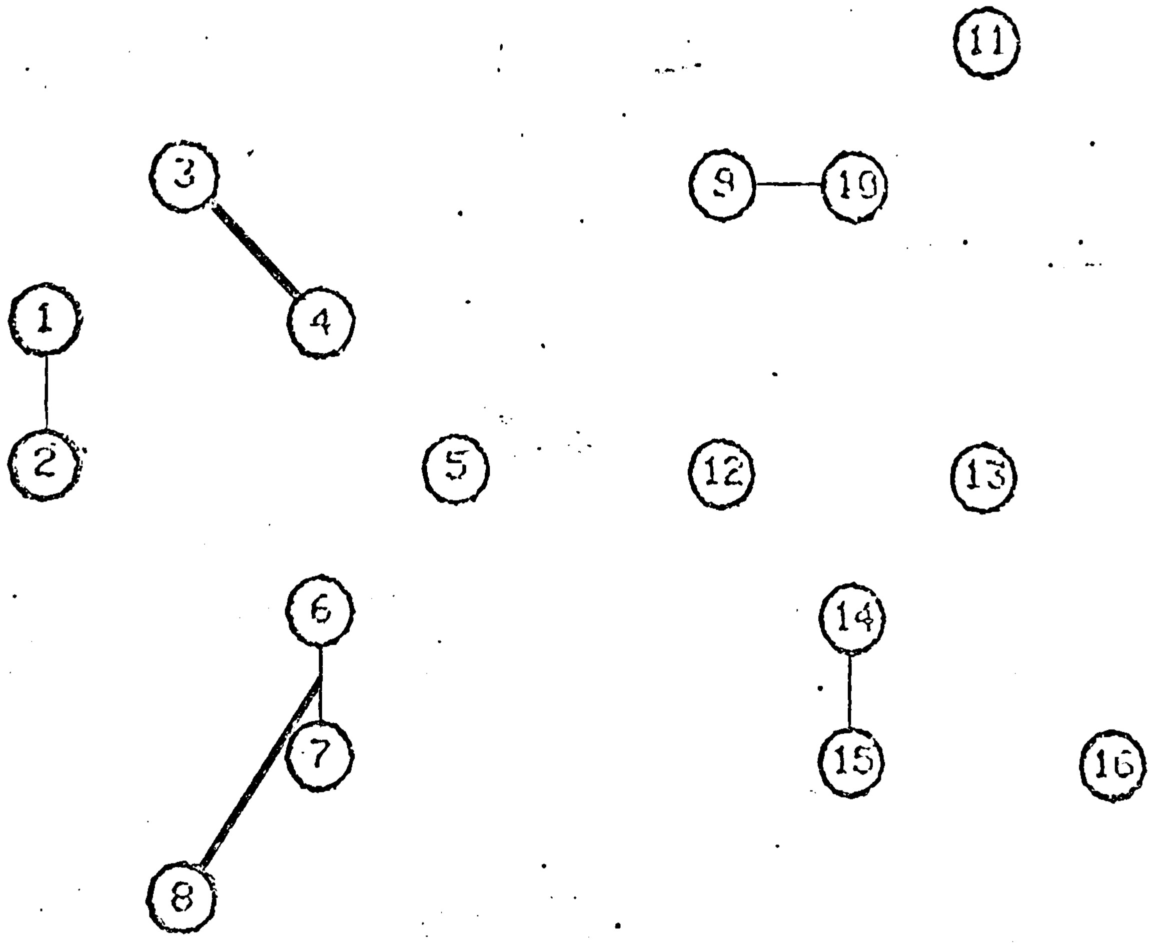


Figure 5.31

Centroid strategy - second and third steps

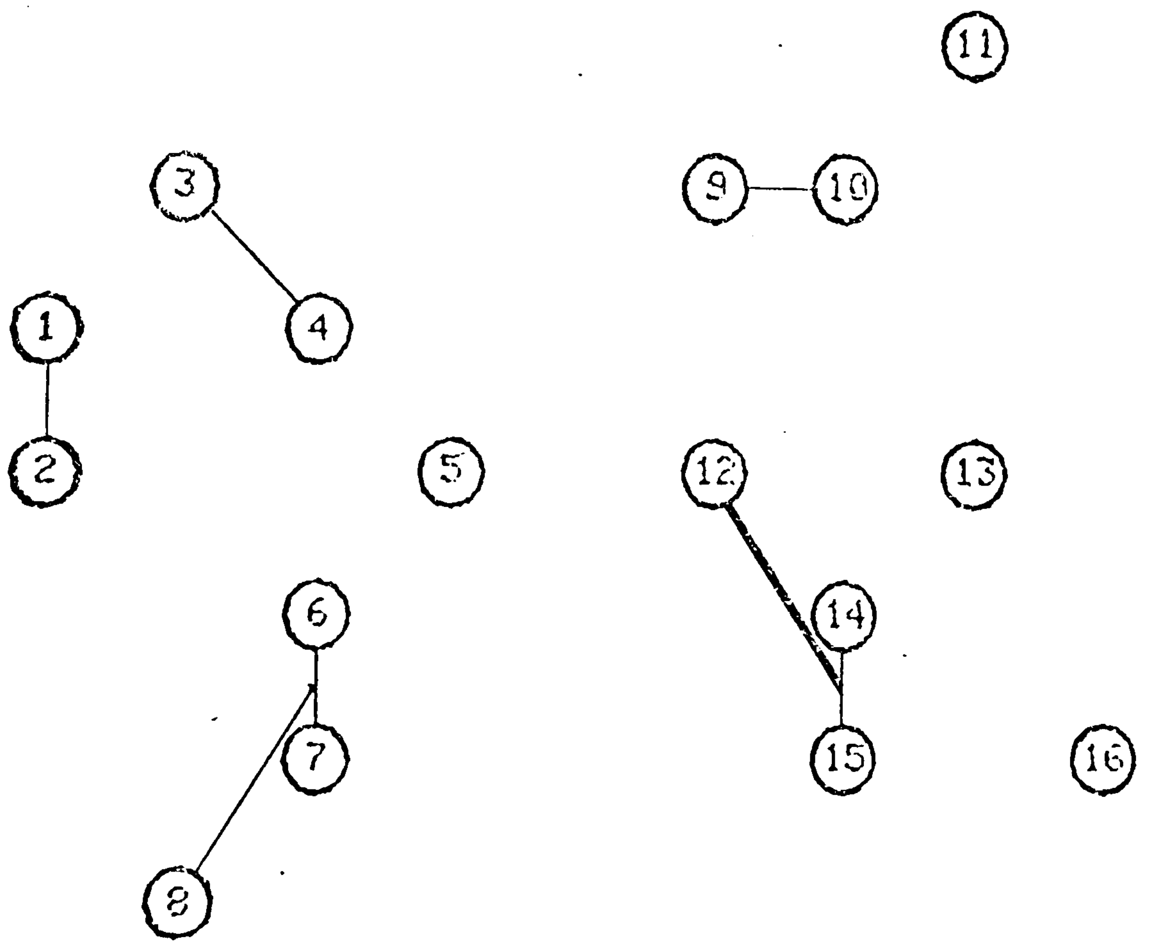


Figure 5.32

Centroid strategy - fourth step

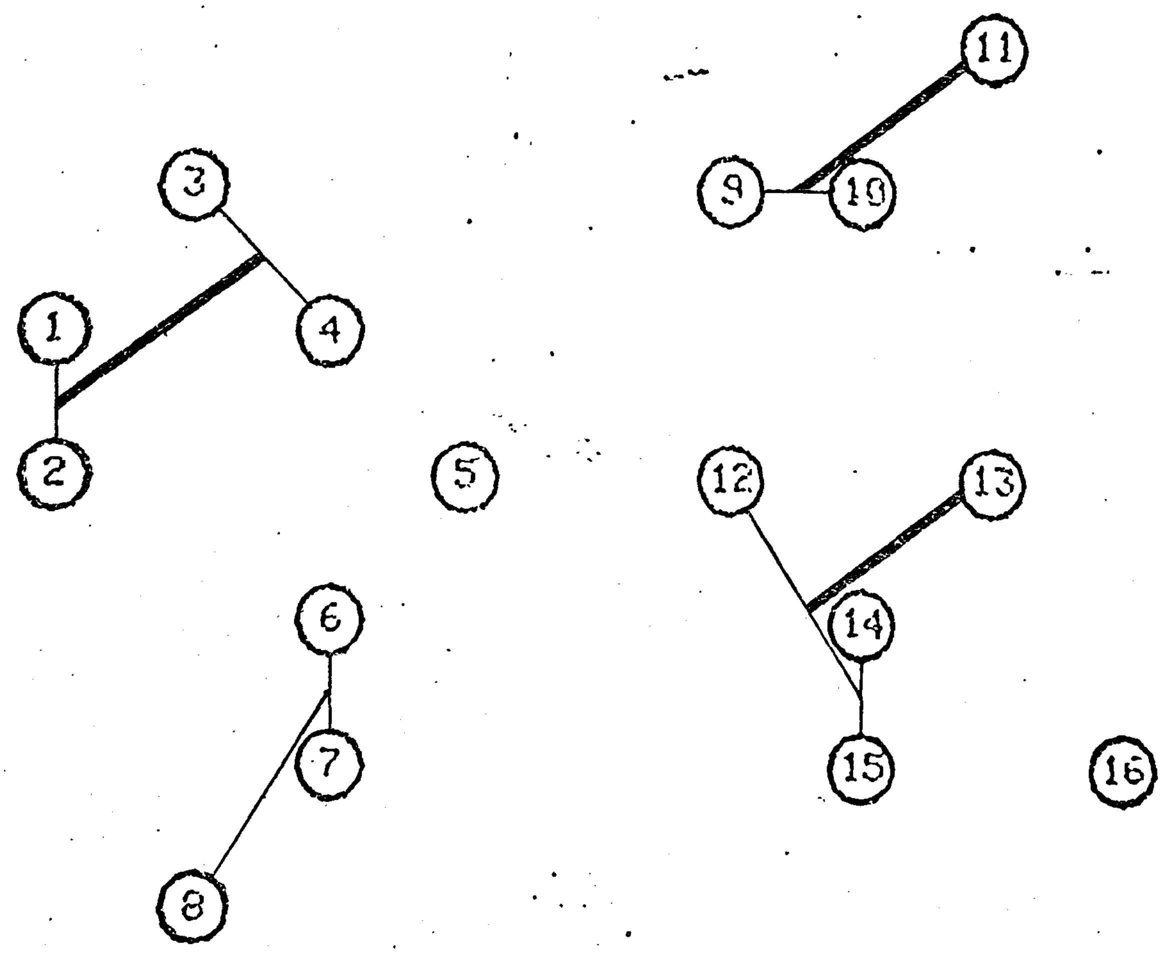


Figure 5.33

Centroid strategy - fifth, sixth and seventh steps

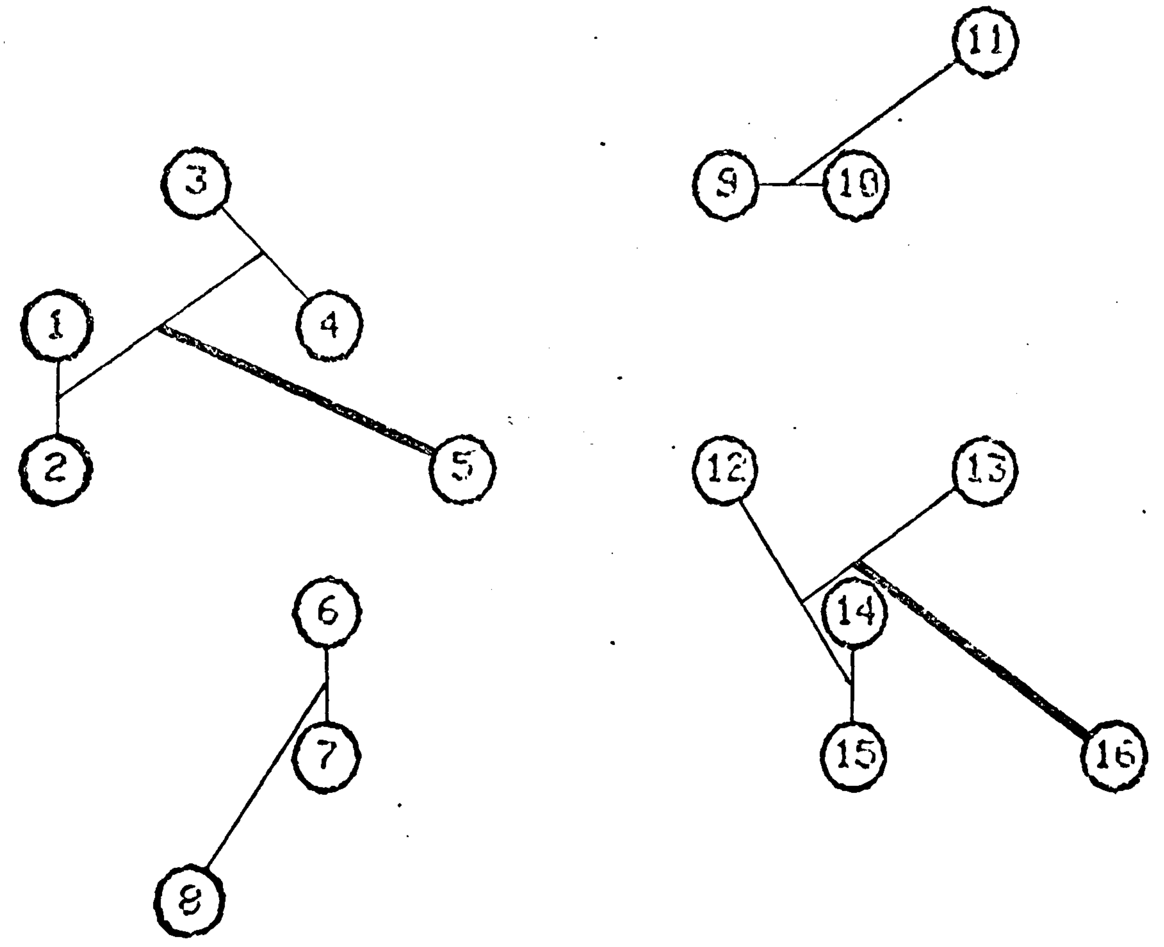


Figure 5.34

Centroid strategy - eigth and ninth steps

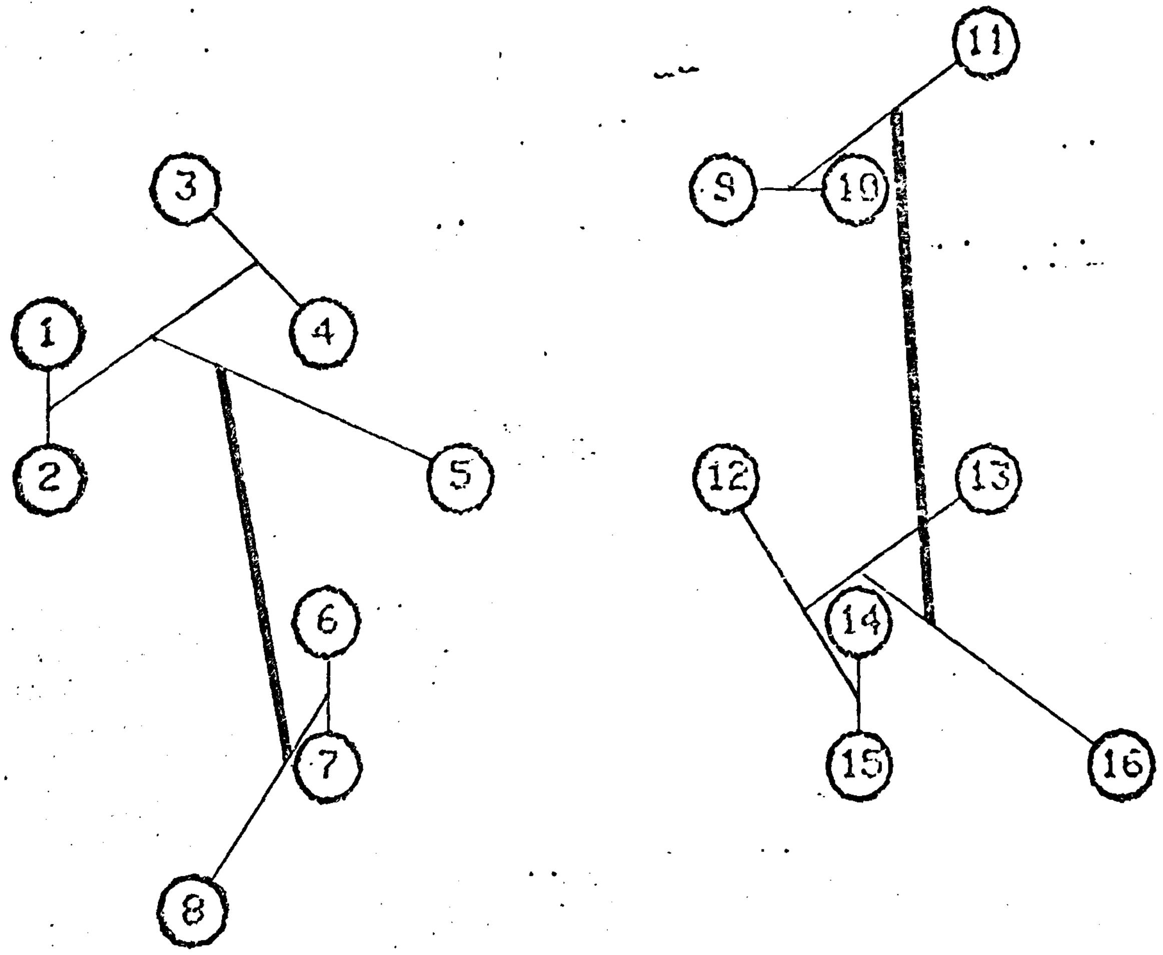


Figure 5.35

Gentroid strategy - tenth and eleventh steps

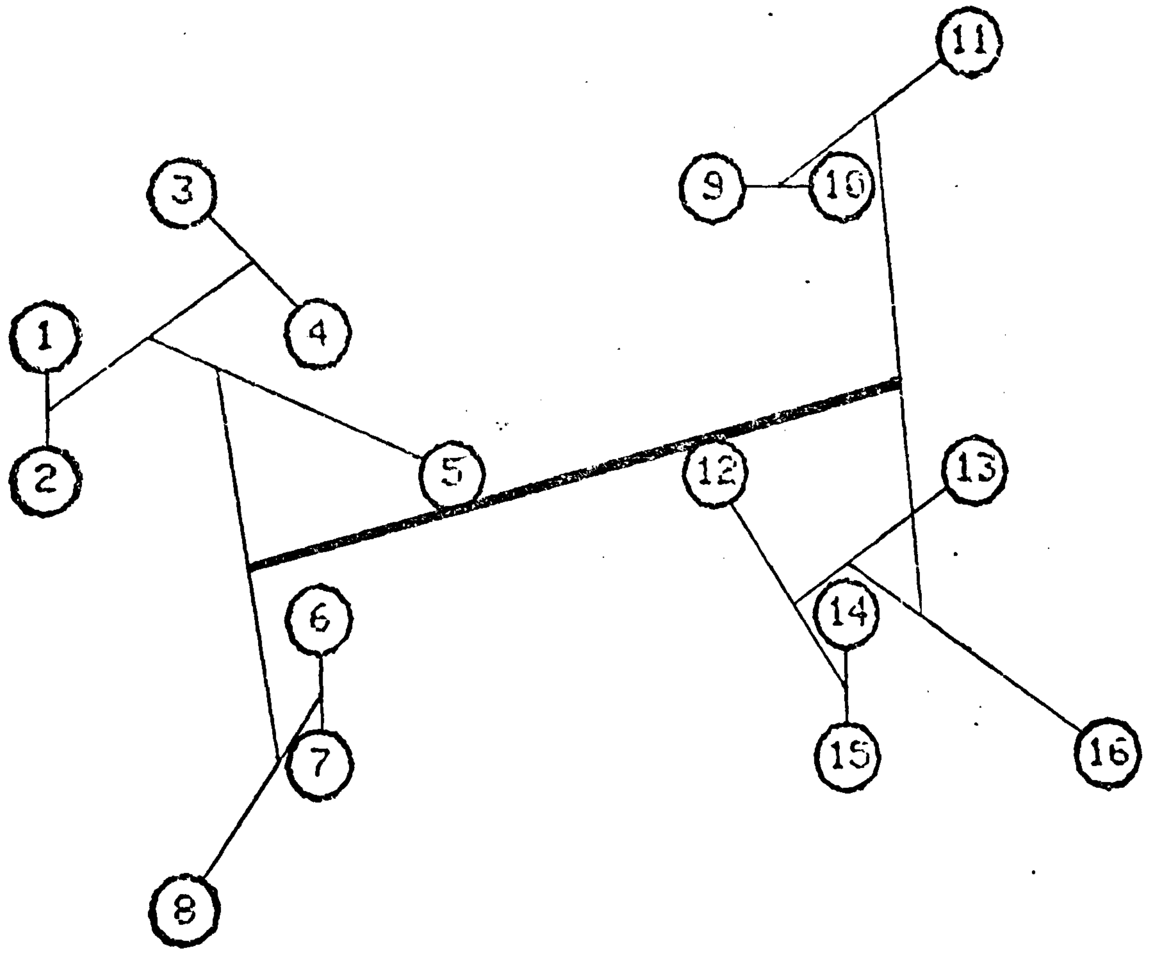


Figure 5.36

Centroid strategy - twelvth step

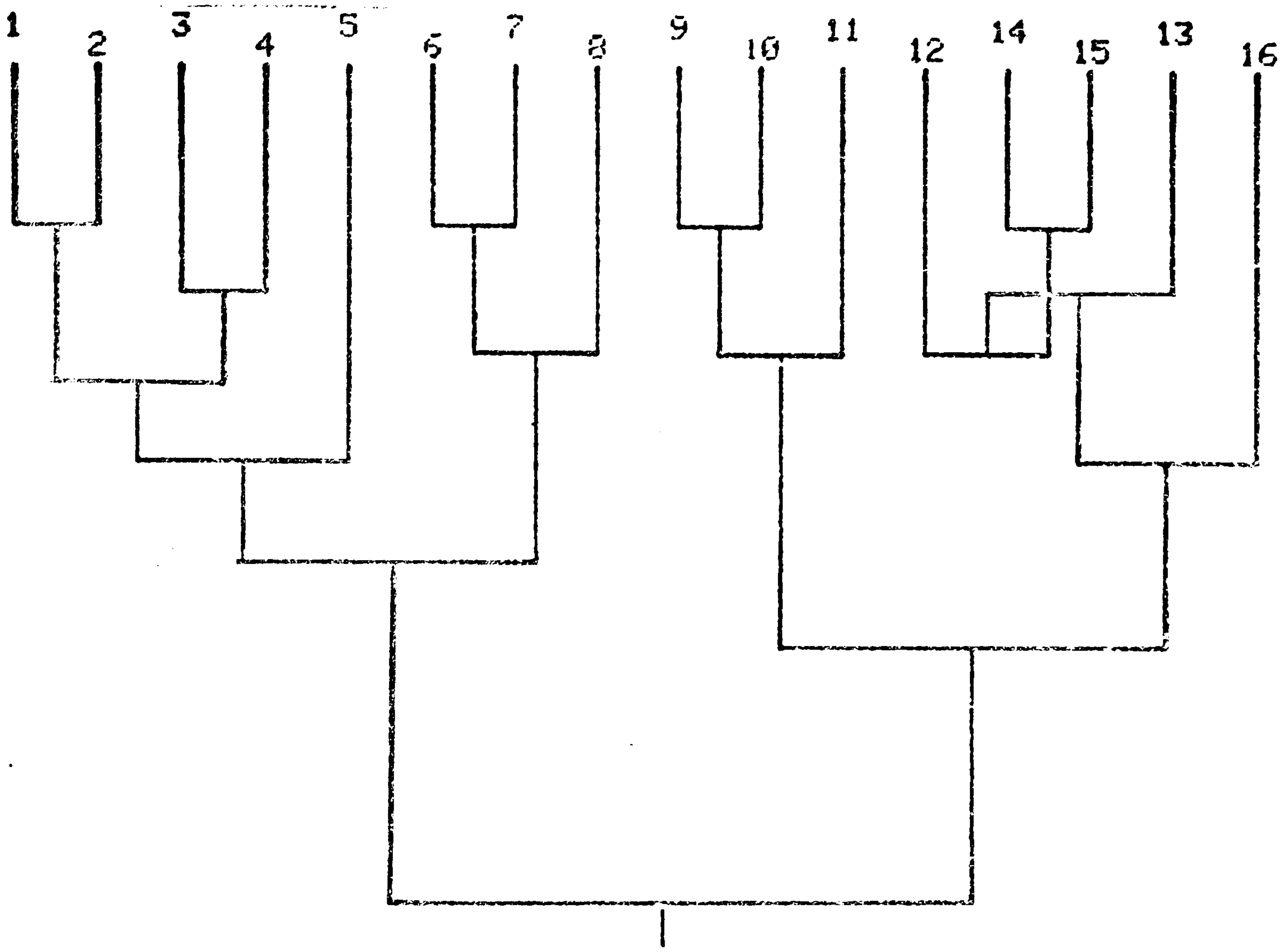


Figure 5.37

Centroid strategy - dendrogram

5.6 MEDIAN CLUSTERING STRATEGY

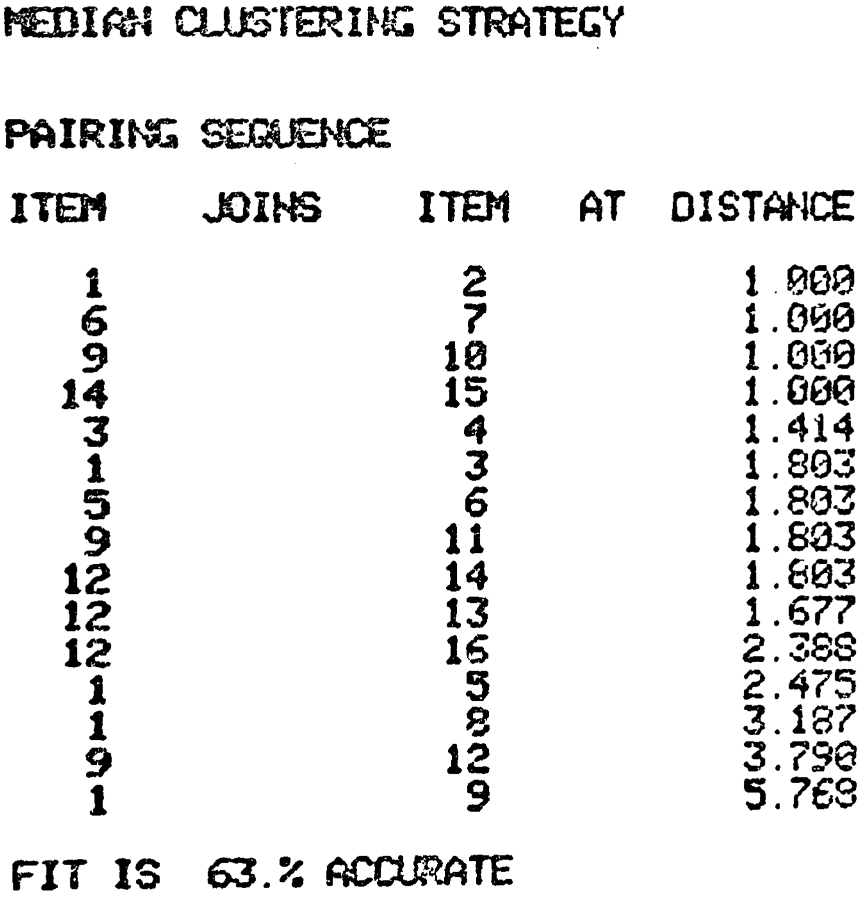
This method was proposed by Gower (1966), and previously known as the Weighted Pairgroup Centroid strategy (Sokal and Sneath 1967), the current name being established by Lance and Williams (1967).

A disadvantage of the Centroid strategy is that if n_i and n_j are very disparate, the centroid of (k) will lie close to that of the largest group, and remain within that group; the characteristic properties of the smaller group are thus lost. The strategy can be made independent of group size by arbitrarily setting $n_i=n_j$; the apparent position of (k) will thus always lie between (i) and (j). The parameters are

$$\alpha_i=\alpha_j=0.5;\ \beta=-0.25;\ \gamma=0$$

$$d_{hk}=0.5(d_{hi}+d_{hj})-0.25d_{ij}$$

The strategy is space-conserving but may be nonmonotonic.



```
MEDIAN CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT   DISTANCE

   1                  2             1.000
   6                  7             1.000
   9                 10             1.000
  14                 15             1.000
   3                  4             1.414
   1                  3             1.803
   5                  6             1.803
   9                 11             1.803
  12                 14             1.803
  12                 13             1.677
  12                 16             2.388
   1                  5             2.475
   1                  8             3.187
   9                 12             3.790
   1                  9             5.768

FIT IS  63.% ACCURATE
```

Figure 5.38

Median strategy - pairing sequence

5.7 INCREMENTAL SUM OF SQUARES CLUSTERING STRATEGY

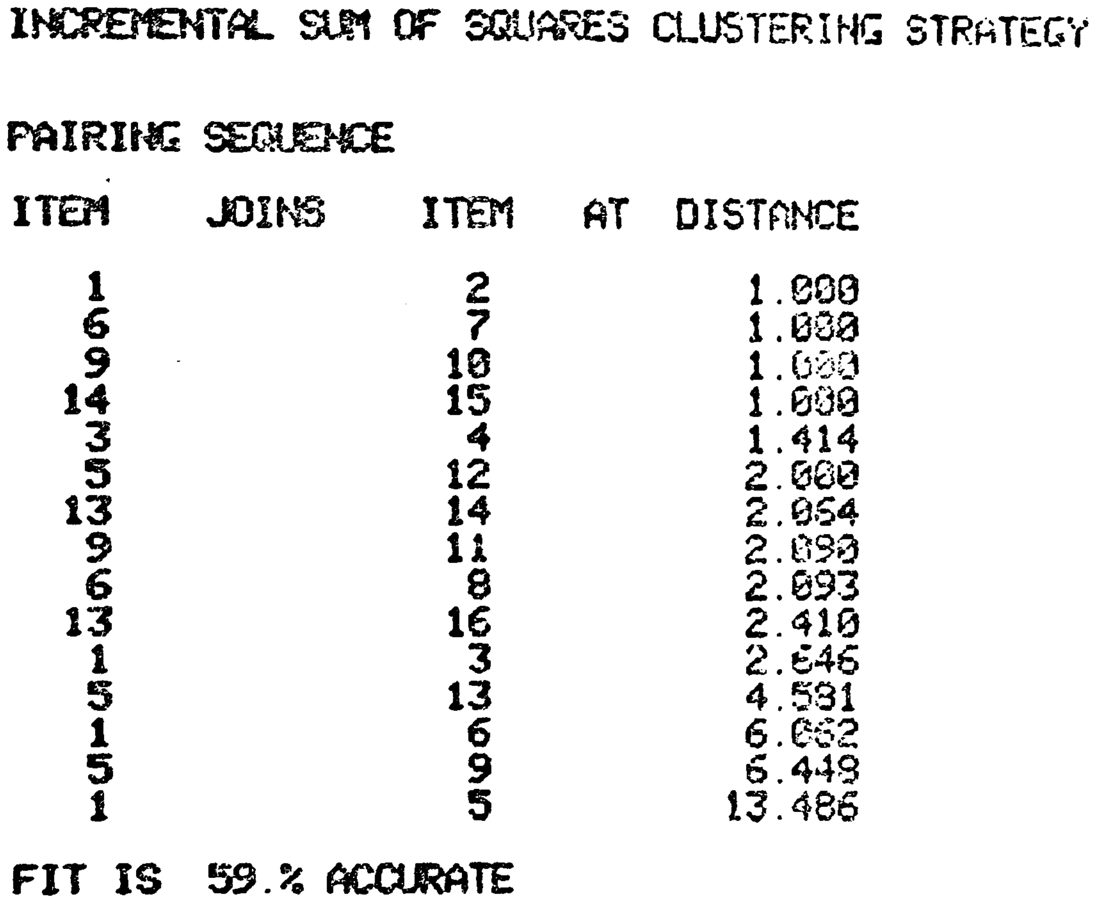
This technique has been known under a number of names: Error Sum of Squares (Ward 1963), Sum of Squares (Orloci 1967). The current name, which seems the most descriptive was proposed by Burr (1968, 1970).

In a Euclidean model, the intergroup distance is defined as the increase in the total within-group sum of squares (of distances from the respective centroids) on fusion. The parameters are:

$$\alpha_i=(n_h+n_i)/(n_h+n_k)$$
$$\alpha_j=(n_h+n_j)/(n_h+n_k)$$
$$\beta=-n_h/(n_h+n_k)$$
$$\gamma=0$$

$$d_{hk}=\frac{1}{(n_h+n_k)}\left((n_h+n_i)d_{hi}+(n_h+n_j)d_{hj}-n_h d_{ij}\right)$$

It is monotonic and space-dilating. Squares of Euclidean distance (D^2) are used as distance measures and after uniting the pair of elements whose D^2 is a minimum, subsequent entities are fused such that the sum of D^2 within a cluster increases by the smallest amount. As the total sum of squares is constant, if the sum of D^2 within a cluster increases minimally, then it follows that D^2 between clusters is increased maximally.



```
INCREMENTAL SUM OF SQUARES CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM     JOINS     ITEM   AT  DISTANCE

   1                  2          1.000
   6                  7          1.000
   9                 10          1.000
  14                 15          1.000
   3                  4          1.414
   5                 12          2.000
  13                 14          2.054
   9                 11          2.090
   6                  8          2.093
  13                 16          2.410
   1                  3          2.646
   5                 13          4.531
   1                  6          6.062
   5                  9          6.449
   1                  5         13.486

FIT IS  59.% ACCURATE
```

Figure 5.39

I.S.S. strategy - pairing sequence

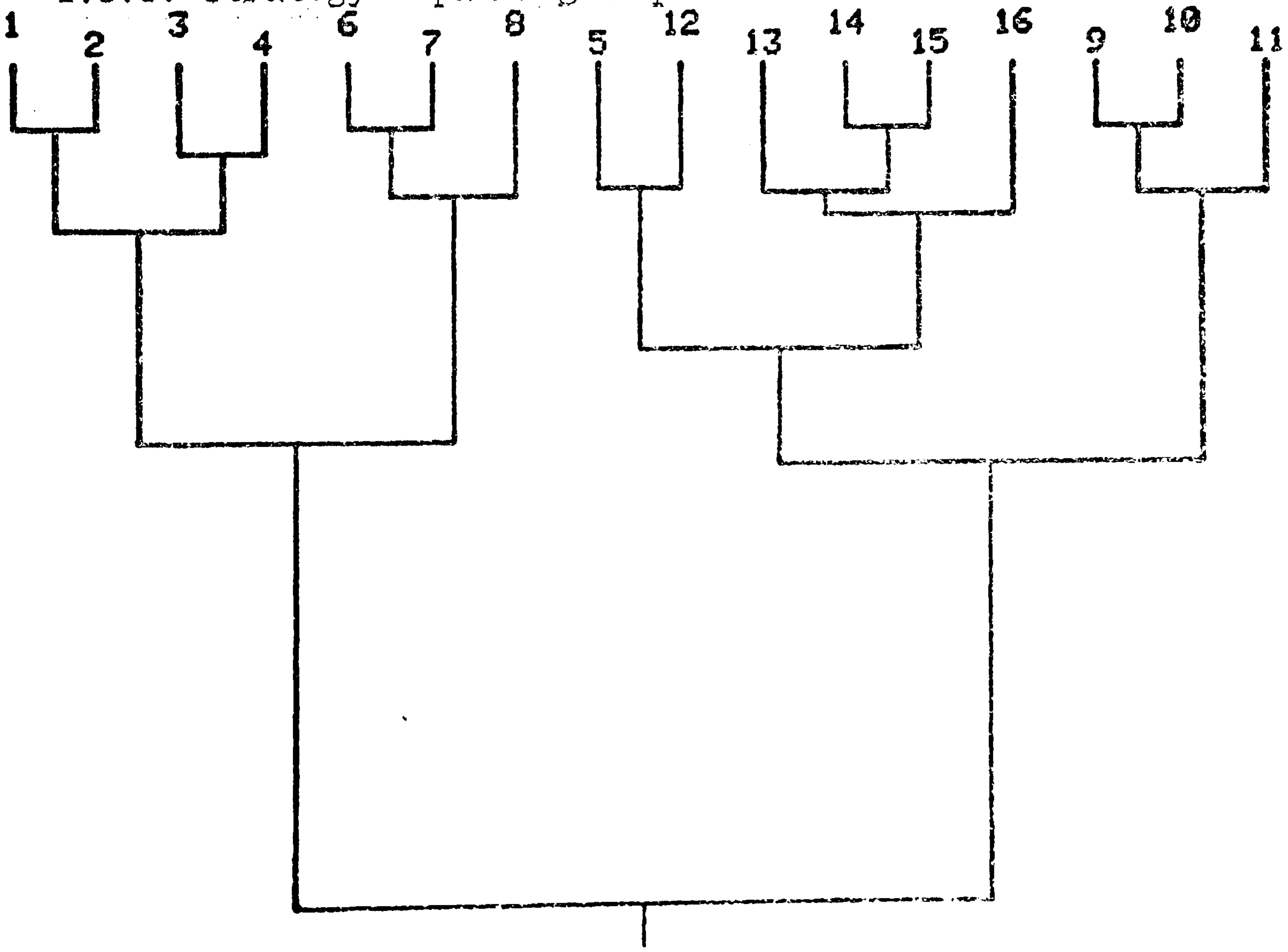


Figure 5.40

I.S.S. strategy - dendrogram

5.8 SIMPLE AVERAGE CLUSTERING STRATEGY

Also known as the Weighted Pairgroup Method Using Arithmetic Averages (Sokal and Sneath 1967), this strategy has a similar relationship to the Group Average method as the Median has to the Centroid. The two groups are given equal weight with n_i set equal to n_j. The method is space-dilating and monotonic. Parameters are:

$$\alpha_i = \alpha_j = 0.5 \text{ and } \beta = \gamma = 0$$
$$d_{hk} = 0.5(d_{hi} + d_{hj})$$

```
SIMPLE AVERAGE CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

  1                  2          1.000
  6                  7          1.000
  9                 10          1.000
 14                 15          1.000
  3                  4          1.414
  5                  6          1.871
  9                 11          1.871
 12                 14          1.871
 12                 13          1.936
  1                  3          2.000
 12                 16          2.622
  1                  5          2.872
  1                  8          3.571
  9                 12          4.161
  1                  9          6.494

FIT IS  63.% ACCURATE
```

Figure 5.41

Simple Average strategy - pairing sequence

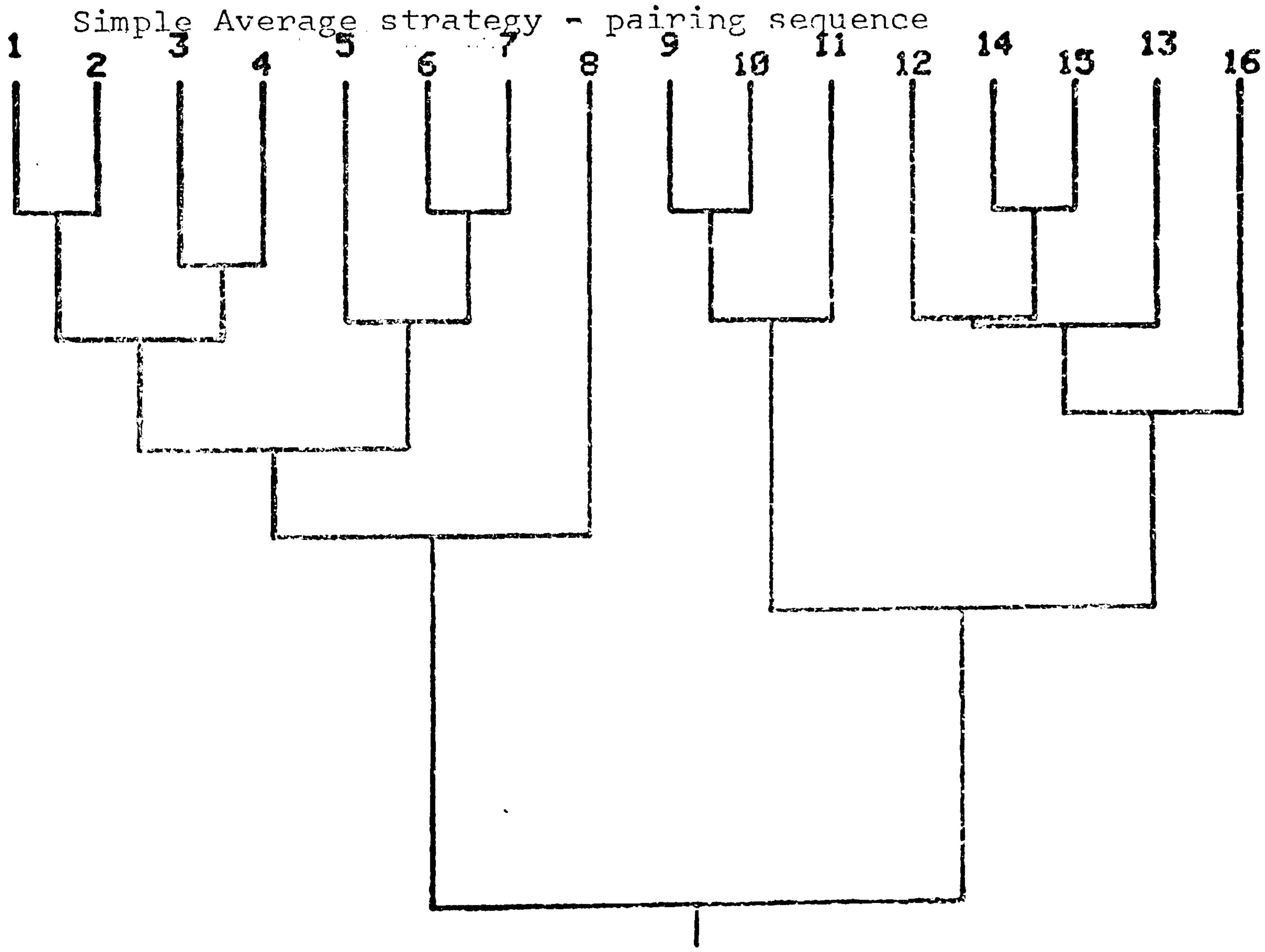


Figure 5.42

Simple Average strategy - dendrogram

5.9 FLEXIBLE STRATEGY CLUSTERING

This is applicable to any dissimilarity measure and is defined by the quadruple constraint:

$$\alpha_i + \alpha_j + \beta = 1$$
$$\alpha_i = \alpha_j$$
$$\beta < 1$$
$$\gamma = 0$$

It is monotonic and its space distorting properties depend entirely upon β. If $\beta=0$ the strategy is space-conserving; as β becomes positive the strategy becomes increasingly space-contracting; as β becomes negative the strategy becomes increasingly space-dilating. In practice a value of $\beta=-0.25$ is commonly used, giving $d_{hk}=0.625(d_{hi}+d_{hj})-0.25d_{ij}$ and thus bearing some resemblance to the Median strategy. Given any value of β the other parameters follow automatically from the constraints.

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.625
BETA =   -0.250


PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

  1                  2           1.000
  6                  7           1.000
  9                 10           1.000
 14                 15           1.000
  3                  4           1.414
  5                 12           2.000
  6                  8           2.031
  9                 11           2.031
 13                 14           2.031
  1                  3           2.332
 13                 16           2.335
  5                  6           4.438
  1                  5           5.028
  9                 13           5.747
  1                  9           9.870

FIT IS  58.% ACCURATE
```

Figure 5.43

Flexible strategy (β = -0.25) - pairing sequence

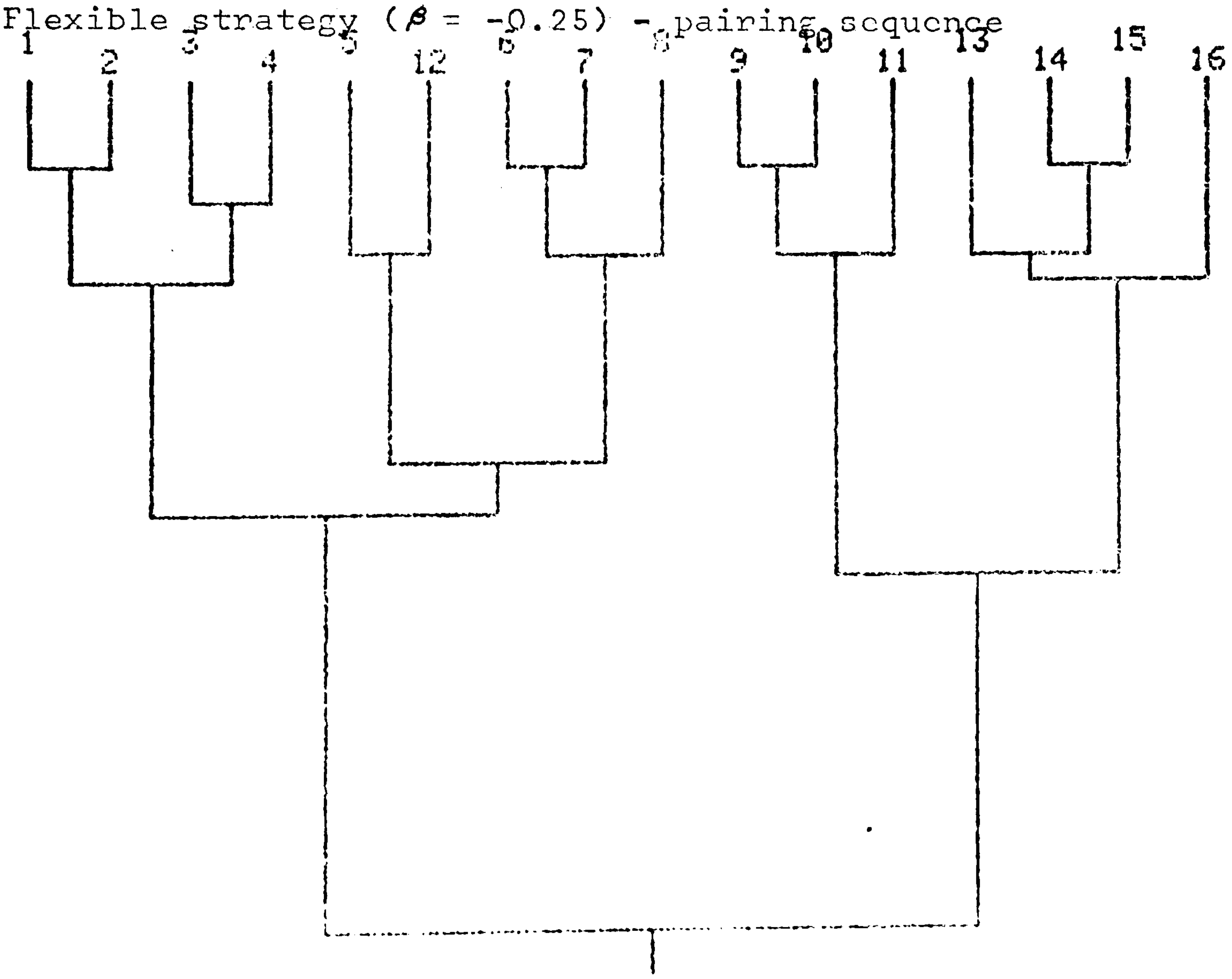


Figure 5.44

Flexible strategy (β = -0.25) - dendrogram

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.990
BETA =   -0.980


PAIRING SEQUENCE

ITEM    JOINS    ITEM   AT  DISTANCE

  1                2          1.000
  6                7          1.000
  9               10          1.000
 14               15          1.000
  3                4          1.414
  5               12          2.000
 13               16          2.236
  6                8          2.439
  9               11          2.439
 13               14          2.973
  1                3          3.432
  5                6          7.130
  1                5         11.634
  9               13         12.075
  1                9         33.730

FIT IS  50.% ACCURATE
```

Figure 5.45

Flexible strategy (β = -0.98) - pairing sequence

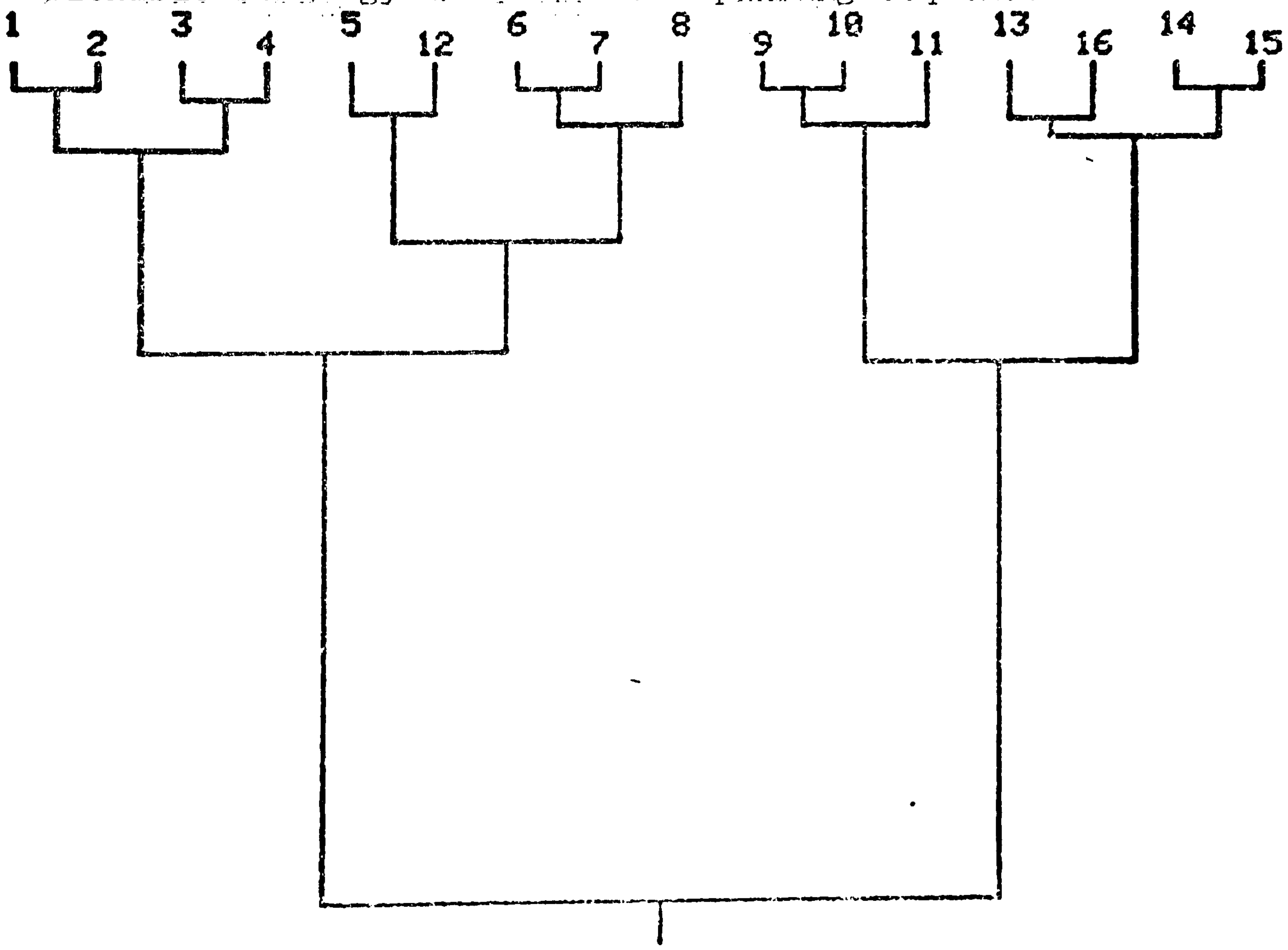


Figure 5.46

Flexible strategy (β = -0.98) - dendrogram

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.750
BETA =   -0.500


PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

  1                  2           1.000
  6                  7           1.000
  9                 10           1.000
 14                 15           1.000
  3                  4           1.414
  5                 12           2.000
  6                  8           2.179
  9                 11           2.179
 13                 14           2.179
 13                 16           2.462
  1                  3           2.693
  5                  6           5.300
  1                  5           6.823
  9                 13           7.548
  1                  9          15.714

FIT IS  58.% ACCURATE
```

Figure 5.47

Flexible strategy (β = -0.5) - pairing sequence

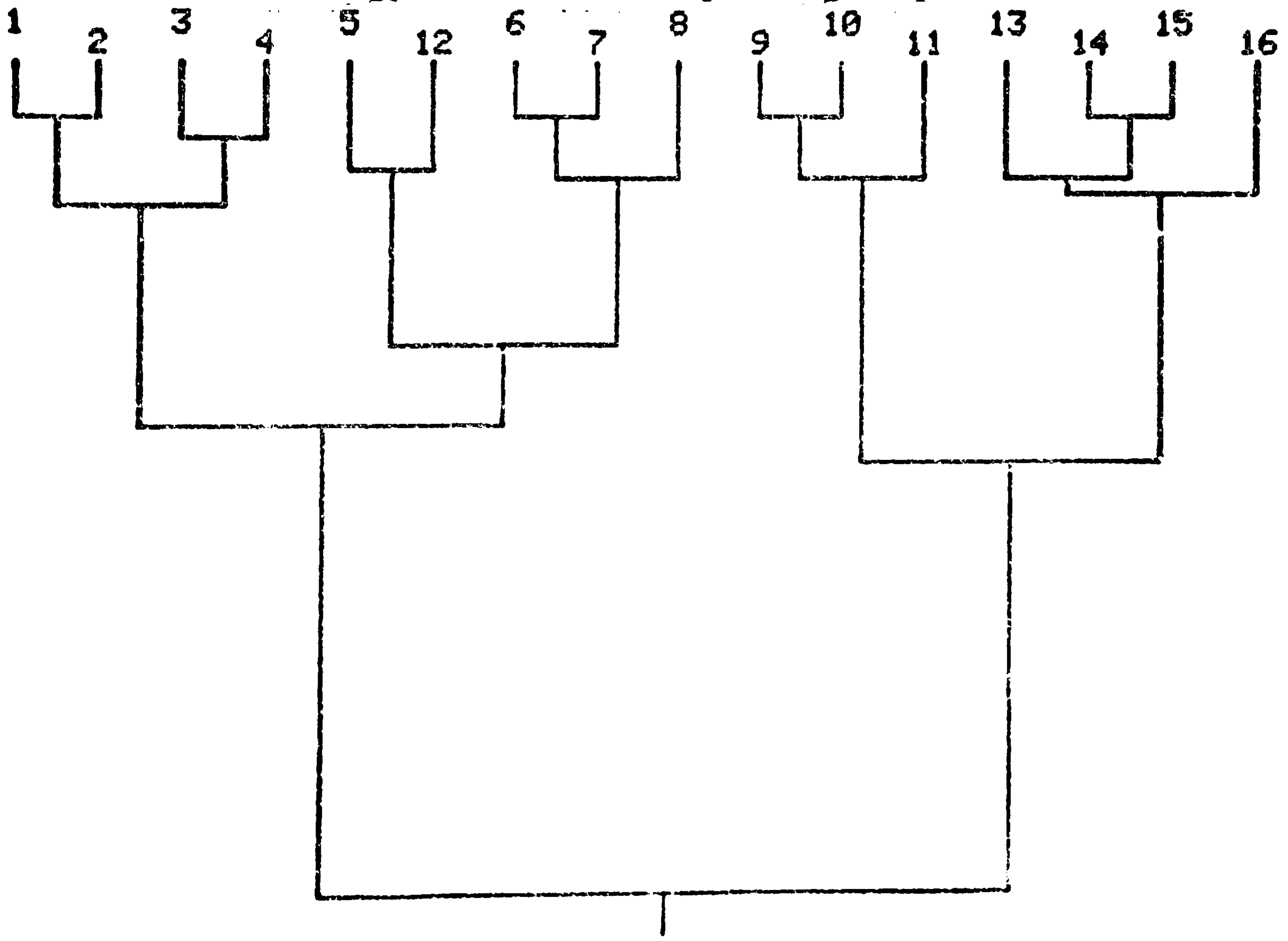


Figure 5.48
Flexible strategy (β = -0.5) - dendrogram

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.500
BETA =    0.000


PAIRING SEQUENCE

ITEM    JOINS    ITEM   AT  DISTANCE

  1                2          1.000
  6                7          1.000
  9               10          1.000
 14               15          1.000
  3                4          1.414
  5                6          1.871
  9               11          1.871
 12               14          1.871
 12               13          1.936
  1                3          2.000
 12               16          2.622
  1                5          2.872
  1                8          3.571
  9               12          4.161
  1                9          6.494

FIT IS  63.% ACCURATE
```

Figure 5.49
Flexible strategy (β = 0.0) - pairing sequence

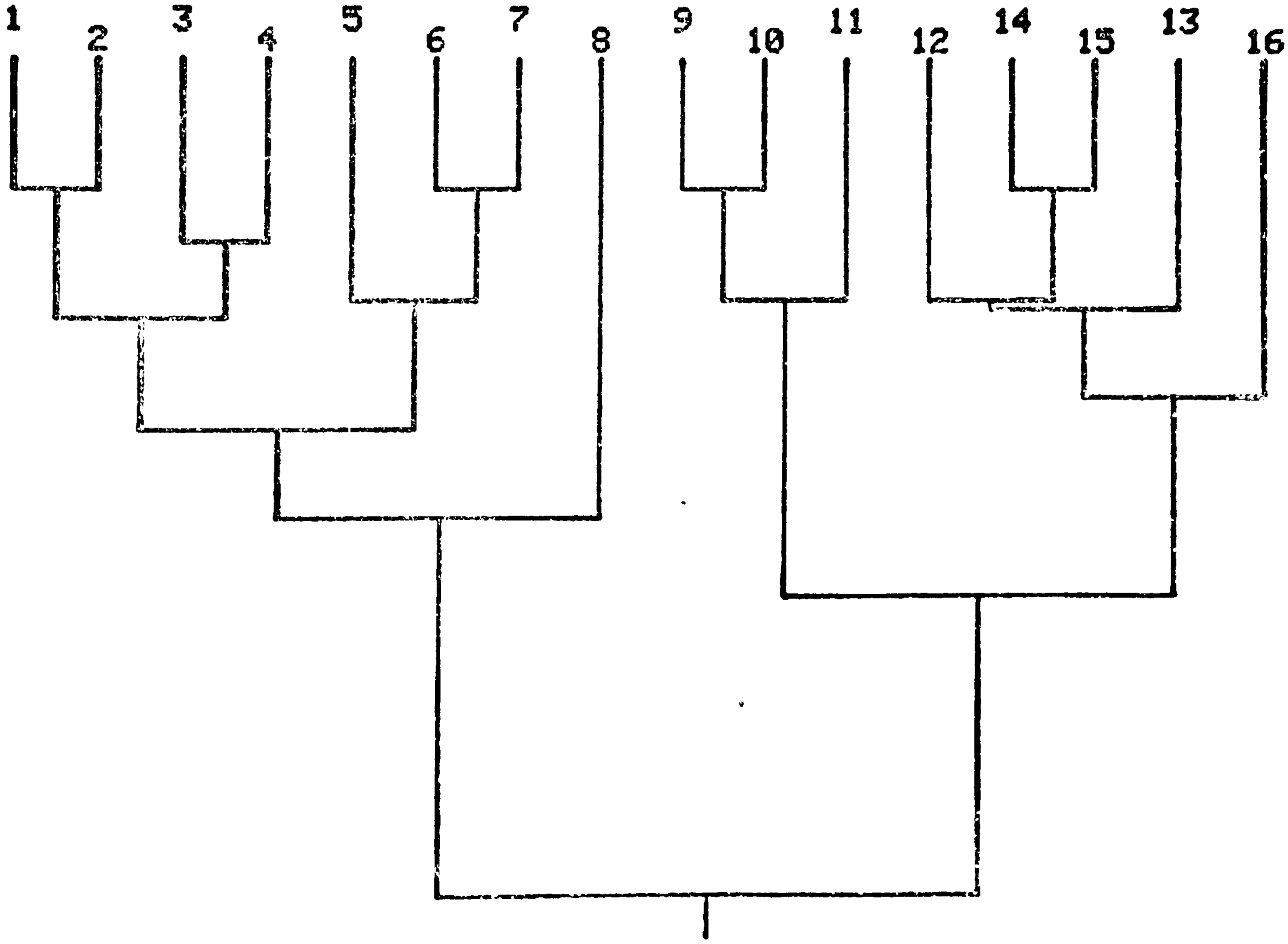


Figure 5.50
Flexible strategy (β = 0.0) - dendrogram

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.250
BETA =    0.500


PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

   1                  2          1.000
   6                  7          1.000
   9                 10          1.000
  14                 15          1.000
   1                  3          1.413
   5                  6          1.413
   9                 11          1.413
  12                 14          1.413
   1                  4          1.450
   1                  5          1.499
  12                 13          1.559
   9                 12          1.669
   1                  9          1.680
   1                 16          1.914
   1                  8          3.210

FIT IS  26.% ACCURATE
```

Figure 5.51

Flexible strategy (β = 0.5) - pairing sequence

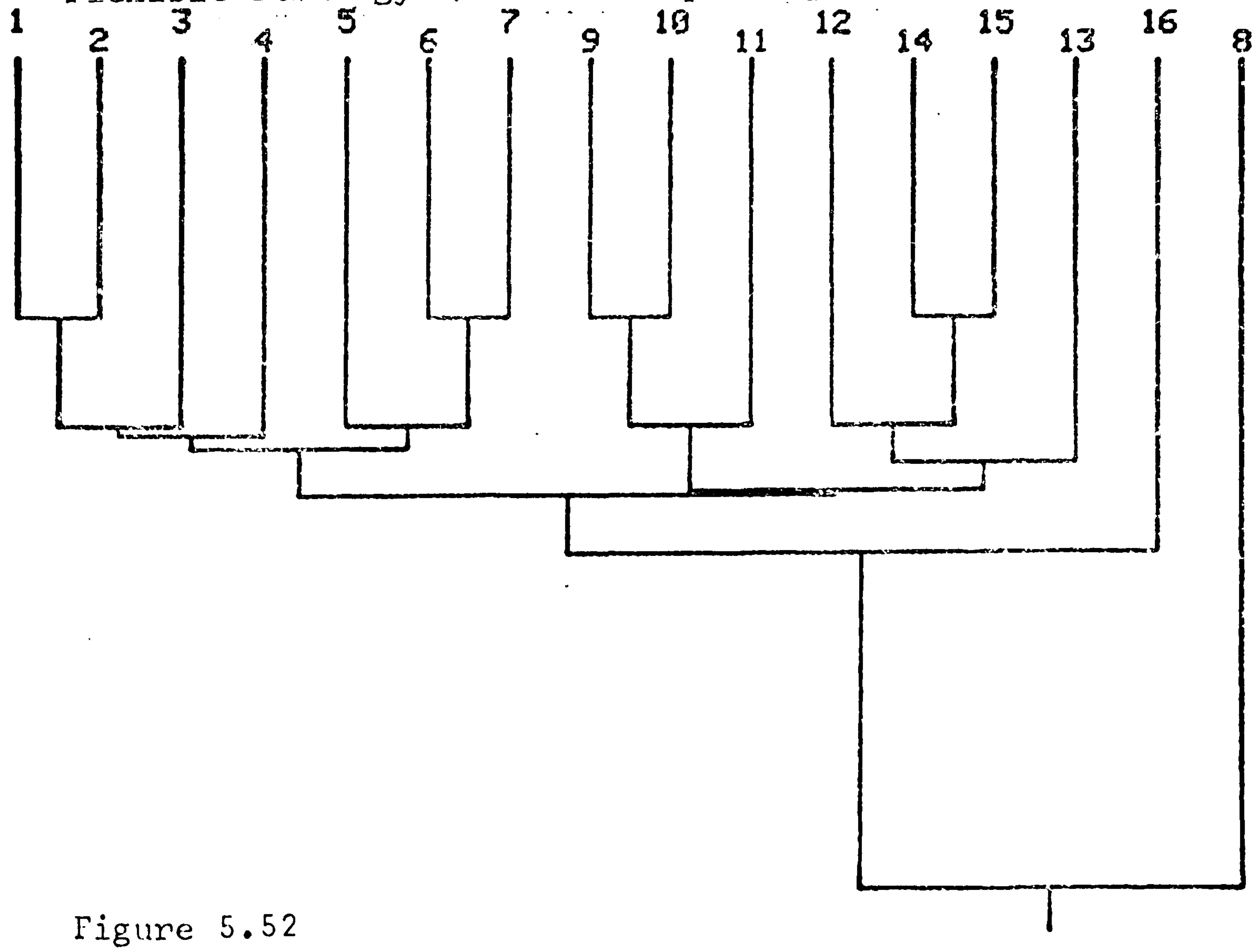


Figure 5.52

Flexible strategy (β = 0.5) - dendrogram

```
FLEXIBLE CLUSTERING STRATEGY

FLEXIBLE STRATEGY COEFFICIENTS:
ALPHA(J) = ALPHA(K) =    0.010
BETA =    0.980


PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

  1                  2           1.000
  6                  7           1.000
  9                 10           1.000
 14                 15           1.000
  1                  6           1.001
  1                 14           1.002
  1                  9           1.002
  1                 11           1.002
  1                 13           1.041
  1                 12           1.055
  1                  5           1.069
  1                  4           1.074
  1                  3           1.078
  1                  8           1.184
  1                 16           1.375

FIT IS  25.% ACCURATE
```

Figure 5.53
Flexible strategy (β = 0.98) - pairing sequence

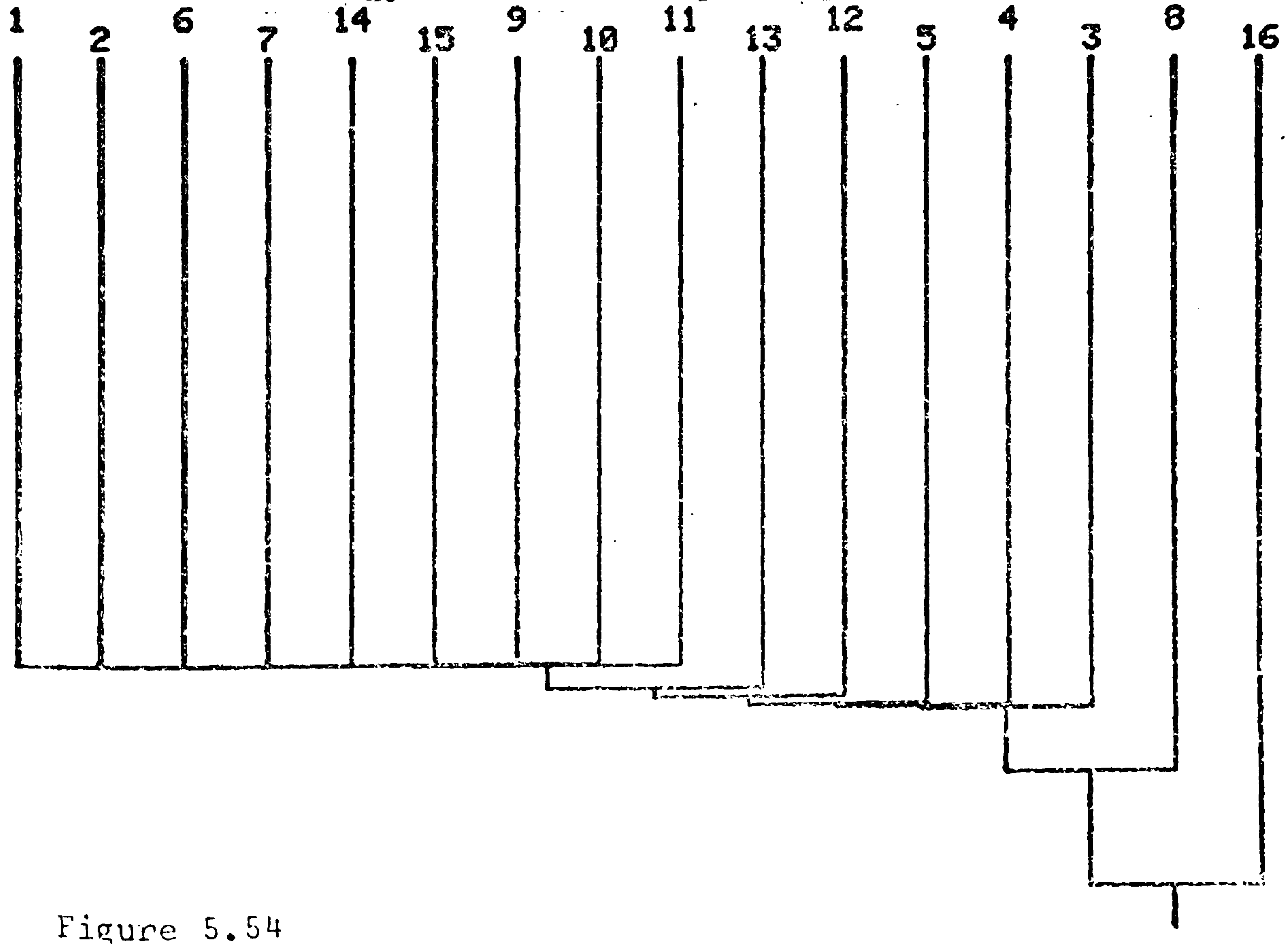


Figure 5.54

Flexible strategy (β = 0.98) - dendrogram

5.10 COMPARISON OF HIERARCHICAL TECHNIQUES

The different hierarchical techniques may be compared in a number of ways. A simple geometric model serves to illustrate the different criteria used in forming clusters in the Nearest Neighbour, Furthest Neighbour, and averaging methods (Group Average, Centroid, Median, Simple Average). Figure 5.55 shows a cluster of four elements collectively labelled J, and a cluster labelled K containing a single element, all about to be joined by another single element cluster labelled L. J and K are assumed to have joined at the last clustering step and it is now desired to compute the dissimilarity of L with the newly formed cluster (J,K), which, for convenience may be termed M. The dissimilarities obtained by the various clustering methods, expressed as Euclidean distance are laid out along the abscissa.

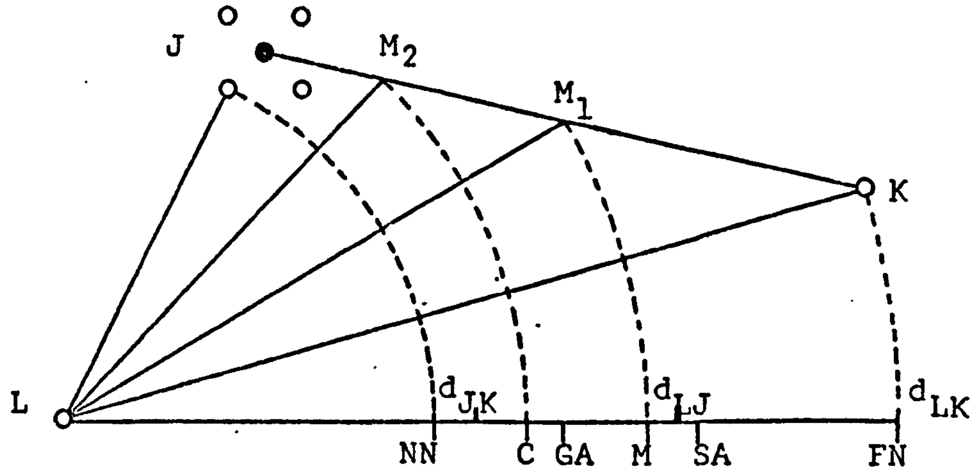


Figure 5.55
The effects of different clustering strategies on the criterion of admitting L (consisting of a single element) to a cluster formed of the four elements in J plus one in K.

It may be seen that Nearest Neighbour shows the least value of d_{LM}, since it is the distance between L and the closest member of M (in fact the nearest element of J). By contrast d_{LM} for the Furthest Neighbour equals the greatest dissimilarity between L and any member of M, namely d_{LK}. The two centroid based methods (Centroid and Median) measure the distance between L and the centroid of clusters J and K. The weighted method (Median) is the median of line JK, shown M_1. In the unweighted (Centroid) method the four elements of cluster J count 4/5 while the single element of K only counts 1/5. The centroid for the five unweighted elements, M_2, therefore lies closer to J (i.e. 0.2 of the distance from J to K). In terms of clustering criteria therefore, the distance LM_2 is less than LM_1.

No similar geometric representation of the Group Average and Simple Average is possible, but the dissimilarities obtained by these strategies are marked off along the abscissa. These distances represent the weighted or unweighted average of the lengths of the vectors from L to each of the five elements of M. It may be seen that in each case the dissimilarities are slightly greater than in the corresponding centroid method.

This becomes obvious by the examination of the respective coefficients in the combinatorial equation. Both centroid methods have $\beta < 0$ while in the averaging methods $\beta = 0$. Since the α coefficients are the same the centroid methods necessarily result in smaller distances.

5.10.1 Comparison Of Cluster Results Obtained

Objective comparison of the results of different strategies necessitates some measure of the "goodness of fit" of the hierarchical structure generated by the procedure to the original data. One approach involves the comparison of the coefficients of similarity d^*_{ij} derived from the hierarchical structure and the similarity coefficients d_{ij} measured on the original data. Clearly, if D^* ($=d^*_{ij}$) and D ($=d_{ij}$) closely resemble one another then the structure of the data is closely modelled by the hierarchical representation. The elements of D^* can be derived from the linkage order of the dendrogram.

The most widely used measure of resemblance for comparing the matrices D and D^* is the product-moment correlation coefficient, known in this context as the

Cophenetic Correlation Coefficient (Sokal and Rohlf 1962). It is calculated in exactly the same way as a normal correlation coefficient with the elements of the strict lower triangles of D and D^* considered as forming linear arrays when read row-wise.

This measure is used in MAGIC and the results on the example data set are summarised in figure 5.56 It may be seen that in this case the space-conserving strategies perform the best, followed by the space-dilating and then the space-contracting, performing worst.

Strategy	% Fit	No. of levels
Nearest Neighbour	50	3
Furthest Neighbour	55	7
Group Average	60	12
Centroid	62	12
Median	63	9
Inc. Sum of Squares	59	12
Simple Average	63	10
Flexible β=-0.25	58	10
-0.98	51	11
-0.5	53	10
0	61	10
0.5	26	9
0.98	17	9

Figure 5.56

Fit of different clustering strategies

The properties of the cophenetic correlation coefficient have been investigated by a number of authors, for example Sneath (1966), Farris (1969), Rohlf (1970), and Sokal and Rohlf (1970). Holgersson (1978) presents a probabilistic study of the statistic which suggests that it may be misleading as an indicator of presence of clusters. Highly seperated clusters are well identified but the measure shows some degree of variability for low seperation clusters. Holgersson used Monte Carlo studies of the characteristics of the coefficient for all the combinatorial strategies. An alternative approach was adopted by Gower and Banfield (1975) who used the Nearest Neighbour method to analyse data drawn from a single multivariate normal distribution and examined the behaviour of various measures of distortion.

A wide range of such measures have been proposed, and selected measures of distortion are summarised in figure 5.57. Sokal and Rohlf's (1962) measure is the correlation between the sets (s_{ij}) and ($\hat{s}_{ij}$), and hence gives an indication of the linear relationship between these two variables. It has also been used to compare the similarities defined by two different dendrograms, ($\hat{s}_{ij1}$) and ($\hat{s}_{ij2}$).

D1	$\frac{\Sigma(s_{ij}-\bar{s}_{ij})(\hat{s}_{ij}-\bar{\hat{s}}_{ij})}{[\Sigma(s_{ij}-\bar{s}_{ij})^2\,\Sigma(\hat{s}_{ij}-\bar{\hat{s}}_{ij})^2]^{1/2}}$	Sokal and Rohlf (1962) Kruskal and Carroll (1969)
D2	$\frac{\Sigma d_{ij}\hat{d}_{ij}}{[\Sigma d_{ij}^2\,\Sigma\hat{d}_{ij}^2]^{1/2}}$	Guttman (1968)
D3	$\Sigma(d_{ij}^2-\hat{d}_{ij}^2)$	Gower (1966,1970)
D4	$\begin{cases}(\frac{1}{2}\Sigma\lvert s_{ij}-\hat{s}_{ij}\rvert^{1/\mu})^{\mu}, & 0<\mu\leqslant 1\\ \max_{i,j}\lvert s_{ij}-\hat{s}_{ij}\rvert\end{cases}$	Jardine et al (1967)
D5	$\Sigma w_{ij}(s_{ij}-\hat{s}_{ij})^2$	Hartigan (1967)
D6	$\Sigma w_{ij}(d_{ij}-\hat{d}_{ij})^2$	Anderson (1971)
D7	As D6 with $w_{ij}=k$	Shepard(1962) Thompson and Woodbury (1970)
D8	As D6 with $w_{ij}=1/d_{ij}\Sigma d_{ij}$	Sammon (1969)
D9	$\Sigma[\hat{d}_{ij}-f(d_{ij})]^2/\Sigma d_{ij}$ where $f(d_{ij})$ is some "regression" function	Kruskal (1964) Kruskal and Carroll (1969)
D10	$N^{-1}\Sigma[d_{ij}/\hat{d}_{ij}]^{2a}\frac{[N^{-1}\Sigma(\hat{d}_{ij})^{2(a-b)}]^{b/(a-b)}}{[N^{-1}\Sigma(d_{ij})^{2(a-b)}]^{a/(a-b)}}$ with a=0.5, b=1 or a=b=0.5	Kruskal and Carroll (1969)

Figure 5.57

Some measures of distortion

Hartigan (1977,1978) discusses a range of different statistics, indicating the difficulty of obtaining distributional results in many cases. Other approaches have been described by Beale (1969), Calinski and

Harabasz (1974) and Mojena (1977).

It seems unlikely that any criteria will find widespread acceptance in a strict hypothesis-testing sense, because of the difficulty of anticipating the behaviour of relevant statistics under the diversity of different structures which may be present in the data. However, if used with discretion, such tests are of use in the investigation of a data set.

5.11 PROPERTIES OF CLUSTERING PROCEDURES

Some of the measures just described were concerned with providing some protection against finding groups in the data when, in fact, none were present. If this can be regarded as analogous to seeking to control the error of the first kind in hypothesis-testing, it is also relevant to investigate something corresponding to an error of the second kind: if a particular type of structure is present in the data, it should be detected.

One approach to this problem has been by simulation studies: investigators have examined the manner in which various clustering criteria have analysed data sets whose structure was known.

Thus, for example, Cunningham and Ogilvie (1972) investigated the performance of seven of the clustering algorithms in recovering the structure in six artificial data sets, each of which contained twenty objects. The data were: (i) random; four groups of five objects, the configuration within each group being identical, and the groups being either (ii) well-seperated, or (iii) close together; data whose dissimilarities were specified by dendrograms which (iv) depicted distict groups, or (v) indicated chaining; (vi) dissimilarities were obtained by distorting slightly the dissimilarities obtained in the fifth data set. Cunningham and Ogilvie (1972) reported that the Group Average method was usually at least as efficient as the other methods in recovering the underlying structure.

Kuiper and Fisher (1975) examined the performance of six of the same algorithms in analysing bivariate and multivariate normal samples, concluding that the Nearest Neighbour performed poorly and the Incremental Sum of Squares methods performed well if there were an equal number of objects from each population, but less well for unequal samples. Other studies were reported by Sneath (1966) and Baker and Hubert (1975).

Simulation studies of this kind can be of assistance in indicating the properties of different clustering criteria, and possibly in identifying unreliable criteria. However, it is unlikely that they will be of more than limited usefulness. No single clustering criterion can be guaranteed to detect correctly all types of structure in data, and even if the most appropriate procedure were known for every conceivable type of structure, the problem remains that in general the precise form of data is not known prior to the analysis: it is precisely in order to establish this that the investigation is undertaken.

5.11.1 A Theoretical Comparison Against Required Criteria

The following discussion assumes a degree of information about the data, or the required properties of the classification. In the context of taxonomy Jardine and Sibson (1971) regarded classification as the mapping from a data set A to a target set Z:

$$D : A \rightarrow Z$$

The mapping from a set of dissimilarities (d_{ij}) into a set of transformed dissimilarities ($\hat{d}_{ij}$) corresponding to a dendrogram can be represented by the transformation

$$D : d \rightarrow \hat{d}$$

where ($\hat{d}_{ij}$) satisfy the ultrametric inequality, i.e.

$$\hat{d}_{ik} \leqslant \max(\hat{d}_{ij}, \hat{d}_{jk}) \text{ for all objects } i,j,k$$

Jardine and Sibson (1971) presented a more general axiomatic formulation specifying various properties which one might require of the function D and the data set and target set. For example:

(i) The method must not depend on any prior labelling of the objects.

(ii) The method must not depend on any scale factor α;

$$D(\alpha d) = \alpha D(d)$$

(iii) Preservation of clusters: if $d \in A$, then there exists $d' \in Z$ such that $d' \leqslant d[D(d) \leqslant d]$. The rationale being that a maximal linked set at level h in d may have other objects added to it at the same level in D(d), but it must not be broken up.

(iv) In order to be able to investigate the effect of small changes in the input dissimilarity matrix, the mapping should be continuous: small changes in the dissimilarities should not give rise to large changes in the classification.

A fuller list of conditions, with discussion, was given by Jardine and Sibson (1971, Chapter 9). These authors showed that if the mapping is specified in the above way, from a set of dissimilarity coefficients to a set of ultrametric dissimilarities, then the Nearest Neighbour method is the only classification method that satisfies their list of axioms. Other authors have investigated Jardine and Sibsons axioms, and few have regarded them as sufficiently important as to so restrict the acceptable classification methods. In particular, it has been queried whether continuity (axiom iv above) should be required as a global property of a clustering method (Cormack 1971). Hierarchical clustering methods other than Nearest Neighbour are subject to discontinuities when analysing certain data sets, but this should not necessarily rule out thir use.

Axiomatic characterisations of other clustering methods would be a valuable development, but these would appear to be a very difficult to obtain. Wright (1973) gave a list of properties which are satisfied by a sum of squares criterion, but did not prove that these properties give a unique characterisation of the method.

In the absence of complete axiomatisations of each clustering method, such listings of the properties of each method can provide useful information. Relevant related work is the admissability approach of Fisher and Van Ness (1971) and Van Ness (1973). Drawing the concept of admissability from decision theory, these authors gave various properties which one might expect "reasonable" clustering procedures, or the groups obtained from applying these procedures, to possess. If A denotes some property to be satisfied, then any procedure which satisfies A is called A-admissable. Some of the properties introduced by Fisher and Van Ness (1971) are listed below:

(1) Convex admissibility: a partition into clusters C_1, C_2, . . , C_g is said to be convex admissible if the convex hulls of C_1, C_2, ..,C_g do not intersect (this condition requires that the original form of the data be such that each object can be represented by a point in some Euclidean space).

(2) Point proportion admissibility: a procedure is said to be point proportion admissible if duplicating one or more objects any number of times and reapplying the procedure to the modified data set does not alter the boundaries of the clusters obtained.

(3) Cluster omission admissibility: suppose that a clustering procedure produces a partition into g clusters, $C_1, C_2, \ldots, C_g$, and all objects in any one of these clusters, say C_j, are removed from the data set, then the reduced data set is re-analysed to obtain the optimal (g-1) clusters using the same procedure. If the (g-1) clusters obtained are always (C_i ($i=1,\ldots,g; i\neq j$)), the procedure is said to be cluster omission admissible.

(4) Monotone admissibility: a procedure is monotone admissible if applying a monotone transformation to each element of the dissimilarity (or similarity) matrix does not change the resulting clustering.

(5) Well-structured (g-group) admissibility: data are defined to be well-structured (g-group) if there exists a partition into g groups for which all within-group dissimilarities are smaller than all between-group dissimilarities. A clustering procedure is well-structured g-group admissible if it produces the correct partition into g groups whenever it is applied to data which are well-structured (g-group).

The rationale of Fisher and Van Ness's admissibility approach is that it is not usually possible to specify a single "best" clustering procedure, but using their data one may select a procedure with known characteristics.

Clustering	Admissibility condition				
Procedure	1	2	3	4	5
N.N.	No	Yes	Yes	Yes	Yes
F.N.	No	Yes	Yes	Yes	Yes
G.A.	No	No	Yes	No	Yes
Centroid	No	No	Yes	No	No
I.S.S.	Yes	No	Yes	No	No

Figure 5.58

Admissibility table of some clustering strategies (see text for details of conditions)

Figure 5.58, adapted from Fisher and Van Ness (1971) and Van Ness (1973), summarises admissibility properties of five clustering strategies. Such tables may be used as follows: if one wanted to use a clustering procedure which was point proportion admissible and monotone admissible, one should not use either the Incremental Sum of Squares or the Group Average procedures. If restricting attention to the criteria described in figure 5.58 one would analyse the data either by the Nearest Neighbour or Furthest Neighbour method, or preferably both.

The admissiblity approach assumes one has some information on the form of data to allow one to reduce the number of clustering criteria which have to be considered. This information often need only be of a very vague nature, but sometimes even such limited information is not available and other approaches to classification are required.

5.12 COMPARATIVE STUDIES

The previous statistical investigations have possibly over-emphasised the extent to which clustering criteria impose their own structure on data. As Cormack (1971) remarks, "if clusters are really distinct, it would be hoped that any strategy worthy of use would find them".

The converse to this argument has been suggested by a number of authors: if the results of several different classification procedures agree closely, then one has more confidence in the reality of any group structure which is indicated; it is less likely to be purely an artifact of the classification criteria used. A wide range of comparitive studies have been carried out. Various authors have been concerned:

(i) to examine the effects of using different measures of dissimilarity possibly based on different standardisations of data, or on different subsets of the variables;

(ii) to compare the results of applying different clustering and/or geometrical procedures to the same data set, or to compare the results suggested by numerical classification procedures with classifications obtained by traditional, non-numerical methods.

Typical of these studies are Sokal and Michener (1967), Moss (1968), and Boyce (1969).

Such comparitive studies can provide useful information about the properties of different clustering methods and measures of dissimilarity in much the same way as the simulation studies described previously. However, when the aim is to obtain an assessment of the structure in the data revealed by different clustering methods, most measures of resemblance between partitions, or between dendrograms, do not explicitly specify where the similarities of the classifications lie; these similarities have tended to be assessed by eye. More formal methods of comparison have been proposed by Adams (1972).

Probably the most fruitful comparative studies to date have been those which have combined clustering with geometrical methods of analysis. This is undoubtedly because the relative strengths and weaknesses of the two approaches are largely complimentary. Thus, geometrical methods do not force a group structure on the data, allowing the observer to assess whether the points fall naturally into distinct clusters. On the other hand, the assessment by eye of two- or three-dimensional representations can be subjective, and it is profitable to examine whether partitioning the data using some clustering criterion indicates the same groups as appear to be present in the geometrical representation.

Many clustering and geometrical classification methods appear to be complimentary in another way, in that studies indicate that clustering methods tend to be more reliable in depicting lower level differences between objects, whereas geometrical representations generally portray the group relationships more reliably. As will be illustrated later a combination of the two approaches can prove helpful in uncovering the structure in multivariate data.

Classification can be a means of reducing large amounts of data to manageable summary form. As the volume and compexity of data increase, the human brain becomes less able to hold in balance all the different factors which are relevant to the assessment of the data. MAGIC performs this balancing act, and hence helps the designer to gain insights into the structure of his data.

CHAPTER 6
EUCLIDEAN CLUSTER ANALYSIS

6.1 NONHIERARCHICAL CLUSTER ANALYSIS

For a data set of m entities the hierarchical methods of chapter 5 give m nested classifications ranging from m clusters of one member each to one cluster of m members. This chapter describes clustering techniques which produce a single classification of k clusters, where k is either specified a priori or is determined as part of the clustering method.

The concept in the majority of these methods is to choose some initial partition of the activity data and then alter cluster membership so as to obtain a better partition. The various algorithms which have been proposed in the literature differ as to what constitutes a "better" partition and what methods may be used for effecting the improvements, but the broad concept for all methods is very similar to that underlying the steepest descent algorithms used for unconstrained optimisation in nonlinear programming. Such algorithms begin with an initial point and then generate a sequence of moves from one point to another, each giving an improved value of the objective function, until a local optimum is found. In terms of the exploratory data analysis approach adopted by MAGIC the techniques may also be compared to the plotting of scatter diagrams. The plotting of these diagrams is a traditional approach to finding patterns in data, but it is essentially a two-dimensional technique (which may be extended to three dimensions with some difficulty). As used in MAGIC nonhierarchical cluster analysis may be regarded as an exploratory technique for doing in n dimensions some of the things that scatter diagrams do so well in two dimensions.

Compared with the hierarchical techniques, nonhierarchical clustering methods optimise intra-group homogeneity, as distinct from optimising a hierarchical route from individual elements to population. The methods of nonhierarchical cluster analysis possess the theoretical advantage that they admit the relocation of elements, which thus allows a poor initial partition to be corrected at a later stage. All hierarchical strategies suffer from what has become known as the "migration problem": links which were correctly made in the early stages of the process may later prove unprofitable in so far as they eventually lead to the possible misclassification of elements further down the tree.

A nonhierarchical clustering system will, in principle, consist of four distinct processes, as follows:

(i) a method of initiating clusters;

(ii) a method of allocating new elements to existing clusters, and/or of fusing existing clusters;

(iii) a method of determining when further allocation may be regarded as unprofitable, so that certain elements remain unallocated as single-element clusters;

(iv) a method of reallocating some or all of the elements to existing clusters when the main classificatory process is completed, thus redressing any misclassification produced by the "migration" process referred to above.

All systems necessarily involve (i) and (ii); but in any particular system either (iii) or (iv), or both, may be lacking. The differences between methods lie primarily in the method of initiation employed, and the criteria used for reallocation. The following general discussion of specific strategies is thus organised from the standpoint of (i) and (iv) above.

6.2 INITIAL CONFIGURATIONS

All of the methods discussed here begin with an initial partition of the elements into groups, or with a set of seed points around which clusters may be formed. The majority of techniques begin by establishing a set of k seed points in the p-dimensional space, which act as initial estimates of cluster centres around which the set of m elements can be grouped. The problem of deciding an appropriate value of k for any set of data is discussed in section 6.5. The following methods are representative examples of how such seed points may be generated.

(i) the simplest procedure is to choose the first k elements in the data set (MacQueen 1967)

(ii) a variation is to subjectively choose any k elements from the data set.

(iii) a further variation is to label the data elements from 1 to m and choose those labelled m/k, 2m/k, ..., (k-1)m/k, and m.

(iv) label the data elements from 1 to m and choose those corresponding to k different random numbers in the range 1 to m (McRae 1971).

(v) take any partition of the data elements into k mutually exclusive groups and compute the group centroids as seed points (Forgey 1965). Methods of generating such partitions are discussed in section 6.3.

(vi) Beale (1969) sets up cluster centres regularly spaced at intervals of one standard deviation on each variable. MAGIC adopts a variation of this method by randomly choosing cluster centres within the range between the maximum and minimum observed values on each variable. Full details of the complete clustering algorithm used in MAGIC are given in section 6.6.

This list of methods is not exhaustive, but does provide the basis to enable a number of observations to be made. The methods in which every seed point is itself a data unit ensure that each cluster will have at least one member - other techniques need to include checks for "empty" clusters. Randomness is another important topic: all the methods described have elements of randomness, either through an implicit assumption of random ordering of data elements within the data set, or through explicit random selection. In terms of exploratory data analysis it is not randomness per se that is of interest but indifference; that is, the goal is an initial configuration free of overt bias. Ultimately indifference is probably best effected through random selection, but the selection of the set of possibilities from which the random selections are made can also affect the problem. In MAGIC the method adopted makes a deliberate attempt to span the data set with seed points as such methods are less prone to distorted or badly balanced configurations than methods involving totally random selection. The adopted method is firther refined by the application of a "pseudo F-test", which is discussed in detail in sections 6.5 and 6.6.

6.3 INITIAL PARTITIONS

In some clustering methods the emphasis is laid on generating an initial partition of the data elements into k mutually exclusive clusters rather than finding a set of seed points, although in many cases a set of seed points is used in that process. Some methods of generating such partitions are considered here together with ways of allocating elements to clusters given an initial set of seed points.

(i) For a given set of seed points, assign each element to the cluster built around the nearest seed point (Forgey 1965). The seed points remain stationary throughout the assignment of the full data set; consequently the resulting set of clusters is independent of the sequence in which the elements are assigned.

(ii) Given a set of seed points, let each seed point initially be a cluster of one member; then assign elements one at a time to the cluster with the nearest centroid; after an element is assigned to a cluster update the centroid so that it is the true mean vector for all the data elements currently in that cluster (MacQueen 1967).

(iii) A hierarchical clustering method may be used to obtain an initial partition. Wolfe (1970) uses the Incremental Sum of Squares method and Lance and Williams (1967) suggest using hierarchical methods on a subset of the data to obtain the nuclei for assignment of the remaining clusters.

(iv) Random partitions may be devised - for example, assign a data element to one of k clusters by generating a random number between one and k. Random allocation to groups is not a particularly useful method as the resulting groups have no properties of internal homogeneity and, indeed, are not clusters at all.

(v) A further option may be to allow the program user to define an initial partition. Friedman and Rubin (1967) provide such an option, but although this may be of interest to the specialist user it could equally confuse the naive or occasional user.

The distance measure used in all cases is the Euclidean metric.

6.4 RELOCATION TECHNIQUES

Once an initial classification has been found a search is made for elements which may be reallocated to another cluster, in an attempt to optimise some clustering criterion. In general relocation proceeds by considering each element in turn for reassignment to another cluster, reassignment taking place if it causes an increase (or decrease in the case of minimisation) in the criterion value. The procedure continues until no further move of a single element causes any improvement. It is possible to reach local optima, and, in general there is no way of knowing if absolute maxima or minima have been achieved. A number of clustering criteria have been developed, based around the matrix equation

$$T = B + W$$

The scatter of two variables is the inner product of their centred score vectors. The total scatter matrix T is a square array in which the typical entry t_{ij} is the scatter of variables i and j computed over all elements in the data set. In a partition of the data set into k clusters, the within group scatter matrix for cluster k, W_k, has the typical entry w_{ijk}, which is the scatter of variables i and j computed over all data elements in cluster k; the within group scatter matrix for the partition is

$$W = \sum_{k=i}^{k=h} W_k$$

The between group scatter matrix B has as its typical element $b_{ij} = \Sigma m_k \bar{x}_{ik} \bar{x}_{jk}$ where $\bar{x}_{ik}$ is the mean (centred about the grand mean in the data set) of the ith variable in the kth cluster, and m_k is the number of data elements in the kth cluster. It can be shown that the three matrices satisfy the relation T = B + W, and a particularly important element in the definition of the various clustering criteria is the determinantal equation $|B - \lambda W| = 0$; the λ_i solutions to this equation being the eigenvectors of the matrix $W^{-1}B$.

Various authors (Friedman and Rubin 1967, MacRae 1971, Scott and Symons 1971, Marriott 1971) have proposed criteria for evaluating whether movements of individual data elements result in an overall improvement of a partition. Four principal criteria have emerged from these studies:

(i) Minimise trace W. The trace of a matrix is the sum of its diagonal elements. It may be shown that this criterion is the same as minimising the total within group sum of squares of the partition, since the minimisation of tr(W) is equivalent to the maximisation of tr(B), as the fundamental matrix equation leads to

$$\text{trace }(T) = \text{trace }(B) + \text{trace }(W)$$

(ii) Minimise the ratio of the determinants $|W|/|T|$. This criterion is widely known as Wilks' lambda statistic (Wilks 1938). Since the matrix T is the same for all partitions, this criterion is equivalent to minimisig $|W|$. Another equivalent criterion is to maximise $|T|/|W|$ which may be shown to be equivalent to maximising $|I+W^{-1}B|$ or maximising $\prod_{i=1}^{n}(1+\lambda_i)$.

(iii) Maximise the largest eigenvalue of $W^{-1}B$. This criterion is known as the largest root criterion.

(iv) Maximise the trace of $W^{-1}B$. This criterion is known as Hotelling's trace criterion and is equivalent to maximising $\sum_{i=1}^{n}\lambda_i$.

The technique adopted in MAGIC follows the first method as there are at least two serious problems associated with criteria (ii), (iii) and (iv). The first problem is that they involve the computation of eigenvalues at each stage which overshadows the rest of the method in terms of computational effort, and secondly, there are no clear statistical advantages in their use anyway.

6.5 STOPPING RULES

The problem of deciding the number of clusters present in the data has already been mentioned. In hierarchical techniques no clear indicator exists, although the examination of various dendrograms may provide an accurate enough empirical technique in our application. With the nonhierarchical methods several attempts to devise reasonable indicators have been made. For example, Beale (1969) gives an F- statistic

$$F(c_2,c_1) = \frac{R(c_1)-R(c_2)}{R(c_2)} \Bigg/ \left\{ \left(\frac{N-c_1}{N-c_2}\right)\left(\frac{c_2}{c_1}\right)^{2/n} - 1 \right\}$$

based on $p(c_2-c_1)$ and $p(N-c_2)$ degrees of freedom. In this formula $R_c = (N-c)S_c^2$ where S_c^2 is the mean square deviation from cluster centres in the sample. A significant result indicates that a subdivision into c_2 clusters is significantly better than a subdivision into some smaller number of clusters c_1. This measure is used in MAGIC and is discussed further in section 6.6.

Marriot (1971) has investigated the properties of the $|W|$ criterion, as proposed by Friedman and Rubin (1967). He proposes the use of $g^2|W|$, where g is the number of groups, as the indicator of group structure, taking as the correct number of groups the value of g for which $g^2|W|$ is a minimum.

The Wilks lambda criterion (above) forms a liklihood ratio test. To test the hypothesis of say c_2 groups against that of c_1 groups it is possible to use the statistic $-2\log\lambda$ where λ is the ratio of likelihoods, $\lambda = L_{c_2}/L_{c_1}$, which Wilks (1938) showed, under certain constraints, is asymptotically distributed as chi-square with degrees of freedom equal to the difference in the number of parameters of the two hypotheses.

6.6 THE STRATEGY ADOPTED IN MAGIC

For any given number of clusters MAGIC generates coordinates for the centre of each cluster and assigns each element to one (and only one) cluster, attempting to minimise the sum of squares of the deviations of the elements from their respective cluster centres. Statistically this is equivalent to maximum likelihood if all clusters are assumed to be spherically normally distributed with a common variance. The distance measure used is the Euclidean metric, all observations being represented as points in n-dimensional space.

Given the grouping of observations into clusters, the centres should ideally be chosen at the means of the observations in each cluster. It is, however, difficult to determine the best grouping. What the program does is find a grouping that cannot be improved by moving any single observation into another cluster, even if the cluster centres are repositioned after the re-assignment.

Thus, if an observation in cluster j is at a distance d_j from its cluster centre, and at a distance d_k from the centre of cluster k, then it is an improvement of the grouping to reassign it to cluster k if $d_k^2 < d_j^2$. But it may also be reassigned if

$$\frac{n_k d_k^2}{n_k+1} < \frac{n_j d_j^2}{n_j-1}$$

where n_j and n_k are the current numbers of observations in clusters j and k. This criterion allows an improvement in many situations where the simpler criterion would not.

Having found a solution with one number of clusters, MAGIC will look for a solution with one fewer clusters by finding the pair that can be amalgamated with the smallest increase in the sum of squares of deviations. That is to say, it minimises

$$n_j n_k d_{jk}^2 / (n_j + n_k)$$

where d_{jk} denotes the distance between the centres of clusters j and k whose amalgamation is being considered. This is used as a first trial solution for the new number of clusters; improved solutions are then found by reassigning individual points to other clusters as before.

The program may find local optimum clusterings rather than global optima, particularly as it adopts a form of random initial grouping. The amalgamation process overcomes many of the objections associated with random starting solutions and, by starting the process with three or more clusters more than are required the solutions for all relevant numbers of clusters should be good ones.

Just what the "relevant number of clusters" is is a difficult problem, already mentioned in passing. Since the clusters are essentially descriptive statistics, and not based on any specific distributional form for the observations, the question cannot be answered precisely.

It is, however, possible to get some guidance as to whether any given set of data can reasonably be interpreted as c clusters.

Suppose we have N observations in n dimensions, and let R(c) denote the residual sum of squares when the observations are divided into c clusters. One might then try an F-test to decide whether a subdivision into c_2 clusters was significantly better than a subdivision into c_1 clusters, where $c_1<c_2$, taking $R(c_1)-R(c_2)$ as having $n(c_2-c_1)$ degrees of freedom. This test would be appropriate if, for any given number of clusters, the observations had been assigned to clusters a priori. But the fact that the points can be assigned so as to minimise R(c) means that this test always suggests that the larger number of clusters is very significantly better.

Nevertheless, this approach may be modified to give intuitively sensible results, by using a large sample correction factor for the expected reduction of R(c) as c increases in the absence of any definite clustering. Returning to the concept of the observations as points in n-dimensional space we may consider the observations as a sample from a population covering a volume V in that space. The clusters will then divide this volume into c regions of approximately similar sizes such that all observations in a region form a cluster. So, if σ_c^2 denotes the mean square distance from any point to the centre of its region, the value of σ_c^2 will decrease as c increases according to

$$c\sigma_c^n = k_o$$

where k_o is some number that depends on V (and possibly n) but not on c. This implies that $\sigma_c^2 = kc^{-2/n}$, where $k = k_o^{2/n}$, i.e. another constraint independent of c. Hence,

$$E(R(c)) = k(N-c)c^{-2/n}$$

The term (N-c) is of little practical importance, being almost independent of c, but is logical as a "degrees of freedom" effect, since the cluster centres within each region are chosen as the sample means. Hence

$$E\left(\frac{R(c_1)-R(c_2)}{R(c_2)}\right) = \left(\frac{N-c_1}{N-c_2}\right)\left(\frac{c_2}{c_1}\right)^{2/n} - 1$$

enabling us to compute the statistic quoted above in section 6.5, i.e.

$$F(c_2,c_1) = \frac{R(c_1)-R(c_2)}{R(c_2)} \Bigg/ \left\{\left(\frac{N-c_1}{N-c_2}\right)\left(\frac{c_2}{c_1}\right)^{2/n} - 1\right\}$$

and treat it as as F-ratio with $n(c_2-c_1)$ and $n(N-c_2)$ degrees of freedom. The statistic is computed for all $c_1 < c_2 \leqslant c_{max}$; and if, for a given c_1, it is significant for c_2, we may say that the representation in terms of

c_1 is not entirely adequate. In practice the significance level does not usually depend much on c_2 for $c_2 > c_1+2$.

6.7 THE IMPLEMENTATION OF THE ALGORITHM

(1) Allocate the points x_{ij} to the cluster having the nearest centre. If this is the initial allocation the centres are chosen randomly between the maximum and minimum observed values on each variable. Distances are calculated by $d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 /p$.

(2) After assignment redefine the cluster centres as the centroids of the clusters by

$$y_{kj} = 1/n_k \sum_{i=1}^{n_k} x_{ij} \quad (j=1,2,..p; \quad k=1,2,..,c_{max})$$

in which n_k represents the number of elements assigned to cluster k, and y_{kj} are the coordinates of the kth cluster centre on the jth variable (axis).

(3) At this stage the elements are moved in turn to other clusters to see if the total squared distance from the points to the cluster centres is reduced, when, at the same time, the cluster centres are themselves moved to take account of the relocation of the points. That is point i is moved from cluster j to cluster k if

$$\frac{n_k \Sigma (x_{im} - y_{km})^2}{n_k + 1} < \frac{n_j \Sigma (x_{im} - y_{jm})^2}{n_j - 1}$$

If this condition is satisfied the move is made permanent and the values of y_{jk} recomputed as in step (2). When no further moves produce any improvement that configuration provides the solution for c_{max} clusters.

(4) The number of clusters, c, is reduced by one unless $c=c_{min}$, in which case this stage is omitted. The pair of clusters to be merged is found by locating that combination of two clusters which minimises the increase in the squared deviations of the observations from their cluster centres, i.e.

$$\frac{n_i n_j}{n_i + n_j} \Sigma (y_{ik} - y_{jk})^2 \quad (i=1,c-1; \quad j=i+1,c)$$

This value is calculated for all (i,j) and the minimum chosen. If the minimum is found when i = ml and j = m2 then clusters ml and m2 are amalgamated and the centroid of the resulting cluster calculated as in step (2).

(5) At this stage all clusterings have been performed for $c_{min} \leqslant c \leqslant c_{max}$ and associated residual sums of squares calculated for each c in the range, by

$$RSS(c) = S_c^2/p(n-c)$$

where S_c is the root mean square deviation of points from the cluster centre, i.e.

$$S_c = \overline{(n-c)\ \Sigma\, d_{ik}^2}^{\,1}\ )^{1/2}$$

where c is the current number of clusters and d_{ik}^2 is the squared distance of the ith point from the centre of cluster k, to which it had been assigned, i.e.

$$d_{ik}^2 = \Sigma(x_{ij} - y_{kj})^2 \quad (i=1,n_k;\ k=1,c)$$

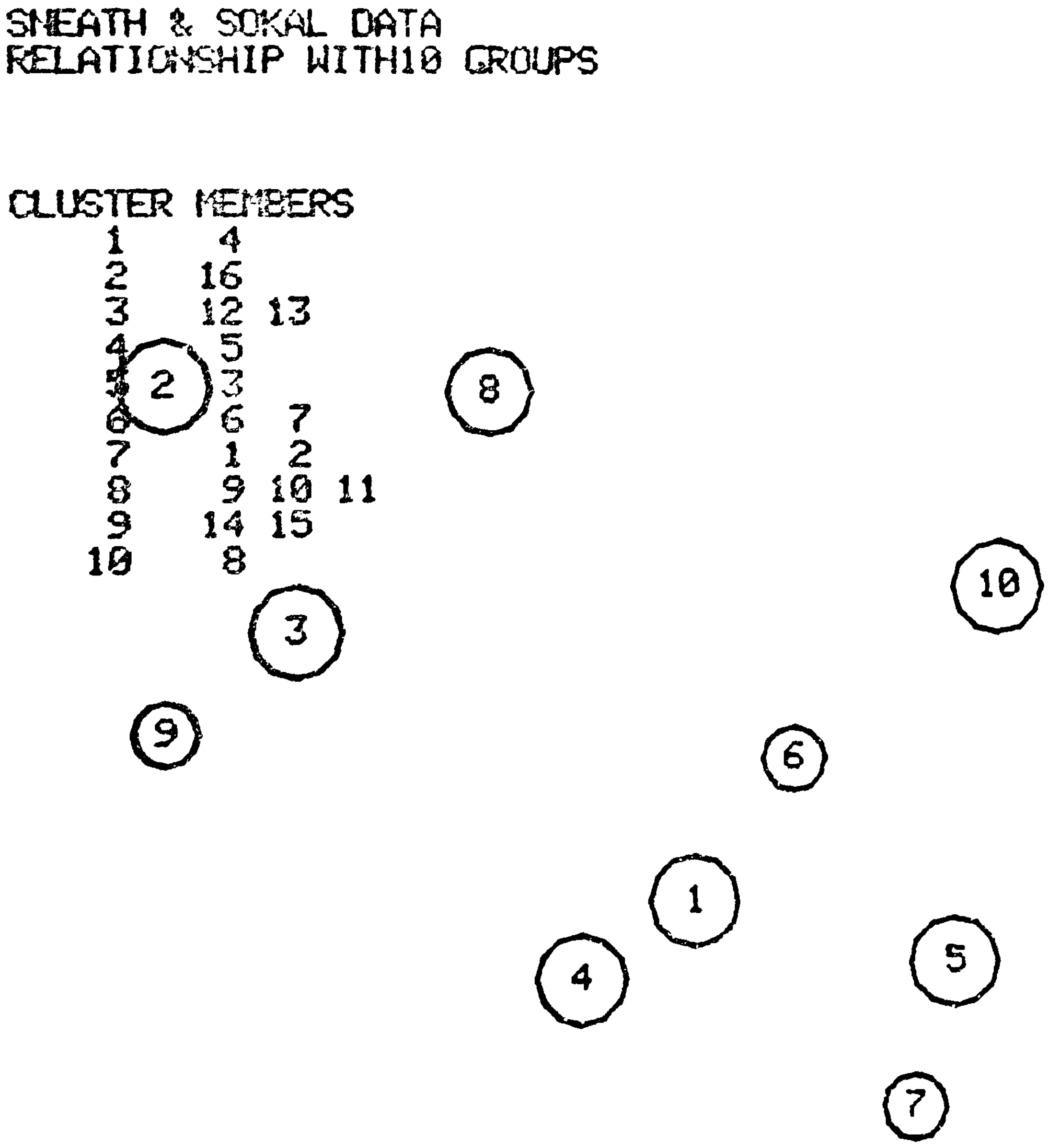
These values may then be used to compute the F-ratio test

$$F(n1,n2) = \frac{RSS(n2) - RSS(n1)}{RSS(n1)} \Big/ \frac{n-n2}{n-n1} \cdot \left(\frac{n1}{n2}\right)^{2/p} - 1$$

which has degrees of freedom of p(n1-n2) and p(n-n1).

6.8 EXAMPLE OF EUCLIDEAN CLUSTERING

The Euclidean clustering is illustrated by the output from MAGIC using the Sneath and Sokal data set, as in the hierarchical clustering. Ten clusters were requested initially and figures 6.1 to 6.9 show the resulting clustering into ten groups down to two groups.

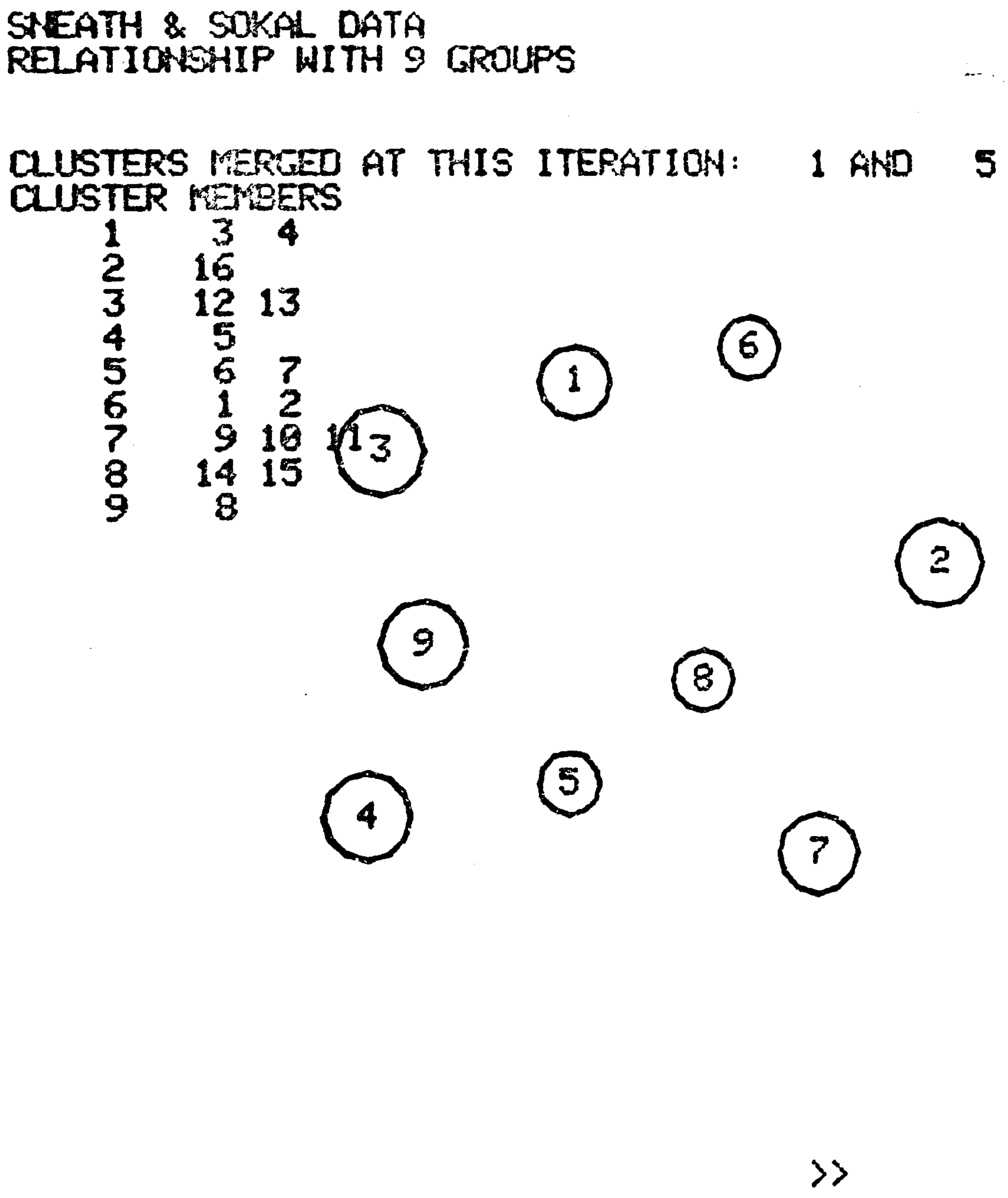


SNEATH & SOKAL DATA
RELATIONSHIP WITH10 GROUPS

CLUSTER	MEMBERS		
1	4		
2	16		
3	12	13	
4	5		
5	3		
6	6	7	
7	1	2	
8	9	10	11
9	14	15	
10	8		

Figure 6.1

Euclidean clustering - 10 groups

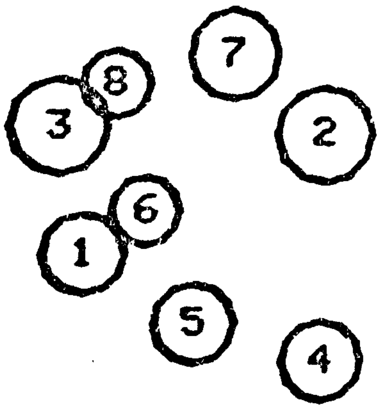


```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 9 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   1 AND   5
CLUSTER MEMBERS
     1    3  4
     2   16
     3   12 13
     4    5
     5    6  7
     6    1  2
     7    9 10 11
     8   14 15
     9    8
```

Figure 6.2

Euclidean clustering - 9 groups

```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 8 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   5 AND   9
CLUSTER MEMBERS
     1    3  4
     2   16
     3   12 13
     4    5  6
     5    7  8
     6    1  2
     7    9 10 11
     8   14 15
```

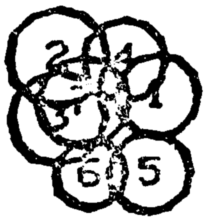


>>

Figure 6.3

Euclidean clustering - 8 groups

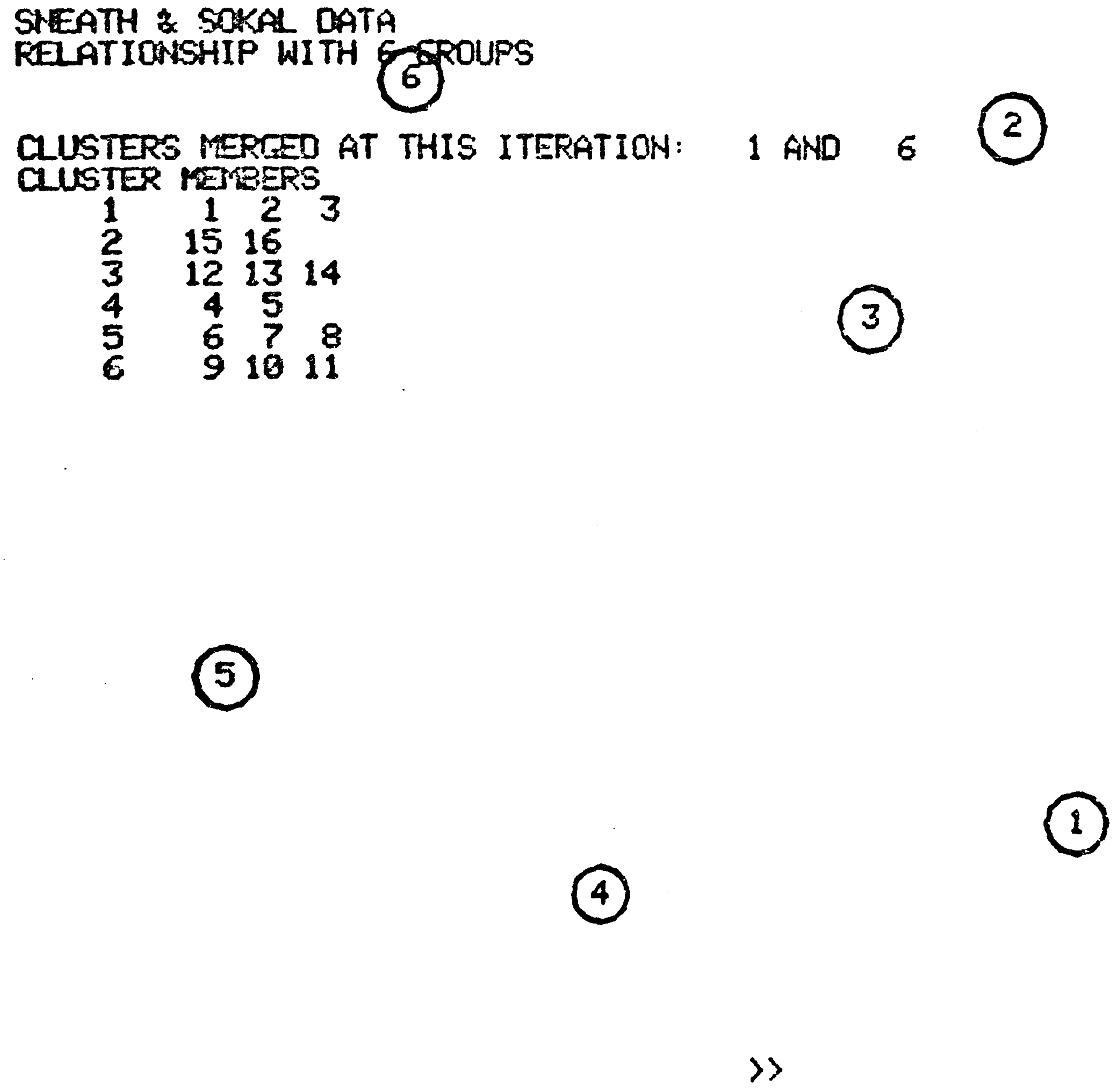
```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 7 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   3 AND   8
CLUSTER MEMBERS
    1     3  4
    2    15 16
    3    12 13 14
    4     5  6
    5     7  8
    6     1  2
    7     9 10 11
```



>>

Figure 6.4

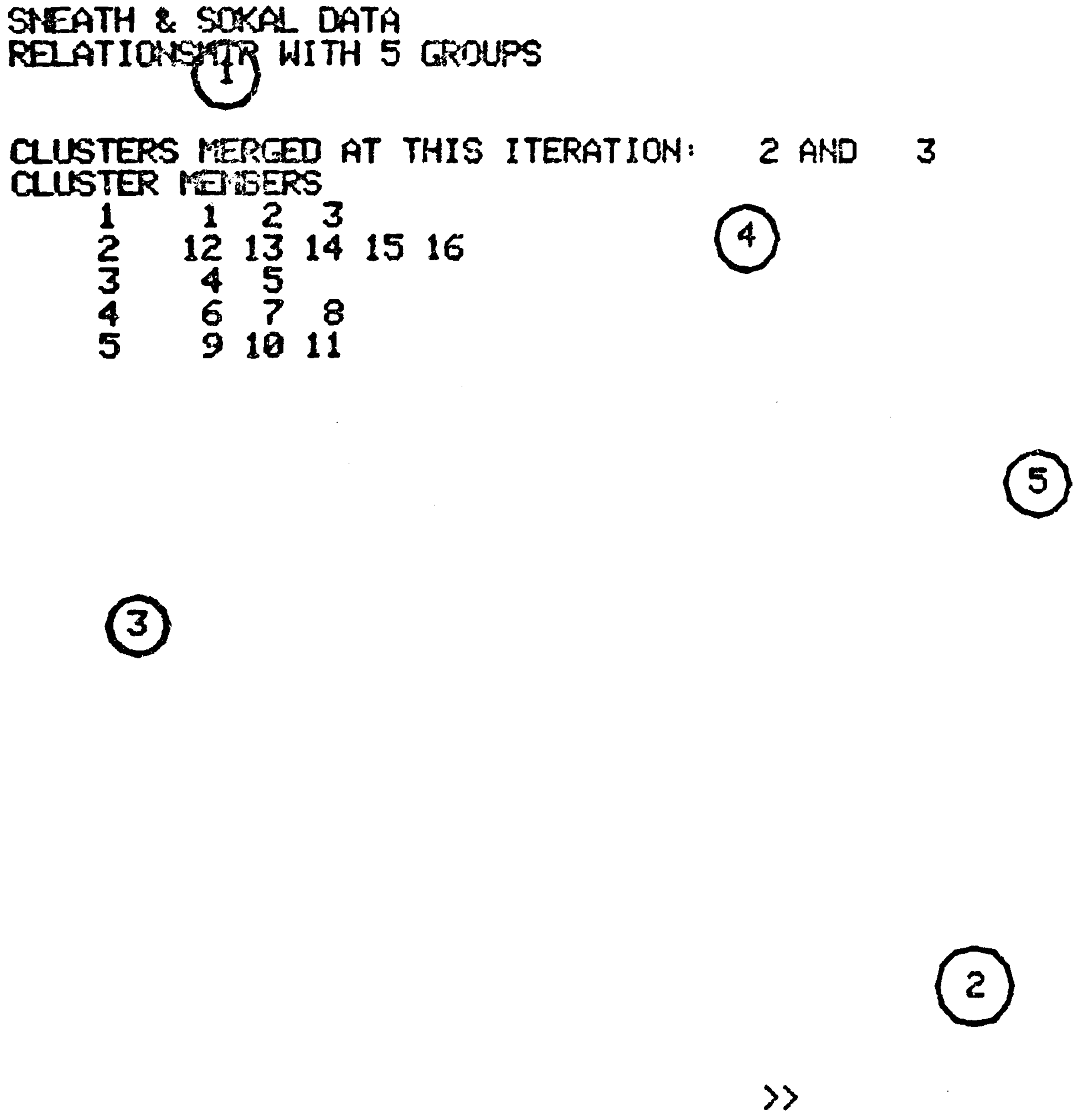
Euclidean clustering - 7 groups



```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 6 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   1 AND   6
CLUSTER MEMBERS
    1    1  2  3
    2   15 16
    3   12 13 14
    4    4  5
    5    6  7  8
    6    9 10 11
```

Figure 6.5

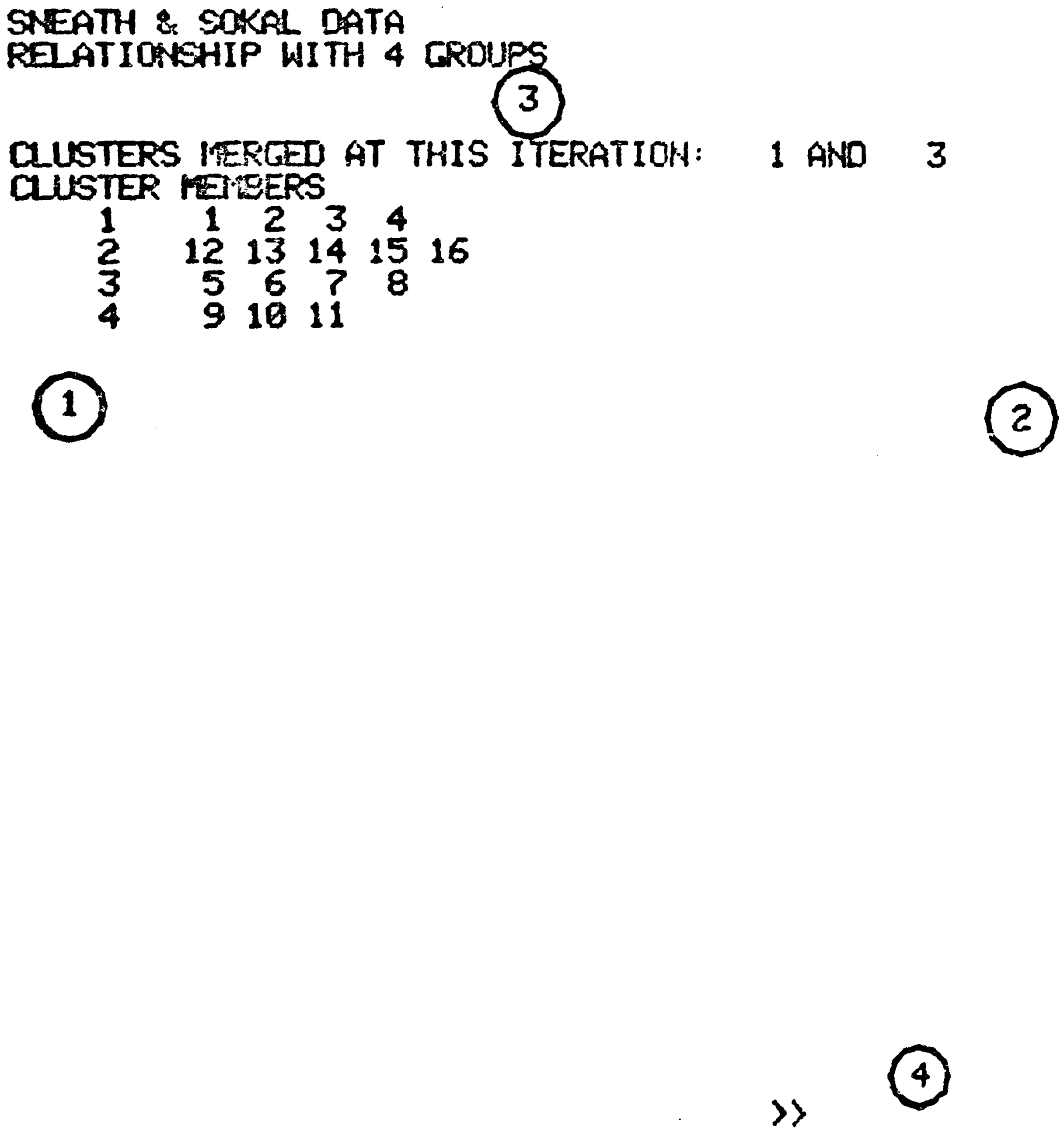
Euclidean clustering - 6 groups



```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 5 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   2 AND   3
CLUSTER MEMBERS
     1    1  2  3
     2   12 13 14 15 16
     3    4  5
     4    6  7  8
     5    9 10 11
```

Figure 6.6

Euclidean clustering - 5 groups



```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 4 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   1 AND   3
CLUSTER MEMBERS
     1    1  2  3  4
     2   12 13 14 15 16
     3    5  6  7  8
     4    9 10 11
```

```
                                    >>
```

Figure 6.7

Euclidean clustering - 4 groups

```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 3 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   1 AND   3
CLUSTER MEMBERS
     1    1  2  3  4  5  6  7  8
     2   12 13 14 15 16
     3    9 10 11
```

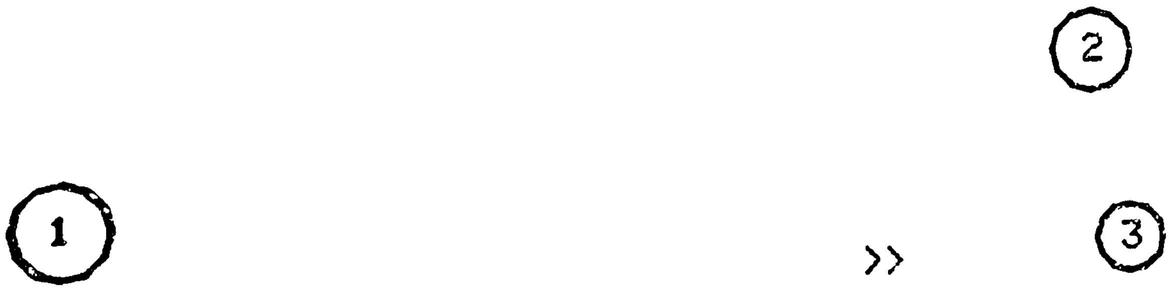


Figure 6.8

Euclidean clustering - 3 groups

```
SNEATH & SOKAL DATA
RELATIONSHIP WITH 2 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   2 AND   3
CLUSTER MEMBERS
    1    1  2  3  4  5  6  7  8
    2    9 10 11 12 13 14 15 16
```

Figure 6.9

Euclidean clustering - 2 groups

CHAPTER 7

ORDINATION TECHNIQUES

7.1 INTRODUCTION

The clustering methods described in the previous three chapters have concentrated on investigating the relationships within a set of objects by imposing some structure on the data · the dendrogram tree or a set of partitions - thus implying that the activities on which observations were recorded fall into one or more classes, which may be arranged either hierarchically or in the form of nonoverlapping clusters. This may be misleading for no such structure may actually exist in the data. It is therefore useful to make available methods of analysis which do not present their results in such a clear cut way. Ordination methods do not require such assumptions, but instead attempt to represent the distance (dissimilarity) relationships among the activities in a space of reduced dimensionality. Any groupings present in the data should then be apparent from visual examination of scatter plots, provided that the distortion introduced by the low dimensional representation is small and that the number of activities is not excessive. As Cormack (1971 p.340) remarks: "When the data have not been forced into clusters, the observer can assess better whether clusters exist".

This chapter describes several methods of analysing a set of objects, in which the basic aim is to represent each object by a point in some Euclidean space so that the objects which are similar to one another are represented by points which are close together. The configuration of points is then investigated in an attempt to detect any underlying structure in the data. As the interpretation of high dimensional data is extremely difficult two or three dimensional representations are derived in such a way as to retain as much of the high dimensional information as possible. There are various ways of measuring information loss, some of which are described later, but it should be noted that these measures are not used in any formal statistical manner as it is not appropriate to regard the data as coming from an underlying population with certain associated distributional properties.

As is the case with clustering methods several different techniques have been developed to achieve the same end result of a configuration of points representing activities. The techniques that transform the high dimensional data into a two or three dimensional space are generally known as mapping techniques, and fall into two distinct types: iterative and noniterative. The noniterative mapping is a unique representation calculated by a precise mathematical formula. The iterative techniques utilise search procedures to determine the low dimensional representation through a series of transformations.

This chapter discusses a number of techniques, examining the advantages and disadvantages of each method. The general problems involved in the representation and interpretation of data in two dimensional space are also discussed. Two particular ordination techniques are incorporated in MAGIC and are described in detail in chapters 8 and 9.

7.2 SIMPLE ORDINATION

This method of reducing the dimensionality of data is perhaps the simplest possible as it only involves arithmetic operations on the data, and, as it introduces a number of basic concepts is described in some detail here. The basic idea is as follows. From the original high dimensional space, choose the two points that are furthest apart. Let those two points be denoted by X_a and X_b. A straight line passing through the two points is chosen as the first ordination axis. To determine the second ordination axis, a straight line perpendicular to the first and passing through a third point, denoted by X_c, that is furthest removed from the first axis, is constructed. When the projections of the points in the original space are plotted on these two new axes, the resulting two dimensional display represents a projection of the high dimensional data into a space defined by the two ordination axes. The projection of the points into a k-space ($k \geqslant 2$) can be similarly accomplished utilising k new ordination axes.

7.2.1 Efficiency Of The Method

The reduction of dimensionality from the original space to a 2-space is achieved at the expense of distance relationships between the points. Since the distance relationship cannot be exactly preserved, it is of interest to determine the amount of distortion resulting from this technique. The distortion due to projecting the points onto the two ordination axes is maximum when the first axis coincides with the direction of maximum variation between the points and the second axis is so positioned that it accounts for a maximum portion of the variation of the points.

To determine the efficiency of the simple ordination method it is possible to examine how each of the two ordination axes accounts for the interpoint distances. First, if h orthogonal axes were constructed, the distance relationship is preserved exactly. That is,

$$\hat{d}_{ij} = d_{ij}$$

where

$$\hat{d}_{ij} = |X_i - X_j|$$

and

$$d_{ij} = |Y_i - Y_j|$$
$$= \left[\sum_{k=1}^{h}(Y_{ik} - Y_{jk})^2\right]^{1/2}$$

However only the first two ordination axes are constructed giving

$$\hat{d}_{ij} \geqslant d_{ij}$$

where

$$d_{ij} = \left[\sum_{k=1}^{2}(y_{ik} - y_{jk})^2\right]^{1/2}$$

The efficiency of the kth axis (k = 1,2) in accounting for the original interpoint distances can be defined by the ratio

$$r_k = \frac{\sum_{i<j} d_{ij,k}}{\sum_{i<j} \hat{d}_{ij}^2} \qquad i = 1,2, \text{ and } k = 1,2$$

where $d_{ij,k} = y_{ik} - y_{jk}$ which is the difference between the projection of X_i and that of X_j onto the kth axis. The sum of the two ratios

$$r = r_1 + r_2$$

expresses the overall efficiency of the two ordination axes in accounting for the interpoint distances.

Another method for determining the efficiency of the ordination axes is to define an error function E. This function should measure how well the N vectors in the 2-space fit with the N vectors in the h-space on the basis of the interpoint distances. One such error function is defined by

$$E = f(\hat{d}_{ij} - d_{ij})$$
$$\frac{1}{\sum\limits_{i<j} \hat{d}_{ij}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{\hat{d}_{ij}}$$

A smaller value of E means a good fit and the corresponding axes may be considered to be efficient. Other error functions of this type may be similarly defined The measure above is the one used in the Nonlinear Mapping described in chapter 8.

7.3 PRINCIPAL COMPONENTS

There is an obvious deficiency with the simple ordination method. The mapping of the N points from the h-space into the 2-space is determined by only three reference points. Clearly, if some structural relationship is to be preserved, the entire collection of the N points, or a characteristic summary of these points, should be used. The method of principal components is one solution to this problem.

Let a typical point X_i in the h-space be represented by

$$X_i = (x_{i1}, x_{i2}, \ldots., x_{ih})$$

Given a collection of N such points in the h-space, these N points can be represented by an h x N matrix X.

$$X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \cdot & \cdot & \cdots & \cdot \\ x_{1h} & x_{2h} & \cdots & x_{Nh} \end{bmatrix}$$

Each column of the matrix represents a data point in the h-space. Let the sample mean of the N points be $\bar{X}$ where

$$\bar{X} = 1/N \sum_{i=1}^{N} X_i$$

with the kth component in $\bar{X}$ calculated by

$$\bar{x}_k = 1/N \sum_{j=1}^{N} x_{jk}$$

Each point X_i measured as a deviation from the sample mean is denoted by

$$X_i - \bar{X} = \begin{bmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ x_{ih} - \bar{x}_k \end{bmatrix}$$

The N points as measured from $\bar{X}$ can be represented by a matrix X_c

$$X_c = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdot\ \cdot & x_{N1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & \cdot\ \cdot & x_{N2} - \bar{x}_2 \\ \cdot \quad\quad \cdot & & \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot & & \cdot \quad\quad \cdot \\ X_{1h} - \bar{X}_h & \cdot\ \cdot & x_{Nh} - \bar{X}_h \end{bmatrix}$$

The total scatter matrix (with respect to the centroid) may be defined as

$$S = x_c x_c^t = [s_{ij}]$$

The element s_{ij} in the matrix S is calculated by

$$\begin{aligned} s_{ij} &= \sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \\ &= \sum_{k=1}^{N} x_{ki} x_{kj} - N\bar{x}_i \bar{x}_j \quad i,j = 1,\ldots, h \end{aligned}$$

To map a point in the h-space to a point in the d-space (d=2 for a two-dimensional display), an origin and d orthogonal axes passing through the origin must be selected. Assuming d axes are selected such that the sum of squares of the projections from the points to the axes is a minimum, it can be shown the d axes pass through the centre of gravity of the N points. Furthermore, the sum of squares of the projections onto the d axes defined by d orthogonal vectors $Q_1, \ldots, Q_d$ is

$$\sum_{i=1}^{h} s_{ii} - \sum_{i=1}^{d} Q_i^t SQ_i$$

Minimising this expression is equivalent to maximising $\Sigma Q_i^t SQ_i$. This maximum is obtained if the d axes are chosen to satisfy

$$Q_i = P_i \quad i = 1, \ldots, d$$

where P_i, i = 1, 2, ..., d, are the first d eigenvectors of the scatter matrix S.

The actual representation of the N points in the d-dimensional space, defined by the d eigenvectors P_1, ..., P_d, is accomplished by computing the d coordinates for each point X_i and forming the vector Y_i.

$$Y_i = (P_1X_i,\ P_2X_i, \ldots, P_dX_i) \quad i = 1, \ldots, N$$

The N points in the d-space can be represented by a Y matrix.

$$Y = \begin{bmatrix} P_1X_1 & P_1X_2 & \cdots & P_1X_N \\ P_2X_1 & P_2X_2 & \cdots & P_2X_N \\ \cdot\ \cdot & \cdot\ \cdot & \cdots & \cdot\ \cdot \\ \cdot\ \cdot & \cdot\ \cdot & \cdots & \cdot\ \cdot \\ P_dX_1 & P_dX_2 & \cdot\ \cdot & P_dX_N \end{bmatrix}$$

To plot the N points in a two-dimensional space with the centre of gravity as the origin, the first two coordinates of N points are computed. This corresponds to the first two rows of the Y matrix. Transferring the origin to the centre of gravity

$$\begin{bmatrix} P_1\bar{X} \\ P_2\bar{X} \end{bmatrix}$$

the two coordinates of the N points in the 2-space are obtained

$$\begin{bmatrix} P_1(X_1-\bar{X}) & \cdots & P_1(X_N-\bar{X}) \\ P_2(X_1-\bar{X}) & \cdots & P_2(X_N-\bar{X}) \end{bmatrix}$$

7.3.1 A Brief Critique Of Principal Components

Although an improvement on simple ordination insofar as all the data set is considered Principal Components Analysis is still not without its problems. Scaling, for example, affects the results, and, unless all the variates in X are measured in the same units, different results will be obtained for a change in scale. Principal Components Analysis is not alone in this and the effect of scale is discussed in section 7.7.2, however it is worth noting now that the distance measures used in 7.2.1 and 7.3 have nonsensical physical dimensions when the variates are measured in different scales. To evade this difficulty it is common practice to normalise variates by dividing each by its sample standard error. Bartlett (1951) discusses the effect of normalisation. Other normalisers could equally well be used: Jolicoeur (1963), for instance, has shown that a simple logarithmic transformation of all the variates will also eliminate the effects of scaling. None of these techniques, however, are entirely satisfactory when we know that the majority of our data is measured on different scales. The solution adopted in MAGIC makes use of Gower's General Coefficient of Similarity, described in 4.5.3, and an extension of Principal Components Analysis, known as Principal Coordinates

Analysis, especially developed to operate on mixed data sets.

7.3.2 Principal Coordinates Analysis

This technique was developed by Gower (1966). The starting point is the matrix D of dissimilarity coefficients (such as Euclidean distances) which is transformed into the matrix E by the relationship

$$e_{ij} = -0.5d^2_{ij}$$

Alternatively a matrix of similarity coefficients E can be computed. In MAGIC Gower's general similarity coefficient is used. This has the advantages that (i) it can handle quantitative, binary or qualitative variables, and (ii) the resulting matrix E is always symmetrical and positive semi-definite, that is, the n objects can be represented as a set of points in Euclidean space (Sneath and Sokal 1973, p. 136). The coefficients e_{ij} are computed seperately for the three types of variable and are then weighted by the reciprocal of the number of variables involved and the resulting values summed. That is,

$$e_{ij} = e^Q_{ij} /p^Q + e^B_{ij} /p^B + e^M_{ij} /p^M$$

where e^Q_{ij}, e^B_{ij} and e^M_{ij} are the values of the coefficient for quantitative, binary and multistate variahles respectively and p^Q, p^B, and p^M are the numbers of such variables.

E is then modified so that the mean of each row and column is removed, since the mean is unimportant in the determination of the distance between any two points. The modified matrix $F = (f_{ij})$ is obtained by

$$f_{ij} = e_{ij} - \bar{e}_i - \bar{e}_j + \bar{e}$$

where $\bar{e}_i$, $\bar{e}_j$ and $\bar{e}$ are the means of the ith column and the jth row, and the overall mean, respectively. A specified number of the largest eigenvalues of F, and the corresponding eigenvectors are then determined. The magnitude of the kth eigenvalue gives the relative importance of the kth dimension in the determination of the variation in interpoint distances. Published results (e g. Blackith and Reyment, 1971, p. 167) indicate that much of this variation is contained in the first two or three dimensions. The eigenvectors give the coordinates of the n points. These coordinates may then be plotted. A more extended discussion of this method is contained in chapter 9.

7.4 MULTIDIMENSIONAL SCALING

An alternative approach to ordination argues that because of the problems in deriving dissimilarities, their precise values are unreliable, and may contain little useful information beyond their rank ordering. This is the only information about the dissimilarities used in the method of nonmetric multidimensional scaling developed by Shepard (1962a,b) and Kruskal (1964a,b). As much use is made of this method in quantitative psychology to reduce the dimensionality of problems and a large literature has developed it is worth looking at this often ill-understood and misapplied technique in some detail. A brief history of the early work in multidimensional scaling is given by Shepard (1972).

The technique, in essence, follows the basic ordination method of 7.2 with the use of an error function to assess goodness of fit. There are a number of alternative formulations, but the following derivation is based on Lingoes and Roskam (1973).

7.4.1 The Underlying Assumptions

Shepard (1962a,b) argued that in obtaining a geometrical representation, one wanted to ensure that the interpoint distances ($\hat{d}_{ij}$) were monotonically related to the given dissimilarities (d_{ij}); the relationship might not be exactly monotone for distances based on a low-dimensional configuration of points, but one wanted to ensure that, on the whole, the larger the dissimilarity, the larger the corresponding distance.

This monotonic model means that it is not assumed that the set of dissimilarities contains any metric information; all that is used is their rank ordering. The method has thus been called a nonmetric multidimensional scaling method, but as the result is a geometrical configuration of points - which certainly contains metric information - it may be more appropriately described as an ordinal scaling method.

Shepard (1962a,b) presented a heuristic algorithm for seeking a configuration approximately satisfying the monotonicity requirement; his approach did not involve an explicit minimisation of some function measuring the departure from perfect monotonicity between dissimilarities and distances. These ideas were formalised by Kruskal (1964a,b), who proposed the method of least squares monotone regression. This is a general method of comparing two sequences of real numbers, and it will be convenient to introduce it in a general context before discussing its application to comparing sets of dissimilarities and distances.

Assume that a, b and c are three sequences, each containing m real numbers, $(a_1,....,a_m)$, $(b_1,....,b_m)$ and $(c_1,....,c_m)$, respectively. In the following description, a is a sequence in which only the ordering is of interest: a and b will later be identified with (d_{ij}) and $(\hat{d}_{ij})$, respectively; c is a sequence which will be used in the comparison of a and b.

Two possible definitions of monotonicity are:

(i) c is primarily monotone increasing (PMI) over a if

$$a_k < a_l \text{ implies that } c_k \leqslant c_l \quad (1 \leqslant k,l \leqslant m)$$

(ii) c is secondarily monotone increasing (SMI) over a if

$$a_k \leqslant a_l \text{ implies that } c_k \leqslant c_l \quad (1 \leqslant k, l \leqslant m)$$

These two definitions of monotonicity differ only in their treatment of ties in the sequence a. In the secondary definition, these ties must be preserved in c: if a_k equals a_l, then c_k must equal c_l. In the primary definition of monotonicity, ties in a may be broken in either direction in c.

Having defined monotonicity, c is required to be monotone (either PMI or SMI) over a and, subject to this constraint, to resemble b as closely as possible. An example may clarify this idea. Assume that a = (1,2,4,4,6,8,9,10,11,15) and b = (1,4,5,6,7,8,12,13,13,14). The points $\{(a_k,b_k), k=1,....,10\}$ are plotted as crosses in figure 7.1.

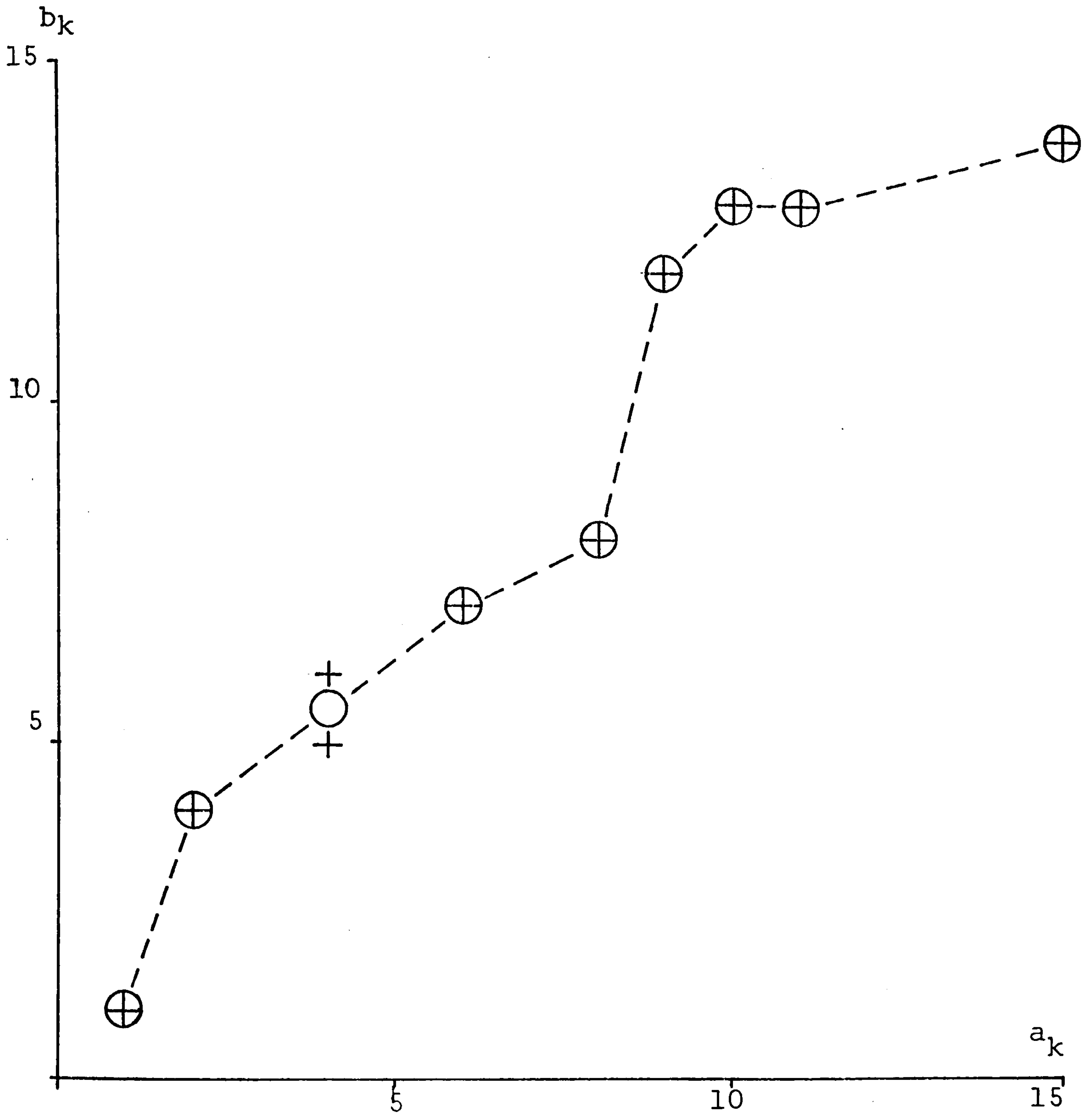


Figure 7.1

A plot of the artificial sequence $(a_k,b_k),k=1,\ldots,10$ described in the text, and the secondary least squares monotone regression c of b on a: the points (a_k,b_k) and (a_k,c_k) are represented by crosses and open circles

If the monotonicity requirement is PMI, the equality between a_3 and a_4 can be broken, and by choosing $c_k = b_k (k=1,\ldots.10)$ perfect resemblance between c and b, with c satisfying the primary monotonicity requirement is obtained. If the monotonicity requirement is SMI it is necessary for $c_3=c_4$, and - because b_3 does not equal b_4 - a perfect resemblance between c and b is not possible. To determine the optimal shared value for c_3 and c_4 a definition of what is meant by the requirement that c should resemble b "as closely as possible" is necessary. To measure the departure from a perfect fit Kruskal (1964a) suggested that a sum of squares criterion should be used:

$$S^*(c) = \sum_{k=1}^{m}(b_k - c_k)^2$$

This criterion will be minimised when c_3 and c_4 are both chosen to be 5.5, the mean of b_3 and b_4; in the general solution, c will be split up into a set of blocks containing elements with consecutive indices, e.g. $(c_{r+1}, c_{r+2}, \ldots., c_s)$, such that each element in the block equals the mean of the corresponding set of values in b - in this case

$$\sum_{k=r+1}^{s} b_k/(s - r)$$

and such that the common value increases from block to block.

For given sequences a and b, the sequence c which reduces $S^*(c)$ to its minimum value (S^*, say) subject to being PMI (SMI) over a is called the primary (secondary) least squares monotone regression of b on a; S^* is called the primary (secondary) raw stress. The secondary least squares monotone regression c of b on a for the artificial example is shown by the set of open circles in figure 7.1. The form of the regression "function" in between successive circles is only required to be monotone increasing; for illustrative purposes, straight line sections have been drawn in figure 7.1.

The description thus far has been in terms of computing a pair of sequences, a and b, and the least squares monotone regression method may be regarded as an alternative to other regression methods. However, the theory may also be applied to comparing a set of dissimilarities (d_{ij}) and a set of distances ($\hat{d}_{ij}$). Thus, the m(=10) elements of a could be the n(n-1)/2 (= 10 for n = 5) pairwise dissimilarities for a set of five objects, and the elements of b could be the interpoint distances for five points representing the same five objects.

This provides a method of measuring, in terms of raw stress S*, the resemblance between given sets of dissimilarities (d_{ij}) and distances $(\hat{d}_{ij})$. The value of S* is not invariant under uniform dilation of the geometrical configuration; this undesirable property is removed by dividing by a normalising factor, $T^* = \Sigma \hat{d}_{ij}^2$. Then, the normalised stress is defined by

$$\text{Stress, } S = (S^*/T^*)^{\frac{1}{2}} = \left[\frac{\Sigma(\hat{d}_{ij} - c_{ij})^2}{\Sigma \hat{d}_{ij}^2}\right]^{1/2}$$

thus ensuring that S is bounded by 0 and 1. In this expression, (c_{ij}) is the (primary/secondary) least squares monotone regression of $(\hat{d}_{ij})$ on (d_{ij}), i.e. the set of values which minimises $S^*(c)$ subject to being (primarily/secondarily) monotone increasing over (d_{ij}); the summation being taken over all (or some) of the pairs of values (i,j). It is possible to formalise this description into a general model.

7.4.2 A General Model

The following terms and matrices are used:

(1) P = a r-element array or vector of arbitrary indices of similarity or dissimilarity between all pairs of n objects or variables, having general element p_{ij}, r = 0.5n(n-1) and the r pairs of subscipts are generated by

taking for each first subscript i a second subscript j = i+1, i+2,....,n(i = 1,2,....,n-1). Thus the elements of a n-square matrix of relations on pairs of objects (where $p_{ij}=p_{ji}$) are systematically ordered in an array. Both the diagonal of this matrix and one half of the off-diagonal elements are ignored.

(2) Δ = a r-element vector of real numbers with elements $\delta_{ij}= f(p_{ij})$, such that whenever $p_{ij} > p_{kl}$ (for similarity data) or $p_{ij} < p_{kl}$ (for dissimilarities) then either: (a) $\delta_{ij} < \delta_{kl}$(semi-strong monotonicity when some P are tied and strong monotonicity when there are no ties in P) or (b) $\delta_{ij} \leqslant \delta_{kl}$ (weak monotonicity for no ties in P and semi-weak monotonicity when ties exist in P), for all i,j,k, and l, where i ≠ j and k ≠ l, i.e. $\Delta \rightarrow$ P monotonically. The Δ vector represents a monotonic transformation of the P vector having certain statistical properties in addition to the mathematical ones defined above, whose function is to weight the iterations for moving the configuration towards its goal and to form the basis for evaluating progress at any given iteration (see (6) below).

(3) X = a nxm matrix of rectangular coordinates (the configuration), where m is the number of dimensions.

(4) D = a r-element vector of distances calculated from X among the n points embedded in a Euclidean space according to the standard distance formula:

(5) $$d_{ij} = \left(\sum_{\alpha=1}^{m} (x_{i\alpha} - x_{j\alpha})^2 \right)^{1/2}$$

Now, given P, some initial configuration X, a fixed m, and the distances calculated from (5) above, the problem of nonmetric multidimensional scaling can be formulated in terms of the minimisation of a function of two sets of unknowns, namely D and Δ. To obtain the D as close as possible to the Δ (possibly with certain restrictions on the Δ vis-a-vis the D) one obvious formulation is in the form of a normalised least-squares function, in this context termed the loss function. Its value is denoted by L. The loss function is defined by:

(6) $$L = \left(\sum_{ij} (d_{ij} - \delta_{ij})^2 / \sum_{ij} d_{ij}^2 \right)^{1/2}$$

This function is formally equivalent to Kruskal's (1964a) stress, but (6) does not assume any particular definition of the Δ apart from (2). By its construction (6) is also similar to a function defined by Guttman (1968).

7.4.3 Nonlinear Mapping

Nonlinear mapping, although developed independently by Sammon (1969), is similar both in concept and execution to multidimensional scaling methods in that n points in a h-dimensional space are projected onto a d-dimensional subspace ($d < h$) with a minimum of distortion. Sammon's computational technique is somewhat simpler than that of Kruskal (1964b). The output consists of the values of a goodness of fit function, termed mapping error, and a two or three dimensional representation of interpoint relationships. Nonlinear mapping does not attempt to ensure monotonicity between observed dissimilarities and calculated distances; rather, the goodness of fit function measures the amount of distortion of interpoint distances introduced by mapping onto a d- (as opposed to a h-) dimensional space. The function minimised is:

$$E = \frac{1}{\sum_{i<j}^{N} \hat{d}_{ij}} \cdot \sum_{i<j}^{N} \frac{(\hat{d}_{ij} - d_{ij})^2}{\hat{d}_{ij}}$$

where $\hat{d}_{ij}$ is an observed dissimilarity and d_{ij} a distance measured in a d-dimensional space. The initial d-space representation of the points, Y, is chosen randomly.

Given the matrix X, from which the interpoint distances D $(=d_{ij})$ are computed, and $\hat{D}$ $(=\hat{d}_{ij})$, the matrix of dissimilarities, the method of steepest descent is used to locate a minimum of E, by computing the d-space coordinates Y at iteration (m+1) from:

$$y_{pq}(m+1) = y_{pq}(m) - MF.\,\phi_{pq}(m)$$

where MF is a parameter termed by Sammon the "magic factor" (a fixed step length which Sammon determined empirically to perform best in the range 0.3 - 0.4), and ϕ_{pq} is the ratio of the first to the second-order partial derivatives of E with respect to y_{ij},

that is $\phi_{pq}(m) = \dfrac{\delta E(m)}{\delta y_{pq}(m)} \Bigg/ \left|\dfrac{\delta^2 E(m)}{\delta y_{pq}(m)^2}\right|$

These derivatives are defined as:

$$\frac{\delta E}{\delta Y_p} = \frac{-2}{c}\sum_{j=1}^{N}\left[\frac{\hat{d}_{pj} - d_{pj}}{\hat{d}_{pj}d_{pj}}\right](Y_p - Y_j)$$

and

$$\frac{\delta^2 E}{\delta Y_p^2} = \frac{-2}{c}\sum_{j=1}^{N}\frac{1}{\hat{d}_{pj}d_{pj}}\left[(\hat{d}_{pj} - d_{pj}) - \frac{(Y_p - Y_j)^2}{d_{pj}}\left(1 + \frac{\hat{d}_{pj} - d_{pj}}{d_{pj}}\right)\right]$$

The algorithm terminates when a fixed number of iterations have been carried out or whenever E has converged to a suitably small value. A more detailed discussion of this technique is contained in chapter 8.

7.5 A COMPARISON OF ORDINATION TECHNIQUES

This section attempts briefly to indicate the similarities and differences between the various techniques discussed in this chapter and to point out some of their advantages and disadvantages.

The first point to make is that the first two "serious" techniques discussed, namely principal components and principal coordinates, are both latent root and vector methods, while multidimensional scaling and nonlinear mapping both operate by minimising a particular function using some iterative algorithm. The former methods have, therefore, obvious computational advantages. Of these, principal coordinates is perhaps the most powerful for obtaining a low-dimensional representation of data since it is not as restrictive as principal components analysis. In particular it is not necessary to consider only data sets for which Euclidean distance is considered appropriate. The only advantage of this particular distance measure is that it allows the principal coordinates to be related linearly to the original variable values. Principal coordinates analysis also has the advantage of being directly applicable to data given in the form of a distance or similarity matrix.

In many respects the mathematical formulations of non-metric multidimensional scaling and nonlinear mapping are similar. However, the mapping criteria, "stress" and "mapping error", are quite different. A major distinction is that multidimensional scaling employs only the ordinal properties of the similarities or distances being used. Gower (1966) discusses the relationship between principal coordinates analysis and nonmetric multidimensional scaling. He concludes that where the former gives an adequate fit in two dimensions, then the solution will be similar to the one that would be found employing the latter method. Gower points out, however, that multidimensional scaling may be able to find a good fit in a low number of dimensions when principal coordinates may not; because of the differing computational complexities of the two methods Gower finally recommends the initial use of principal coordinates analysis, and, where this does not lead to a solution of sufficiently low dimensionality, he suggests using the coordinates found as a starting point for the iterative algorithm of nonmetric multidimensional scaling. Sammon (1969) gives some interesting examples when nonlinear mapping recovers the structure in some specially constructed sets of multivariate data, whilst principal component plots of the same data fail to reveal this correctly. Sibson et al (1981) discussed several possible models for obtaining a set of dissimilarities from an underlying configuration, and described a simulation study which compared the abilities of several scaling methods, including principal coordinates analysis and nonmetric multidimensional scaling. They concluded that, provided the iterations started from a reasonable configuration, nonmetric multidimensional scaling was never significantly worse, and under some models for the dissimilarities was considerably better, than the methods which used the numerical values of the dissimilarities.

Studies have also been carried out to investigate the nature of the differences between the configurations produced when different geometrical methods are used to analyse the same data. For example, Rohlf (1972) noted that nonmetric multidimensional scaling tended to depict differences between similar objects more accurately than principal coordinates analysis, but did not necessarily represent the distances corresponding to smaller and larger dissimilarities in the same scale. Further comparisons are reported by Gower (1972), Fasham (1977) and Prentice (1977).

If the background to the problem indicates that one method of analysis is particularly appropriate, it should be used, but in order to simplify the use of MAGIC the choice is limited to one technique of each type, and, for reasons of computational efficiency, those are principal coordinates analysis and nonlinear mapping. Nonmetric multidimensional scaling, involving the minimisation of a function of nt variables for each value of the number of dimensions, t, makes heavy demands on computing resources compared to both principal coordinates analysis, in which the main work to be carried out involves the eigenanalysis of an (nxn) matrix, and nonlinear mapping. Given this fact, it would seem preferable to use principal coordinates analysis and nonlinear mapping.

7.6 RELATION BETWEEN ORDINATION AND CLUSTERING

Ordination and clustering techniques are both methods of analysing data, but rather than being in competition with each other, are essentially complimentary. They can be used together in several ways and these joint uses are usually desirable. The basic relationship may be seen in terms of the data sets used, and is summarised in figure 7.2.

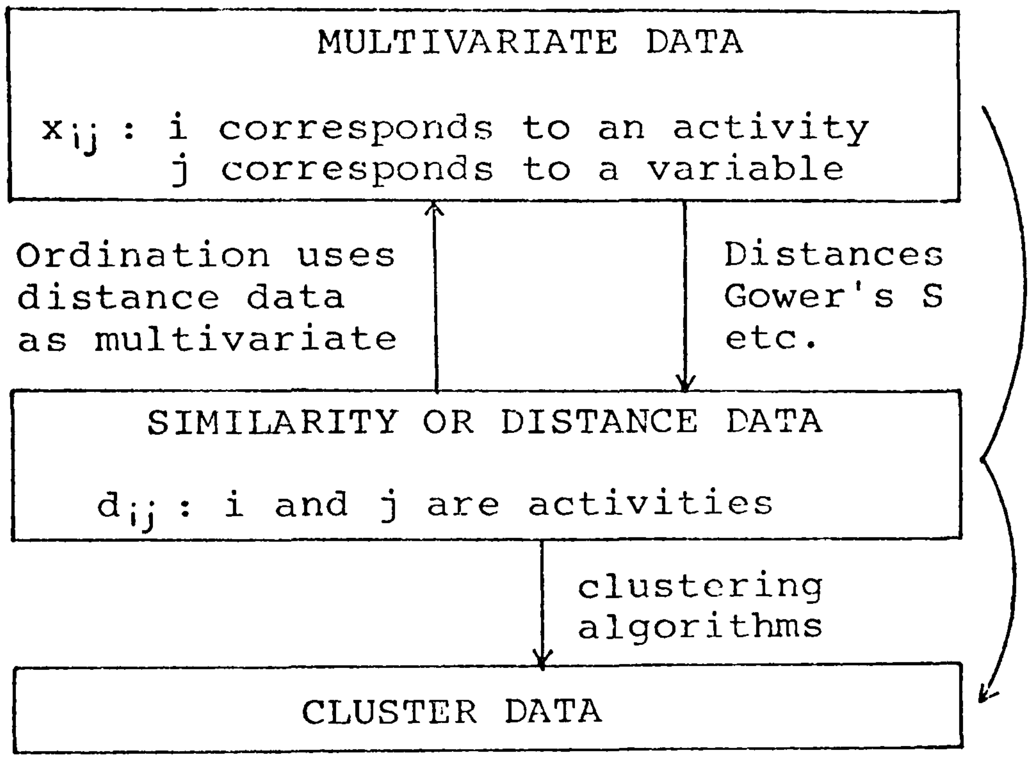

MULTIVARIATE DATA
x_{ij} : i corresponds to an activity
j corresponds to a variable
Ordination uses
distance data
as multivariate
Distances
Gower's S
etc.
SIMILARITY OR DISTANCE DATA
d_{ij} : i and j are activities
clustering
algorithms
CLUSTER DATA

Figure 7.2

Relationship between ordination and clustering

Two main types of data are used - the multivariate data and proximity data. Clustering algorithms may operate directly on proximity data, but if we start with mixed multivariate data, in order to determine clusters we must first convert the multivariate data to proximity data. Strictly speaking therefore, cluster algorithms operate only on the second stage, the first stage being a necessary preliminary. We may, therefore, consider clustering as a procedure which starts with one type of data and converts it to some other type. The conversion of multivariate data to proximity data used in MAGIC utilises Gower's similarity and Euclidean distance (dissimilarity).

Ordination may be thought of as a transformation in the other direction, converting proximity data to multivariate data in the form of a configuration of points in low-dimensional space.

Another dimension is the application. The main purpose is simply, in exploratory data analysis, "to see whats there". A second purpose is to comprehend the data more clearly, and a third is to provide information to aid subsequent design work.

Another aspect is the distinction between "natural clusters" which may exist in the data and "artificial" clusters which may arise as a result of the clustering method used.

The key difference between ordination and clustering is that ordination provides a spatial representation of the proximity data, whilst clustering provides a tree representation. In hierarchical clustering small clusters tend to be well identified and are often meaningful, but large clusters higher up the tree tend not to fit so well. On the other hand ordination deals much more with the overall relationships. Small changes in the data may cause changes in local position and arrangement, but it is the general position of the points within the configuration which is important. For example, the fact that certain points near the middle of the configuration will not change, even though the arrangement at the middle may vary.

Since ordination and clustering are sensitive to complimentary aspects of the data (the large dissimilarities or overall picture in ordination, and the small dissimilarities or local structure in clustering) it is appropriate to use them both on the same data set. It is in fact possible to combine the results into a single diagram using a two-dimensional ordination (figure 7.3). The position of the points are obtained from the ordination, whilst the loops show the groupings obtained from the clustering. The figure uses the Sokal and Sneath data and the results of the group average clustering and a nonlinear mapping.

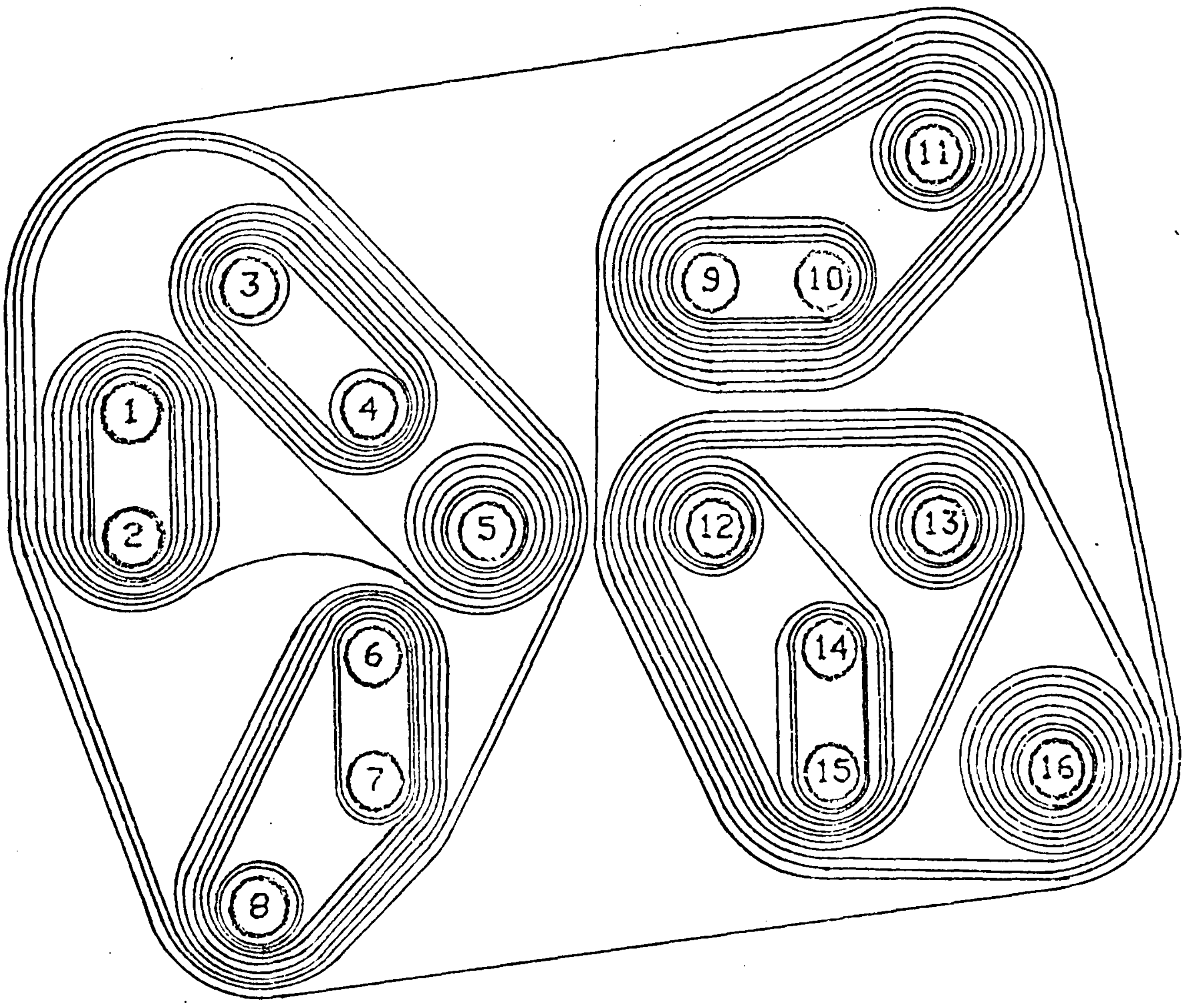


Figure 7.3

Ordination and clustering - combined plot

7.7 TWO-DIMENSIONAL DISPLAYS

A number of difficulties may arise mapping higher-dimensional vectors into a two-dimensional space for visualisation. Particular problems are the iterative nature of the mapping algorithms and the interpretation of the clusters in the two-dimensional space.

7.7.1 The Problem Of Local Minima

In principle the iterative steps of mapping algorithms to determine the final configuration are not difficult to implement. There is, however, a potential difficulty in the criterion of termination. A configuration of N vectors from which no small movement of vectors is an improvement corresponds by definition to a local minimum of the error function E. The difficulty with a local minimum is that it may or may not be the global minimum whose corresponding configuration is really being sought. When searching for the minimum of E using steepest descent or other techniques, there is no sure way to prevent finding a local minimum. Figure 7.4 shows an error function of one variable with several local minima. Only one of them (point B) is the true global minimum.

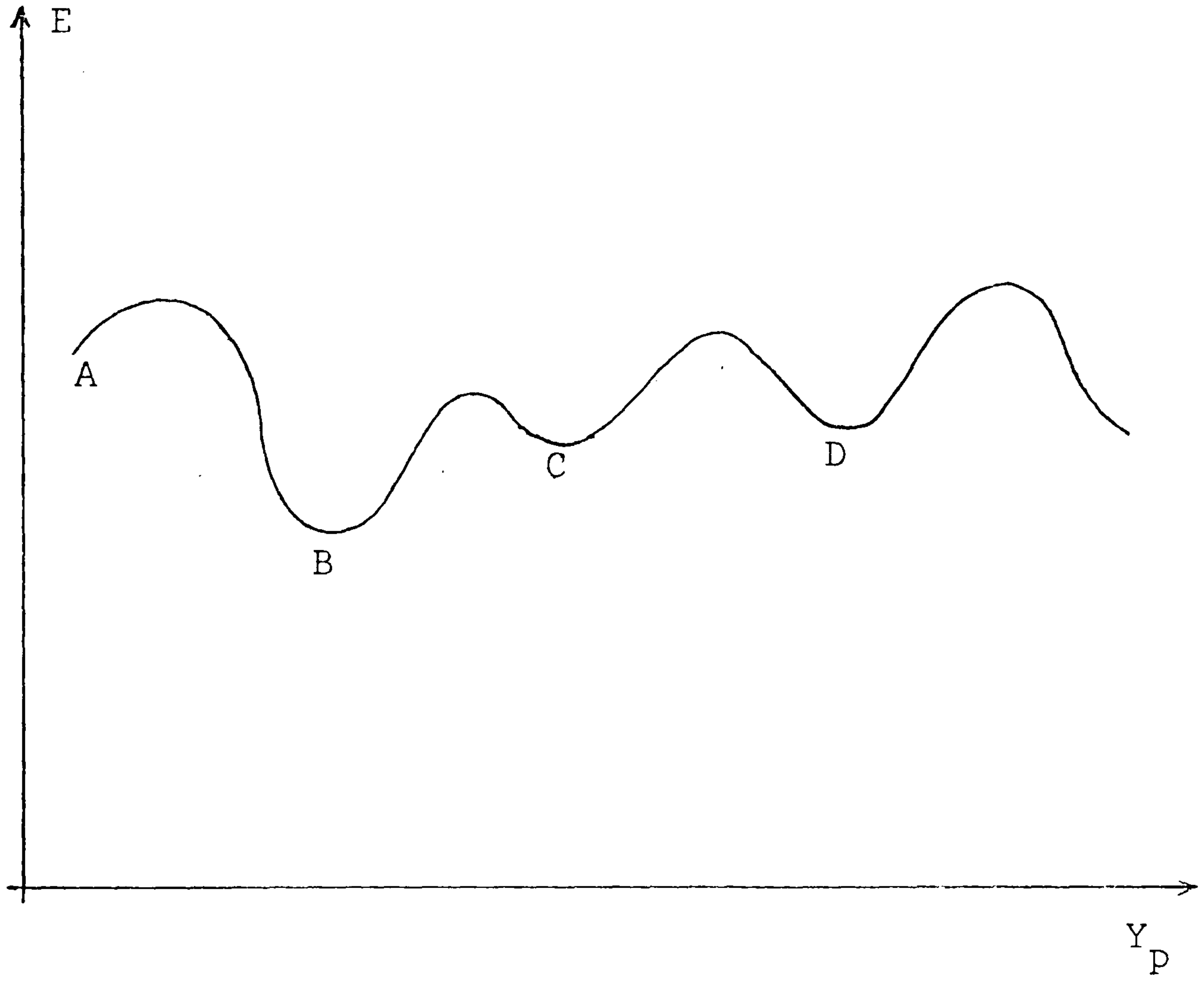


Figure 7.4

An error function with several local minima (A, B, C, D)

This local/global minimum difficulty is of course not unique in mapping algorithms. It is a widely known problem in all search and minimisation problems. In the implementation in MAGIC if the display is drawn the minimisation has reached a preset level of accuracy. If this level is not reached after the default number of iterations the current solution is displayed. This may either be compared with one of the other analyses and accepted if it appears reasonable, or else points may be interactively moved to "jump" the configuration out of the possible loal minimum, and the calculation restarted from the new position.

7.7.2 The Effect Of Scaling

In an architectural problem the measurements in the data are often composed of a variety of units. Using them in their original form in the vector representation of the data has serious complications. first, the units of the measurements define an implicit weighting of the vector components. Second, when different units are combined to achieve a single measure of distance, the meaning of that distance measure is nonsensical.

The following figures illustrate the effects of units of measurement upon the graphical analysis of four elements represented by two-dimensional vectors. Let

$$X_1 = (1,\ 0)$$

$$X_2 = (1,\ 0.5)$$

$$X_3 = (4,\ 0)$$

$$X_4 = (4,\ 0.5)$$

where the first variable is distance (say metres) and the second mass (say kilograms). The four vectors are plotted in figure 7.5 in which two different scales are used for plotting. Visually identified clusters indicate two possible configurations due to different scalings used in the plots. Various normalisation techniques have been developed, but all involve a distortion of the data. MAGIC again overcomes the problem by the use of Gower's general coefficient.

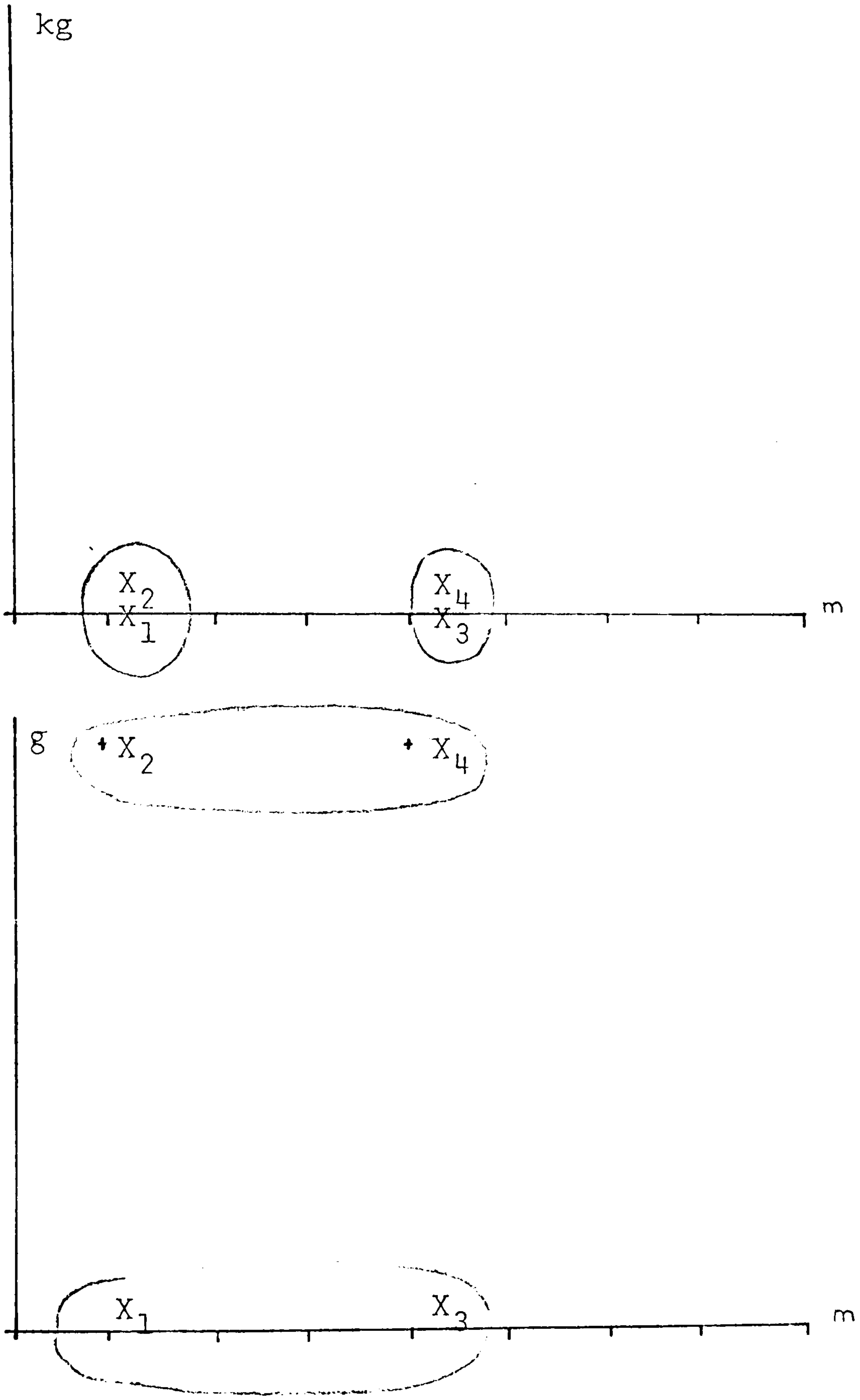


Figure 7.5

Effect of scaling upon graphical analysis of data.

Visually identified clusters are circled.

7.7.3 Sample Size

Another problem occurs when the ratio of the number of activity vectors to the number of variables is small, and deceptive results may be obtained. Foley (1972) showed how misleading mappings may arise, depending on the sample size, n, and the number of variables measured, h. Foley derived expressions of the estimated probability of error as a function of n and h, for a number of underlying probability distributions. Figure 7.6 shows a typical plot of the estimated probability of error as a function of the ratio n/h.

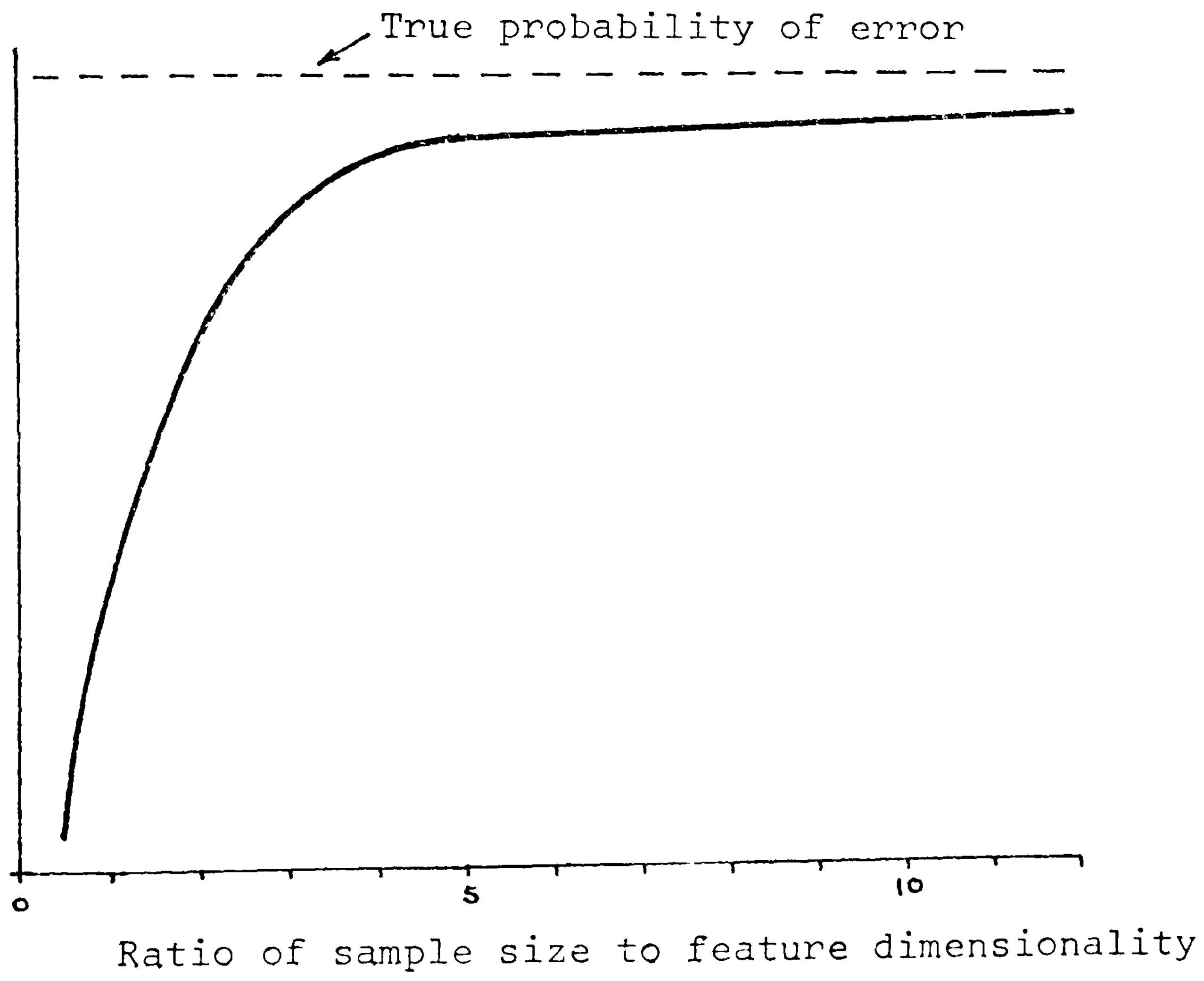


Figure 7.6
Foley plot - typical curve of average probability of error as a function of the ratio n/h

As the ratio increases, the estimated probability of error calculated from the set of given vectors approaches the true probability of error. For ratios less than three, the difference between the estimated and the true performance is noticeably large. However the error curve appears to level off for ratios greater than three. This indicates that a critical value of the ratio n/h exists. In typical architectural applications this critical ratio is not usually of importance as activities usually outnumber variables measured.

7.8 SUMMARY

In this chapter several techniques which are useful for producing a low-dimensional representation of multivariate data have been discussed. The main interest has been explicitly in the two-dimensional solution given by the methods, since the main aim has been to be able to examine the data visually. It should be mentioned, however, that for some data sets it would be unrealistic to expect a two-dimensional representation to give anything but a very approximate indication of the inherent structure present. In other words, two dimensions may just not be sufficient to accommodate the full compexity of the relations in the given data set. Unfortunately no real test exists for the value of the number of dimensions necessary to provide an adequate fit, apart from the informal goodness of fit criteria mentioned in connection with particular methods. However, Gnanadesikan and Wilk (1969) make the following important point which perhaps suggests that a formal test of the number of dimensions is not important:

> Interpretability and simplicity are important in data analysis and any rigid inference of optimal dimensionality, in the light of the observed values of a numerical index of goodness of fit, may not be productive.

Two dimensional solutions certainly have the virtue of simplicity; they are also readily understood by the program user and may, in many cases, provide the basis for the understanding of the overall relationships in the data set; consequently they are likely to be the solutions of the most practical value.

Finally, it should be mentioned again that the techniques described in this chapter should in no way be regarded as methods to be used to the exclusions of the other types of analysis; indeed they will, in general, be most useful when used in conjunction with other forms of analysis.

CHAPTER 8

NONLINEAR MAPPING

8.1 INTRODUCTION

Mapping algorithms all have the basic characteristics of an iterative search for an optimal solution to a given problem. Iterative search techniques usually start off with an arbitary guess of the solution which is then improved upon repeatedly through a systematic mechanism until a satisfactory final solution is obtained This transition from initial guess to final solution implicitly defines the mapping algorithm for a given set of data

The mappings defined in this manner differ from noniterative mappings in three respects. First no specific a priori knowledge, such as the statistical characterisation of the data is used in defining the mapping. Second all iterative algorithms must be provided with a suitable termination criterion which determines when a satisfactory solution has been achieved Third the mapping. which has been iteratively obtained for a set of data. can only apply to that data set When new data is introduced a new mapping must be computed

8.2 NONLINEAR MAPPING ALGORITHMS

The objective of these algorithms is to reduce the dimensionality of the data in the h-space to the d-space so that some inherent "structure" of the data may be displayed and detected Here the word structure refers to the geometrical relationships that may exist among the subsets of data and in particular those relationships that reveal clusters The display space may conveniently be in two or three dimensional space but is here discussed in terms of two-dimensional space.

Given a data base of N activities, each described by h variables the data base may then be represented by a set of N h-vectors

$$X = (X_i) \quad i = 1, 2, \ldots, N$$

where

$$X_i = (x_{i1}, x_{i2}, \ldots x_{ih}), \quad h > 2$$

Chapter 7 presented a number of methods of interpreting this high dimensional data. This section concentrates on one particular approach nonlinear mapping (NLM). The mapping may be defined as

$$\text{NLM}: X \rightarrow Y$$

where Y is a collection of N two-dimensional vectors in the d-space.

$$Y = (Y_i) \quad i = 1 \;\; 2, \ldots, N$$

where

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix}$$

A nonlinear mapping involves the reduction of dimensionality of the activity data in X from h dimensions to two dimensions by means other than linear transformations whilst attempting to preserve as much of the inherent structure as possible This structure preservation is achieved by fitting N h-dimensional vectors in the d-space such that their intervector distances or dissimilarities approximate the corresponding intervector distances of dissimilarity in the h-space Let the distance or dissimilarity between the vectors X_i and X_j in the h-space be denoted by

$$\hat{d}_{ij} = \text{dis}(X_i, X_j)$$

and the distance or dissimilarity between Y_i and Y_j in the d-space be

$$d_{ij} = \mathrm{dis}(Y_i, Y_j)$$

Then the structure of the data is strictly preserved under the mapping NLM: X → Y if for all i and j, $\hat{d}_{ij} = d_{ij}$. Obviously for all but the most trivial cases this strict preservation is impossible to achieve. It is possible, however, to achieve various kinds of approximate preservation without much difficulty, for example, to preserve certain parts of the structure that exists among the data base by requiring

$$\hat{d}_{ij} = d_{ij} \quad \text{or} \quad \hat{d}_{ij} \approx d_{ij}$$

for those X_i and X_j such that $\hat{d}_{ij} < \theta$ where θ is some threshold value and not seeking such faithful preservation for those X_i and X_j with $\hat{d}_{ij} > \theta$. The consequence of this kind of approximate preservation is the introduction of an error e_{ij} where

$$e_{ij} = \hat{d}_{ij} - d_{ij}$$

for some or perhaps all values of i and j.

All NLM algorithms must deal with the problem of how approximate preservation of the data structure may be best achieved. There are two interrelated questions to be considered:

(1) What distance or dissimilarity measure should be used in order to describe the geometric relationship between the N vectors in the h-space and in the d-space

(2) The choice of an error function

$$E = f(e_{ij}) = f(\hat{d}_{ij} - d_{ij})$$

such that the value of this function will reflect the degree of structure preservation, strict or approximate, in a monotonic fashion, i.e. the smaller the value of E, the better the preservation.

Clearly the way these two considerations are dealt with affects directly how well a particular NLM algorithm works in mapping the data to the two-dimensional space for visualisation Such considerations also in part characterise the various NLM algorithms.

All the NLM algorithms discussed here employ an iterative technique. The basic elements of these iterative algorithms consist of a three-step procedure:

(1) Determine an initial set of Y vectors. This set is referred to as the initial configuration of the d-space

It can be selected by random selection.

(2) Adjust the Y's of the current configuration starting with the initial configuration, in such a way that the next configuration (the set of adjusted Y's) will have a smaller value of the error function. The transition from the current configuration to the next configuration is an iteration.

(3) Repeat (2) until one of the termination criteria is met:

a - the error function E has reached a prespecified value.

b - a prespecified number of iterations have been performed.

The various algorithms described below differ primarily in one or more of the following aspects:

1. The selection of the distance measures $\hat{d}_{ij}$ and d_{ij}

2. The selection of an error function E.

3. The method of termination.

8.3 SAMMON'S NLM ALGORITHM

This method was developed by Sammon (1969) and is the method implemented in MAGIC. Let the distance or dissimilarity measure between the vectors X_i and X_j in the h-space be the Euclidean metric

$$\hat{d}_{ij} = \left(\sum_{k=1}^{h}(x_{ik} - x_{jk})^2\right)^{1/2}$$

Similarly let the distance or dissimilarity measure between Y_i and Y_j in the d-space be the Euclidean metric

$$d_{ij} = \left(\sum_{k=1}^{d}(y_{ik} - y_{jk})^2\right)^{1/2}$$

The error function E, which represents how well the present configuration of the N vectors in the d-space fits the N vectors in the h-space, is defined as

$$E = f(\hat{d}_{ij} - d_{ij})$$
$$= \frac{1}{\sum_{i<j}^{N}\hat{d}_{ij}} \sum \frac{(\hat{d}_{ij} - d_{ij})^2}{\hat{d}_{ij}}$$

where $\sum_{i<j}^{N}$ denotes the sum over all i and j such that $i < j$. Sammon's algorithm then works as follows Generate a random set of Y_i's in the d-space. This set of vectors is the initial configuration of the d-space Next compute all the d-space intervector distances d_{ij} which are then used to determine the value of the error function E. Then adjust the N vectors in the d-space so as to decrease the error function and continue to make these adjustments until the minimum value of E is reached or until a prespecified small value has been obtained. The set of Y_i's at the termination of the adjusting process is the final configuration.

It should be noted that the error function E is a function of 2N independent variables y_{ij}, i = 1,2, . . . N and j = 1,2. In Sammon's algorithm these 2N variables must be adjusted simultaneously to yield a new configuration This is achieved by carrying out a steepest descent procedure to search for the minimum of the error function If the current configuration, that is, the set of Y_i's being adjusted, is denoted by

$$Y_1 = \begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix}, \quad Y_2 = \begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix}, \ldots\ldots, \quad Y_N = \begin{bmatrix} y_{N1} \\ y_{N2} \end{bmatrix}$$

The method of steepest descent consists of successively computing the new vectors

$$Y_1' = \begin{bmatrix} y_{11}' \\ y_{12}' \end{bmatrix}, \quad Y_2' = \begin{bmatrix} y_{21}' \\ y_{22}' \end{bmatrix}, \ldots\ldots, \quad Y_N' = \begin{bmatrix} y_{N1}' \\ y_{N2}' \end{bmatrix}$$

governed by the following recursive relation

$$Y_P' = Y_P - \propto \frac{\delta E}{\delta Y_P} \Bigg/ \left| \frac{\delta^2 E}{\delta Y_P^2} \right| \qquad P = 1,2, \ldots. N$$

where $\propto$ is a correction factor and E is the error corresponding to the present configuration. The notation $\delta/\delta Y_p$ denotes taking the first partial derivatives with respect to each component of Y_p and arranging the partial derivatives in a column vector The set of N vectors Y'_p, $p = 1, \ldots, N$, becomes the new or the next configuration in the d-space. Such a transition from the current to the next configuration defines an iteration of Sammon's NLM algorithm.

At each iteration it is necessary to calculate the first and second partial derivatives of the error function with respect to Y_p. For the first derivative,

$$\frac{\delta E}{\delta Y} = \frac{1}{c} \sum_{j=1}^{N} \frac{\delta}{\delta Y_p} \frac{(\hat{d}_{pj} - d_{pj})^2}{\hat{d}_{pj}}$$

where $c = \sum_{i<j}^{N} \hat{d}_{ij}$ is a constant.

Sammon (1969) shows that this reduces to

$$\frac{\delta E}{\delta Y_p} = -\frac{2}{c} \sum_{j=1}^{N} \left(\frac{\hat{d}_{pj} - d_{pj}}{\hat{d}_{pj} d_{pj}} \right) (Y_p - Y_j)$$

$p = 1 \quad 2, \ldots, N$

and, similarly for the second derivative

$$\frac{\delta^2 E}{\delta Y_p^2} = -\frac{2}{c} \sum_{j=1}^{N} \frac{1}{\hat{d}_{pj} d_{pj}} \left[(\hat{d}_{pj} - d_{pj}) - \frac{(Y_p - Y_j)^2}{d_{pj}} \left(1 + \frac{\hat{d}_{pj} - d_{pj}}{d_{pj}} \right) \right]$$

$p = 1, 2, \ldots, N$

An example of the results from Sammon's algorithm is shown in figure 8.1. This shows the two-space representation of the artificial data set shown in figure 8.2. The initial configuration was random, but all of the embedded clusters are easily identifiable in the final display.

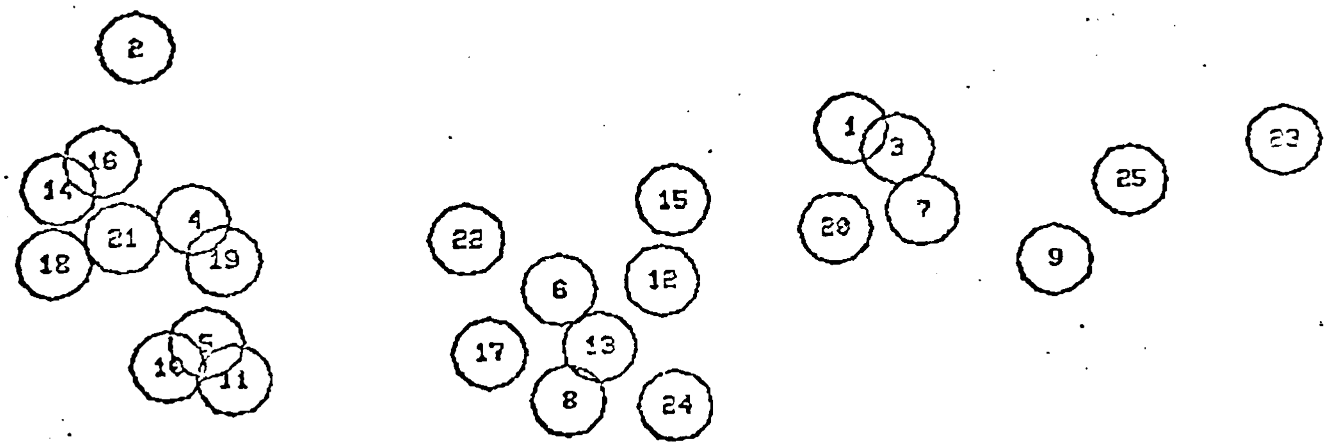


Figure 8.1

Nonlinear mapping plot

ORIGINAL DATA

	1	2	3	4	5
1	48.00	1.00	7.00	1.00	3.00
2	16.00	1.00	5.00	2.00	2.00
3	48.00	1.00	6.00	1.00	3.00
4	24.00	1.00	6.00	1.00	3.00
5	25.00	1.00	2.00	1.00	3.00
6	36.00	1.00	5.00	1.00	5.00
7	48.00	2.00	5.00	3.00	3.00
8	36.00	2.00	4.00	5.00	3.00
9	52.00	2.00	2.00	1.00	3.00
10	24.00	2.00	1.00	1.00	3.00
11	26.00	2.00	1.00	1.00	3.00
12	39.00	2.00	6.00	1.00	3.00
13	37.00	2.00	3.00	2.00	5.00
14	20.00	2.00	6.00	1.00	3.00
15	40.00	2.00	7.00	5.00	3.00
16	21.00	2.00	7.00	2.00	1.00
17	34.00	1.00	3.00	1.00	3.00
18	20.00	2.00	3.00	2.00	2.00
19	25.00	2.00	5.00	3.00	3.00
20	45.00	1.00	6.00	1.00	3.00
21	22.00	1.00	5.00	3.00	2.00
22	33.00	2.00	7.00	1.00	3.00
23	68.00	2.00	6.00	1.00	3.00
24	39.00	1.00	1.00	3.00	3.00
25	55.00	1.00	4.00	1.00	3.00

Figure 8.2

Data set used for figure 8.1

8.4 COMPUTATIONAL ASPECTS

It is of interest to consider the computational aspects of Sammon's algorithm as it is here that most variations have been proposed. First, the algorithm requires the computation and storage of the intervector distances $\hat{d}_{ij}$ for all $i < j$. There are $N(N-1)/2$ such distances. These of course need not be computed for each iteration of the algorithm, it being possible to compute them once and store them for use at each iteration At each iteration, however, all the $N(N-1)/2$ distances d_{ij}, $i < j$, along with the error derivatives, must be computed Thus the overall computational requirement is proportional to $N(N-1)/2$. As the number of vectors increases, the computational requirement (time and storage) grows quadratically

As a means of reducing the computational requirements a variation of Sammon s algorithm was developed by White (1972). Rather than the Euclidean metric, the Hamming metric is used as a distance measure between vectors. In the h-space the Hamming metric between X_i and X_j is defined as

$$\hat{d}_{ij} = \sum_{k=1}^{h} |x_{ik} - x_{jk}|$$

Similarly, the Hamming distance between Y_i and Y_j in the d-space is

$$d_{ij} = \sum_{k=1}^{d} |y_{ik} - y_{jk}|$$

The Hamming metric provides savings in the computational requirements in two ways. First, the Hamming metrc is much simpler to compute than the Euclidean metric. Second the error derivatives as required in each iteration are also simpler to compute However, the use of the Hamming metric as a distance measure has its flaws. If the data in the h-space is known to have a Euclidean structure (i.e the vectors satisfy the conditions of a Euclidean metric), some distortion of the Y_i vectors in the d-space will inevitably occur.

Another problem with the use of the Hamming metric lies in the fact that interpretation of the resulting d-space configuration may be more difficult. With the Hamming metric the usual notion of the Euclidean distance in two dimensions no longer exists. Instead of measuring the length of a line segment joining two vectors, a complicated sum of absolute values would have to be "visualised". The conclusion must therefore be that this "improvement" is of little use in MAGIC.

A different approach has been proposed by Chang and Lee (1973). Like Sammon's algorithm the Chang-Lee relaxation method also seeks to preserve the inherent structure that may exist in the data. The term "relaxation" is borrowed from the relaxation method for solving linear equations. Unlike Sammon's algorithm the minimisation of the error function is carried out by minimising one term of the function at a time.

The basic procedure of the relaxation method as developed by Chang and Lee is very similar to that of Sammon's method but there are two significant differences. First, in the relaxation method the squared Euclidean metric is used as the distance measure, and secondly, the method of adjusting the current Y_i's is different. In Sammon's method all the Y_i's are adjusted simultaneously along the direction of steepest descent so as to reduce the value of the error function E. The idea of the relaxation method is to adjust the Y_i's on a pairwise basis.

Insofar as the computational requirements are concerned, since there are 0.5N(N-1) pairs of vectors to be adjusted in each sequence, each iteration will take N(N-1)

N adjustments. Compared with the N adjustments required in Sammon's algorithm this is a considerable overhead, especially when N is large. However, the computation of the error function is much simpler and does not require the summation of 0.5N(N-1) terms as does Sammon's method.

Several further variations of the relaxation method have been developed to improve the computational requirements. One variation involves the use of heuristics in performing the pairwise adjustments when N is large. This heuristic method is known as the frame algorithm (Chang and Lee 1973). The frame algorithm does not preserve the structure relationship as faithfully as the relaxation method; Chang and Lee give details of experiments comparing the two approaches.

8.5 RELATIONSHIP OF NLM TO OTHER ORDINATIONS

The relationship of nonlinear mapping to nonmetric multidimensional scaling was mentioned in chapter 7. The multidimensional scaling algorithm developed by Shepard (1962) and later improved by Kruskal (1964a,b) seeks to find a configuration of points in a d-space such that the resultant interpoint distances preserve a monotonic relationship to a given set of interelement dissimilarities. Specifically, they wish to analyse a set of interelement dissimilarities given by S_{ij}, i = 1, ..., N, j=1, ..., N. Suppose these dissimilarities are ordered in increasing magnitude, such that

$$S_{p1q1} \leqslant S_{p2q2} \leqslant \ldots\ldots \leqslant S_{pnqn}$$

The Kruskal Shepard algorithm seeks to find a set of N d-dimensional vectors y_i, i = 1, ..., N, such that the order of the interpoint distances d_{ij}=dis[y_i ,y_j] deviates as little as possible from the monotonic ordering of the corresponding dissimilarities. Despite the mathematical formulation of nonlinear mapping being similar, the underlying criteria are quite different, although Kruskal (1971) has shown how his M-D-SCAL program may be modified to produce Sammon's nonlinear mapping.

The nonlinear mapping is preferred here as:

(1) The routine does not depend upon any control parameters that would require a priori knowledge about the data. The only requirements are that the limiting number of iterations and the convergence constant must be set. Both of these values are defaulted in MAGIC.

(2) Nonlinear mapping is highly efficient in identifying complex data structures. Sammon (1968) gives examples.

(3) The resulting mappings are easily evaluated.

(4) The algorithm is simple and efficient.

CHAPTER 9

PRINCIPAL CO-ORDINATES ANALYSIS

9.1 INTRODUCTION

The techniques of classical scaling and principal components analysis were introduced in chapter 7. The limitations of these techniques were noted and particular reference made to the implied distance measures involved in the low dimensional representations and their nonsensical physical dimensions when different variates were measured on different scales. More formally, if the activity data is regarded as defining a set of N points in the high dimensional h-space, these N points can be represented by an h x N matrix X

$$X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{1h} & x_{2h} & \cdots & x_{Nh} \end{bmatrix}$$

and each column thus represents a data point x_i with coordinates $(x_{i1}, x_{i2}, \ldots, x_{ih})$ referred to rectangular axes. Thus the implied distance d_{ij} between x_i and x_j is given by

$$d_{ij}^2 = \sum_{r=1}^{h}(x_{ir} - x_{jr})^2$$

and the spatial configuration of the low dimensional display is of interest only if d_{ij} satisfactorily measures the similarity between x_i and x_j. This usefulness has the obvious defect of depending in a complex manner on the scales of measurement of the different variates. As a solution to this problem Gower (1966) proposed the method of Principal Co ordinates Analysis, where the assumptions are similar, but dissimilarities take the place of distances, and the dissimilarities may be derived from any of the types of variable as described in chapter 4.

9.2 PRINCIPAL CO-ORDINATES ANALYSIS

Let A be a symmetric (nxn) matrix with latent roots λ_1, λ_2, ... λ_n, and associated (nxl) latent vectors c_1, c_2, ..., c_n as shown in figure 9.1.

Root

$$\begin{array}{ll|llll}
 & & \lambda_1 & \lambda_2 & & \lambda_n \\
\hline
 & Q_1 & c_{11} & c_{12} & \cdots & c_{1n} \\
\text{Point} & Q_2 & c_{21} & c_{22} & \cdots & c_{2n} \\
 & \cdot & \cdot & \cdot & \cdots & \cdot \\
 & \cdot & \cdot & \cdot & \cdots & \cdot \\
 & Q_n & c_{n1} & c_{n2} & \cdots & c_{nn}
\end{array}$$

(elements of c_i are c_{1i} , c_{2i} , ..., c_{ni})

Figure 9 1

Latent roots and vectors of symmetric matrix A

Suppose now the elements of the ith row of figure 9.1 are taken as the coordinates of a point Q_i in n-dimensional space. The Euclidean distance, Δ_{ij} , between points Q_i and Q_j in this space is given by

$$\Delta^2_{ij} = \sum_{r=1}^{n} (c_{ir} - c_{jr})^2$$
$$= \sum_{r=1}^{n} c^2_{ir} + \sum_{r=1}^{n} c^2_{jr} - 2\sum_{r=1}^{n} c_{ir}\, c_{jr}$$

If the latent vectors are normalised so that the sums of squares of their elements are equal to their corresponding latent roots, i.e. so that

$$\sum_{i=1}^{n} c_{ir}^2 = \lambda_r$$

then

$$A = c_1c_1' + c_2c_2' + \ldots + c_nc_n'$$

and therefore

$$a_{ii} = \sum_{r=1}^{n} c_{ir}^2 \quad \text{and } a_{ij} = \sum_{r=1}^{n} c_{ir}\, c_{jr}$$

and, substituting these results in the distance equation

$$\Delta_{ij}^2 = a_{ii} + a_{jj} - 2a_{ij}$$

Suppose now that the matrix A had elements $a_{ij} = -0.5d_{ij}^2$ and $a_{ii} = 0$, where d_{ij} is some measure of inter-element distance. From the above equation it may be seen that Δ_{ij} is now simply equal to d_{ij}, and consequently the above procedure gives a method of finding coordinates for a set of points given their interpoint distances d_{ij}. In particular if d_{ij} was Euclidean distance the method is directly analogous to principal components analysis. However, the advantage of this method is that it may be used to find a set of coordinates for observations where the d_{ij} s are not considered to be Euclidean.

If A was a similarity matrix so that elements a_{ii} were unity then

$$\Delta_{ij}^2 = 2(1 - a_{ij})$$

and principal coordinates analysis would lead to a spatial representation of the similarities in which Δ_{ij} functions as Euclidean distance, although not all similarity measures would be suitable.

It has now been established that given a symmetric matrix A with elements a_{ij}, a set of coordinates may be found in the n-dimensional space such that the Euclidean distance between the points in this space is given by Δ_{ij}, and that this procedure may be used to find the coordinates of a set of observations given their interpoint distances (not necessarily Euclidean), or their similarities. Gower shows further that it is legitimate to use principal components on these coordinates to find the best fit in fewer dimensions. The whole process therefore involves two stages, each stage requiring the determination of the latent roots and vectors of an nxn matrix. That is, at stage one the matrix A, and at stage two the nxn matrix of n-dimensional coordinates resulting from the first stage. Gower shows that these two stages may be collapsed into one as follows

(1) Calculate the matrix A. In the case of similarities, A is simply the inter-element similarity matrix; with distance measures, A may be formed by

taking $a_{ii} = 0$ and $a_{ij} = -0.5d^2_{ij}$. Gower (1966) also shows that a convenient representation of the n activities in Euclidean space can be obtained from transforming distances using $a_{ij} = (2(1-d_{ij}))^{\frac{1}{2}}$. This is the transformation used in MAGIC.

(2) Transform this to a matrix α , the elements of which are given by

$$\alpha_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$

where

$$\bar{a}_{i.} = 1/n \sum_{j=1}^{n} a_{ij}$$

$$\bar{a}_{.j} = 1/n \sum_{i=1}^{n} a_{ij}$$

$$\bar{a}_{..} = 1/n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$$

(3) Find the latent roots and vectors of α , scaling each vector so that the sum of squares of its elements is equal to its corresponding latent root.

The elements of the k-th latent vector now give the coordinates of the n points on the k-th principal axis. The first two coordinates may now be used to obtain a visual representation of the original distance or similarity matrix. A measure of the adequacy of fit of, say, the first p principal coordinates is given by

$$T = \sum_{i=1}^{p*} \gamma_i / \mathrm{trace}\ (\alpha)$$

where the γ_i are the latent roots of the matrix arranged in descending order of magnitude.

Gower shows that this method may be used only with similarity measures which give rise to an α matrix with no negative latent roots; i.e. α must be positive semi-definite. Gower (1971a) also shows that this condition holds for a wide class of measures, but in particular that Gower's general coefficient of similarity is always positive semi-definite.

Principal coordinates analysis may thus be seen to have a considerable advantage over principal components analysis when seeking a visual representation of data. It operates directly on similarity and distance matrices and is not restricted to Euclidean distances.

9.3 EXAMPLE OF PRINCIPAL COORDINATES ANALYSIS

Figures 9.2 and 9.3 show the two dimensional and three dimensional ordinations of the test data set.

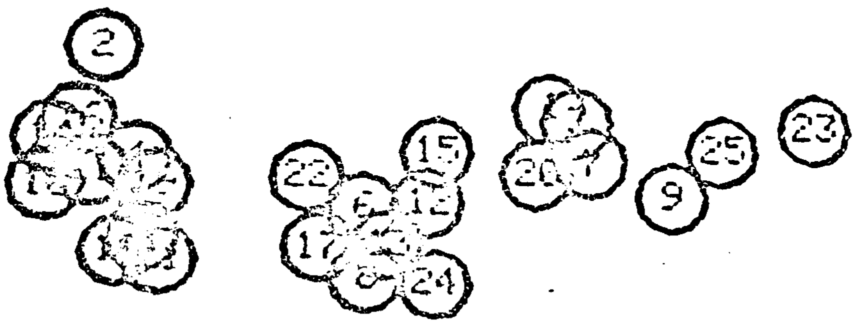


Figure 9.2

Two-dimensional Principal Co-ordinates plot

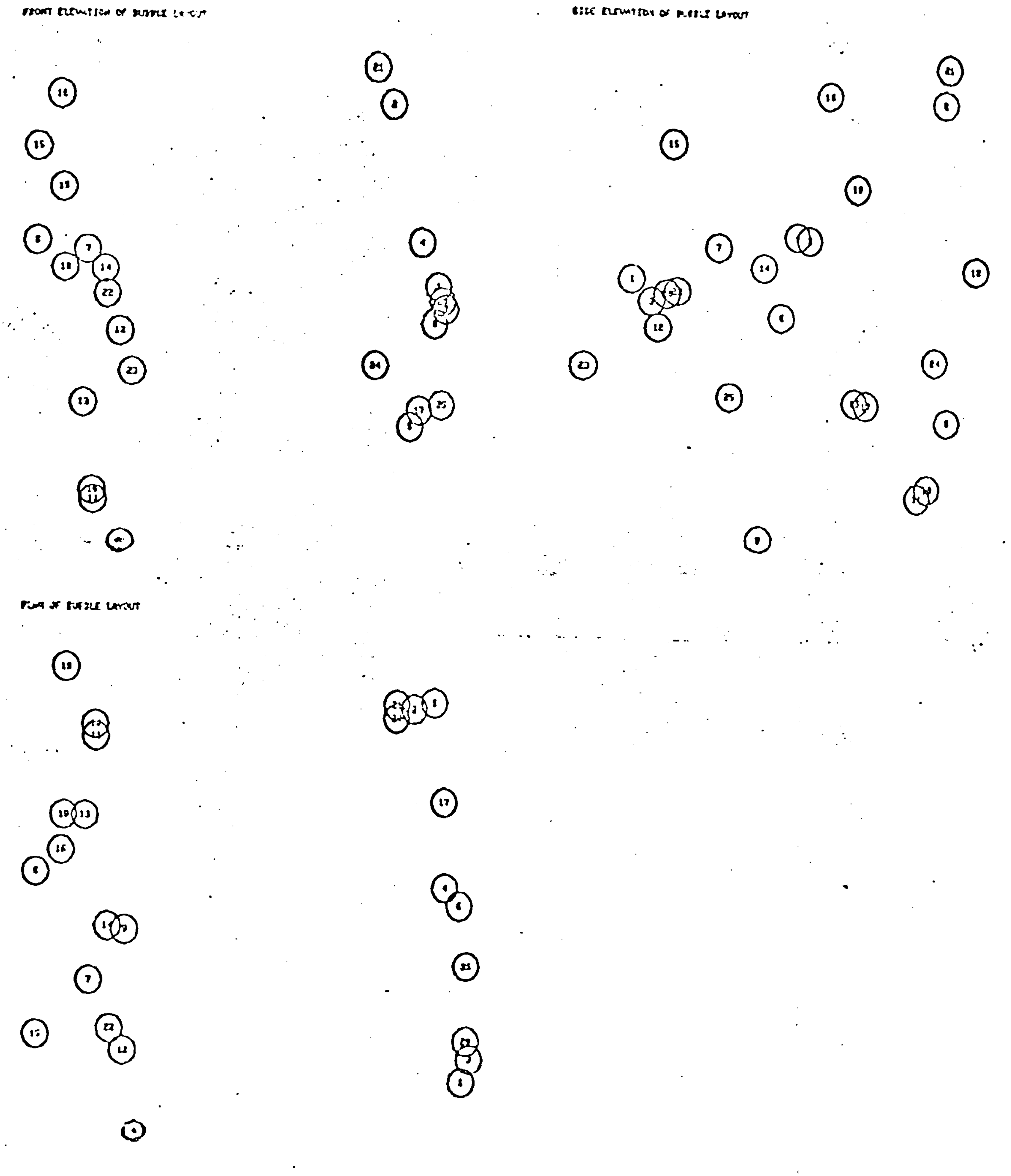


Figure 9.3

Three-dimensional Principal Co-ordinates plot

CHAPTER 10

DISPLAY OF RESULTS

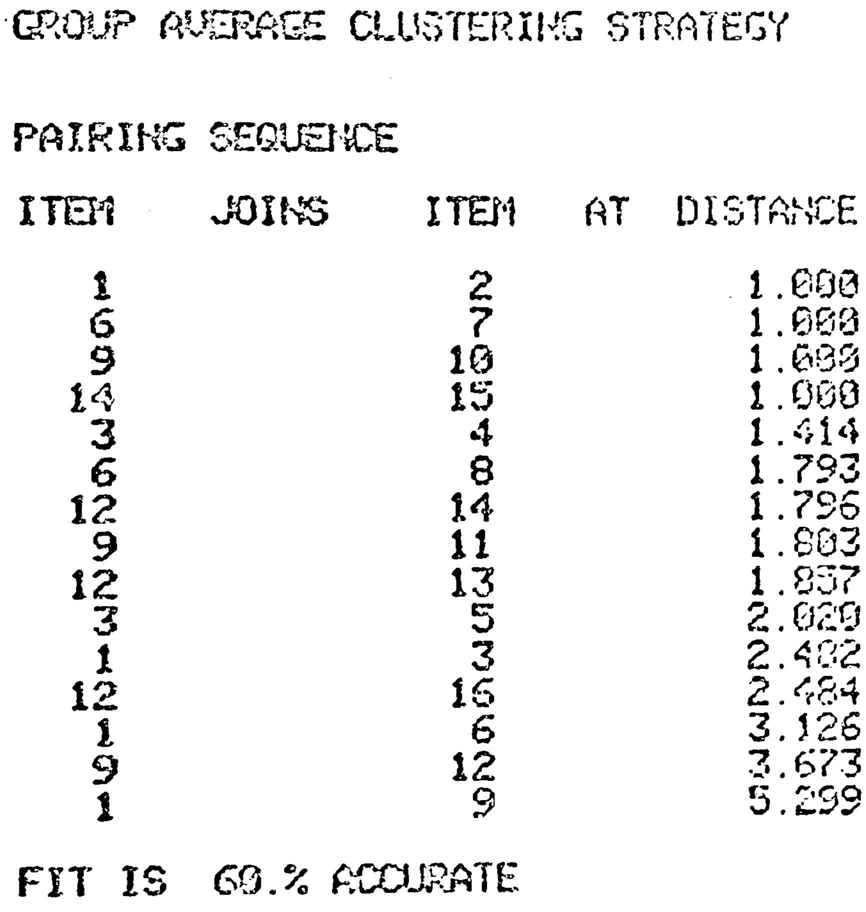
10.1 INTRODUCTION

MAGIC has been designed to operate interactively and to present all results graphically as an aid to interpretation of the data. A number of important techniques have been devloped to enable this - an efficient dendrogram plotting routine, a unique presentation method to display the results of the Euclidean cluster analysis, a number of options for the manipulation of two-dimensional ordinations, and a graphical method of comparing results of different ordinations. This chapter describes these techniques in the same order as the analytical results they are designed to display have been presented.

10.2 DENDROGRAM PLOTS

The results of hierarchical cluster analysis are usually displayed in the form of dendrograms (Sneath and Sokal, 1973, give a general account), but, surprisingly, no efficient algorithm for the automatic display of these diagrams had been developed.

The general form of such diagrams is for the activity labels to be plotted across the top of the page, and vertical lines drawn down to the successive clustering levels where a horizontal line joins those activities clustering at that level. These fusions at successive hierarchical levels are printed in MAGIC in a linkage order table. An example is shown in figure 10.1, and the resulting dendrogram in figure 10.2.



```
GROUP AVERAGE CLUSTERING STRATEGY

PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT  DISTANCE

   1                  2           1.000
   6                  7           1.000
   9                 10           1.000
  14                 15           1.000
   3                  4           1.414
   6                  8           1.793
  12                 14           1.796
   9                 11           1.803
  12                 13           1.857
   3                  5           2.020
   1                  3           2.402
  12                 16           2.484
   1                  6           3.126
   9                 12           3.673
   1                  9           5.299

FIT IS  60.% ACCURATE
```

Figure 10.1

Linkage order table

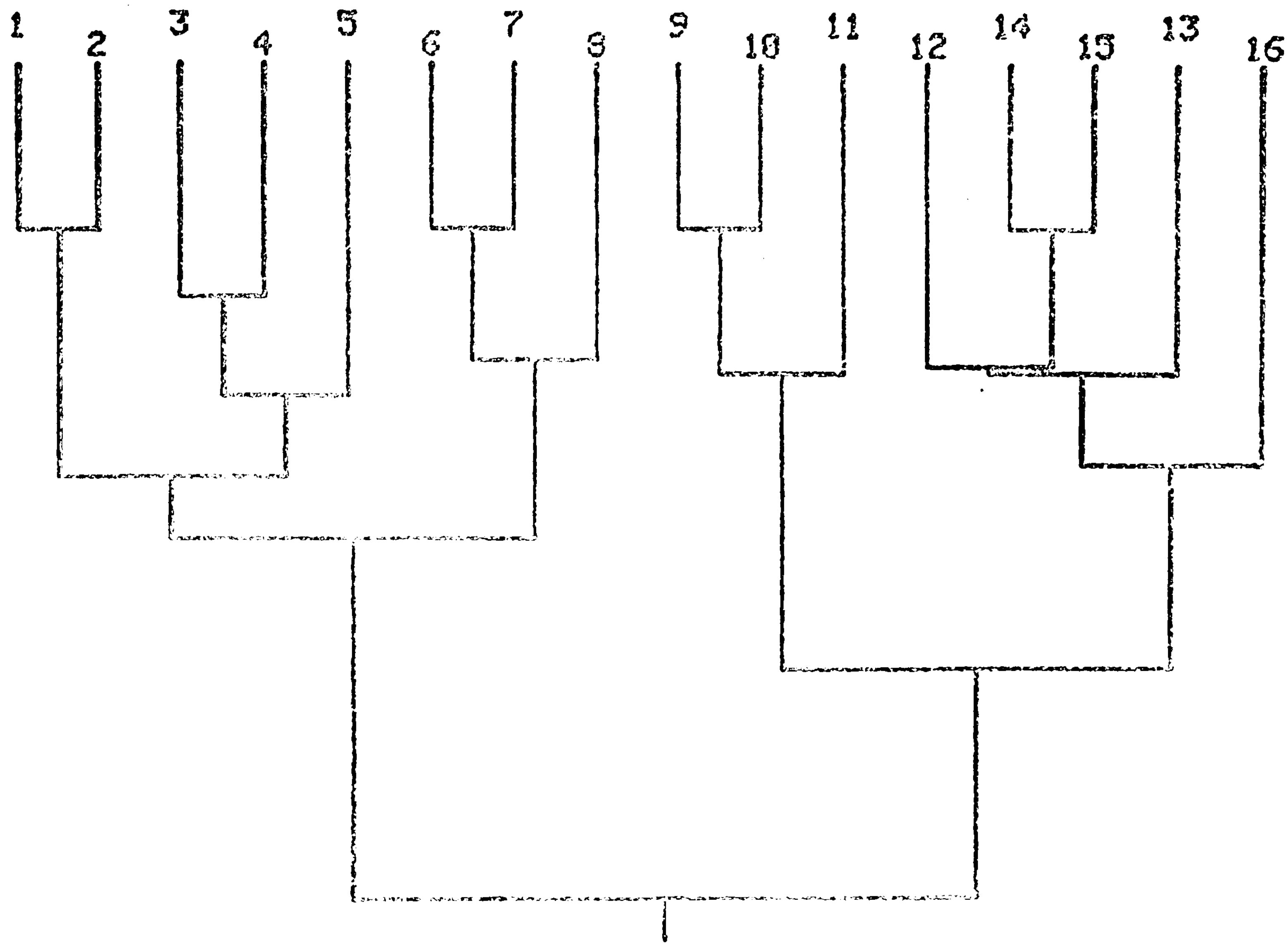


Figure 10.2

Dendrogram of data in figure 10.1

This structure may be defined by two lists - one containing the labels of the n objects in the order in which they are to appear across the top of the dendrogram (an order which must be calculated to prevent crossing lines occuring in the dendrogram), and the other containing the n-1 numbers showing the level at which the successive activities join together in the tree. This second list is conveniently accessed from a further "linkage order" list, containing the first two columns of figure 10.1.

The algorithm to draw the dendrogram is simple. Given the two lists as described above, LAB for the activity labels, LEV for the clustering levels, the linkage order lists, L1 and L2, such that L1 joins L2 at LEV, and assuming screen scaling is carried out elsewhere, there are four basic steps:

(1) Plot a vertical line for each of the i=1, 2, ..., n activities from below each label (the coordinates of which may be defined as (i, y_{max})) to a level (i, $\{\max LEV_i, LEV_{i-1}\}$). Store the bottom coordinates of each line in two arrays XC and YC.

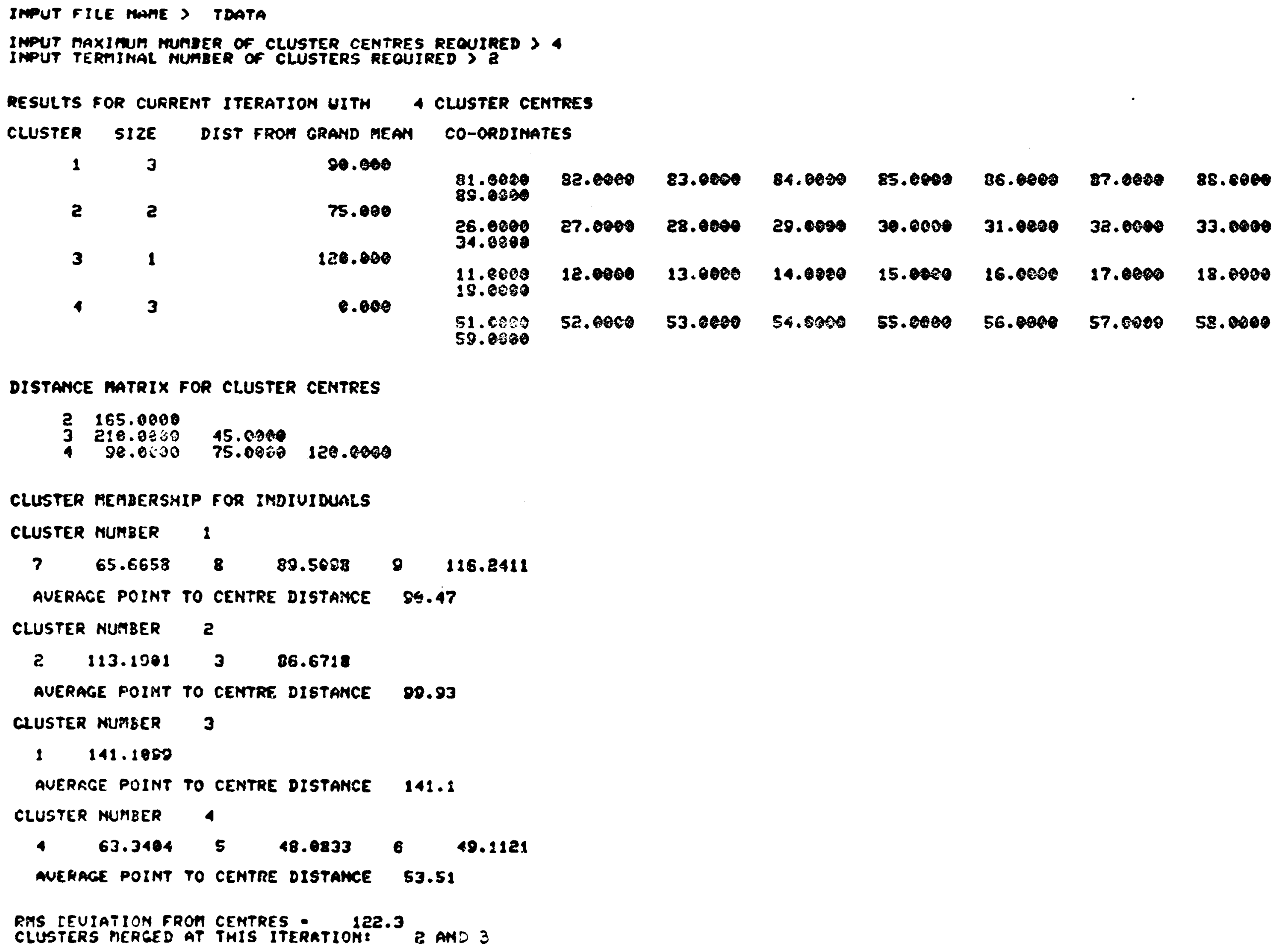
(2) From L1 and L2 find the first $L1_i$ (i = 1, 2, ..., n-1) for which $LEV_{L1_i} > LEV_{L1_{i+1}}$.

(3) Plot a horizontal line from L1 (XC_i, YC_i) to L2 (XC_i, YC_i). Plot a vertical line from the centre of this line down to $\max\{LEV_i, LEV_{i+1}\}$. Store the coordinates of the bottom of this line in XC_{i+1} and YC_{i+1}. This step represents two clusters being merged, so the ith entries in LEV, XC and YC may be deleted and the pointers revised.

(4) Set n=n-1. If $n > 1$ go to step (2), else draw final tail and finish.

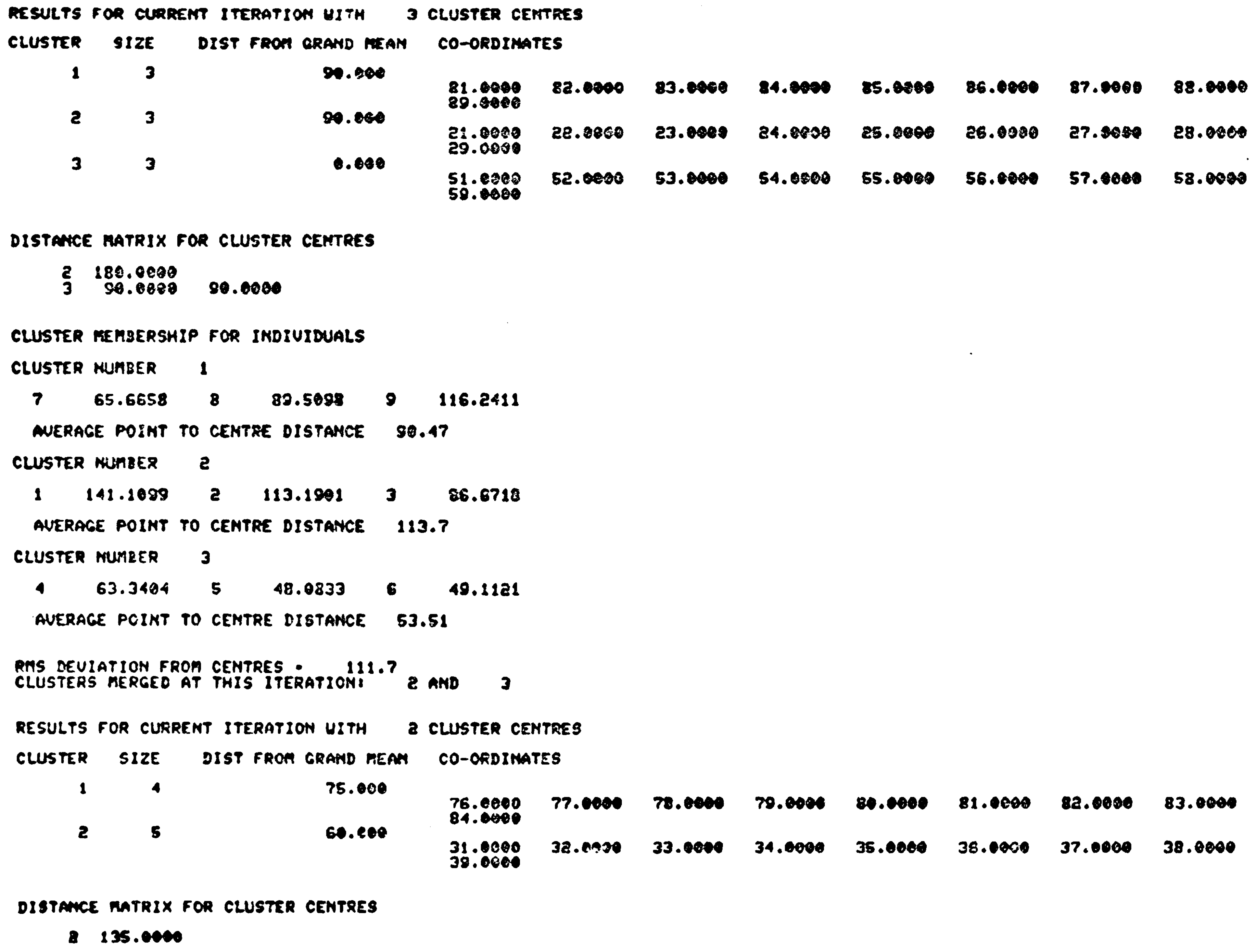
10.3 DISPLAY OF EUCLIDEAN CLUSTER ANALYSIS

Euclidean cluster analyses have never been presented graphically - figure 10.3 shows the typical form of program output. This information is difficult to fully comprehend even for experts used to multivariate analysis. It is possible however to identify those elements which are of importance to the user of MAGIC. These are the cluster membership at each level of clustering, the cluster density (whether the cluster is a compact group of activities, or only loosely connected), and the relative disposition of the clusters one to another. The cluster membership is easily displayed in tabular form, the cluster density may be obtained from the average point to centre distance of each cluster and the cluster disposition is contained in the matrix of cluster centre to centre distances. As this matrix is a simple distance matrix it is possible to pass it through the nonlinear mapping analysis to obtain a convenient two-dimensional representation. If the bubbles of this display are then scaled according to the average point to centre distance applicable to each cluster all of the essential data may be displayed in the form shown in figure 10.4.



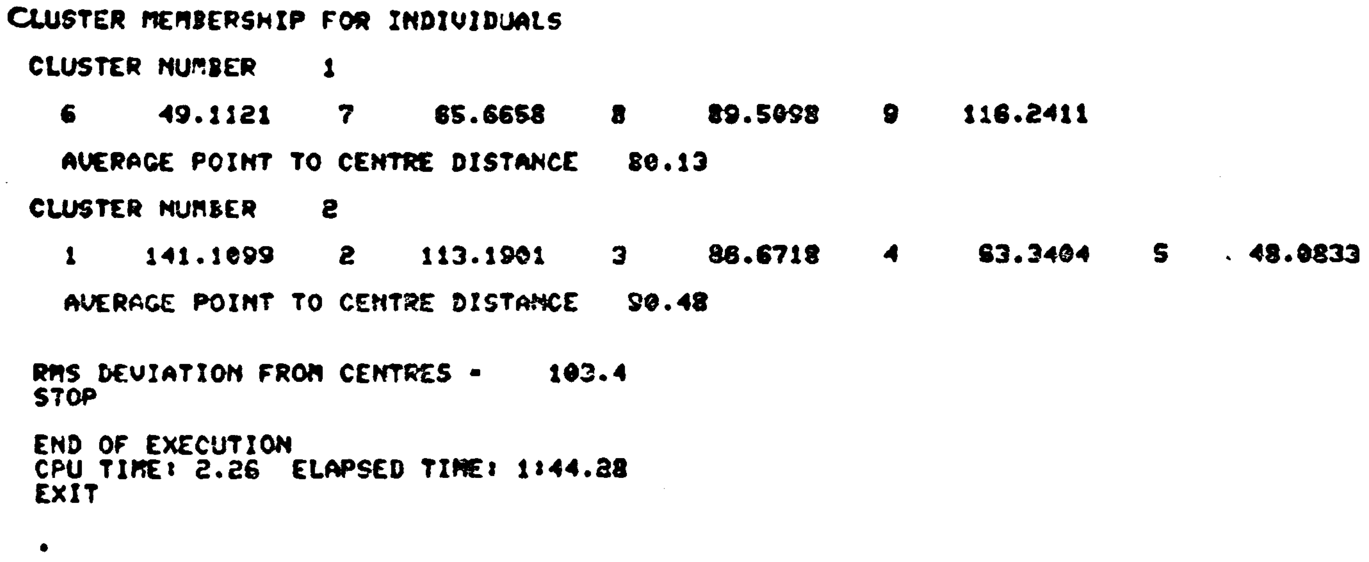
```
INPUT FILE NAME >  TDATA

INPUT MAXIMUM NUMBER OF CLUSTER CENTRES REQUIRED > 4
INPUT TERMINAL NUMBER OF CLUSTERS REQUIRED > 2

RESULTS FOR CURRENT ITERATION WITH    4 CLUSTER CENTRES

CLUSTER    SIZE    DIST FROM GRAND MEAN   CO-ORDINATES

      1       3                 90.000
                                           81.0000   82.0000   83.0000   84.0000   85.0000   86.0000   87.0000   88.0000
                                           89.0000
      2       2                 75.000
                                           26.0000   27.0000   28.0000   29.0000   30.0000   31.0000   32.0000   33.0000
                                           34.0000
      3       1                120.000
                                           11.0000   12.0000   13.0000   14.0000   15.0000   16.0000   17.0000   18.0000
                                           19.0000
      4       3                  0.000
                                           51.0000   52.0000   53.0000   54.0000   55.0000   56.0000   57.0000   58.0000
                                           59.0000


DISTANCE MATRIX FOR CLUSTER CENTRES

     2  165.0000
     3  210.0000   45.0000
     4   90.0000   75.0000  120.0000

CLUSTER MEMBERSHIP FOR INDIVIDUALS

CLUSTER NUMBER    1

  7     65.6658    8     89.5098    9    116.2411

  AVERAGE POINT TO CENTRE DISTANCE   90.47

CLUSTER NUMBER    2

  2   113.1901    3     86.6718

  AVERAGE POINT TO CENTRE DISTANCE   99.93

CLUSTER NUMBER    3

  1   141.1099

  AVERAGE POINT TO CENTRE DISTANCE   141.1

CLUSTER NUMBER    4

  4    63.3404    5     48.0833    6     49.1121

  AVERAGE POINT TO CENTRE DISTANCE   53.51

RMS DEVIATION FROM CENTRES =    122.3
CLUSTERS MERGED AT THIS ITERATION:    2 AND 3
```

Figure 10.3 - Typical Euclidean cluster analysis output



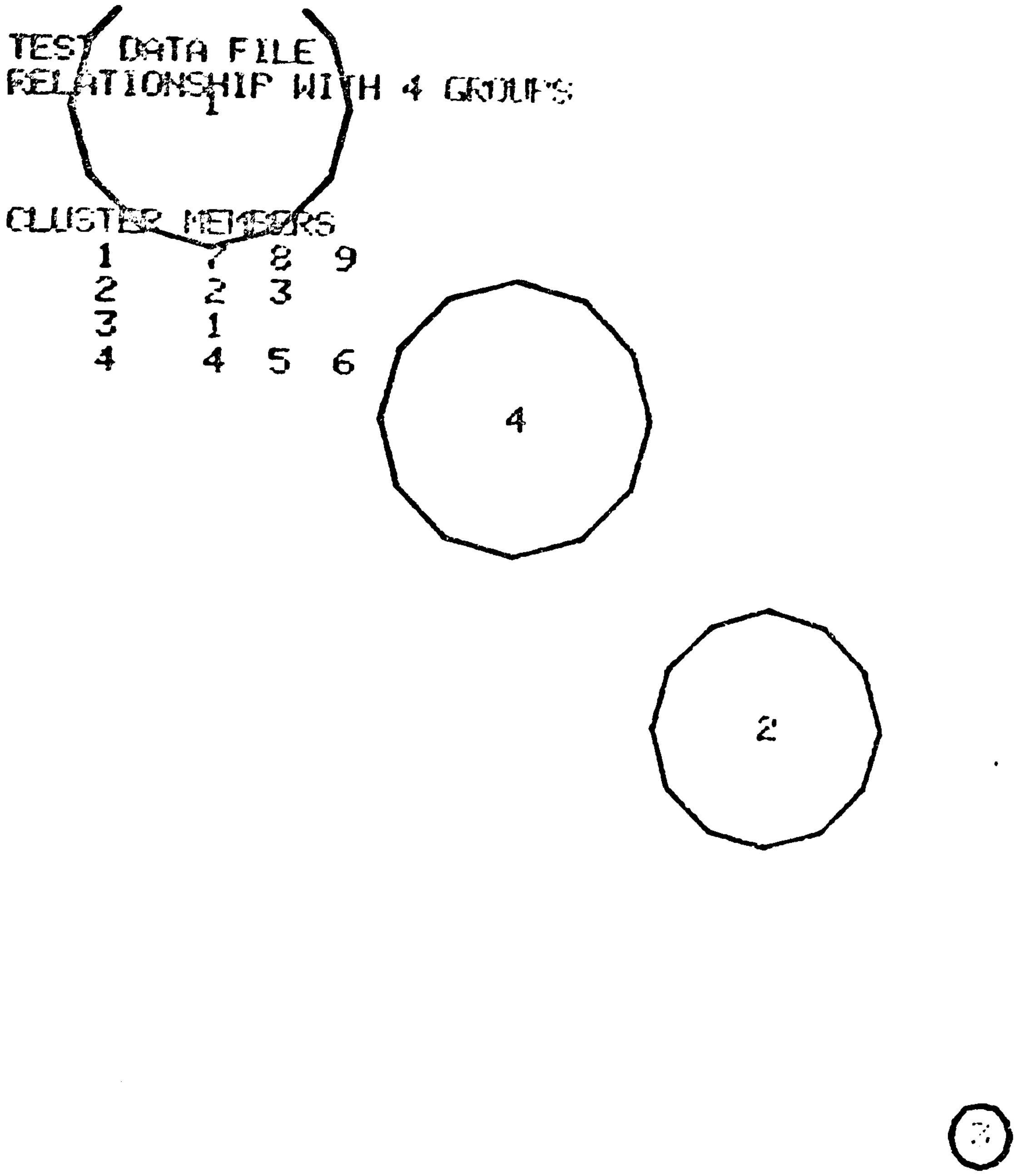
```
RESULTS FOR CURRENT ITERATION WITH    3 CLUSTER CENTRES

CLUSTER    SIZE     DIST FROM GRAND MEAN   CO-ORDINATES

     1       3                    90.000
                                            81.0000    82.0000    83.0000    84.0000    85.0000    86.0000    87.0000    88.0000
                                            89.0000
     2       3                    90.000
                                            21.0000    22.0000    23.0000    24.0000    25.0000    26.0000    27.0000    28.0000
                                            29.0000
     3       3                     0.000
                                            51.0000    52.0000    53.0000    54.0000    55.0000    56.0000    57.0000    58.0000
                                            59.0000

DISTANCE MATRIX FOR CLUSTER CENTRES

     2  180.0000
     3   90.0000   90.0000

CLUSTER MEMBERSHIP FOR INDIVIDUALS

CLUSTER NUMBER    1

  7     65.6658     8     89.5098     9     116.2411

  AVERAGE POINT TO CENTRE DISTANCE    90.47

CLUSTER NUMBER    2

  1    141.1099     2    113.1901     3     86.6718

  AVERAGE POINT TO CENTRE DISTANCE    113.7

CLUSTER NUMBER    3

  4     63.3404     5     48.0833     6     49.1121

  AVERAGE POINT TO CENTRE DISTANCE    53.51

RMS DEVIATION FROM CENTRES =    111.7
CLUSTERS MERGED AT THIS ITERATION:    2 AND    3

RESULTS FOR CURRENT ITERATION WITH    2 CLUSTER CENTRES

CLUSTER    SIZE     DIST FROM GRAND MEAN   CO-ORDINATES

     1       4                    75.000
                                            76.0000    77.0000    78.0000    79.0000    80.0000    81.0000    82.0000    83.0000
                                            84.0000
     2       5                    60.000
                                            31.0000    32.0000    33.0000    34.0000    35.0000    36.0000    37.0000    38.0000
                                            39.0000

DISTANCE MATRIX FOR CLUSTER CENTRES

     2  135.0000
```

Figure 10.3 - Typical Euclidean cluster analysis output (continued)



```
CLUSTER MEMBERSHIP FOR INDIVIDUALS
 CLUSTER NUMBER    1
   6     49.1121    7     65.6658    8     89.5098    9    116.2411
   AVERAGE POINT TO CENTRE DISTANCE   80.13
 CLUSTER NUMBER    2
   1    141.1099    2    113.1901    3     88.6718    4     83.3404    5     48.0833
   AVERAGE POINT TO CENTRE DISTANCE   90.48

 RMS DEVIATION FROM CENTRES =    103.4
 STOP

 END OF EXECUTION
 CPU TIME: 2.26  ELAPSED TIME: 1:44.28
 EXIT
```

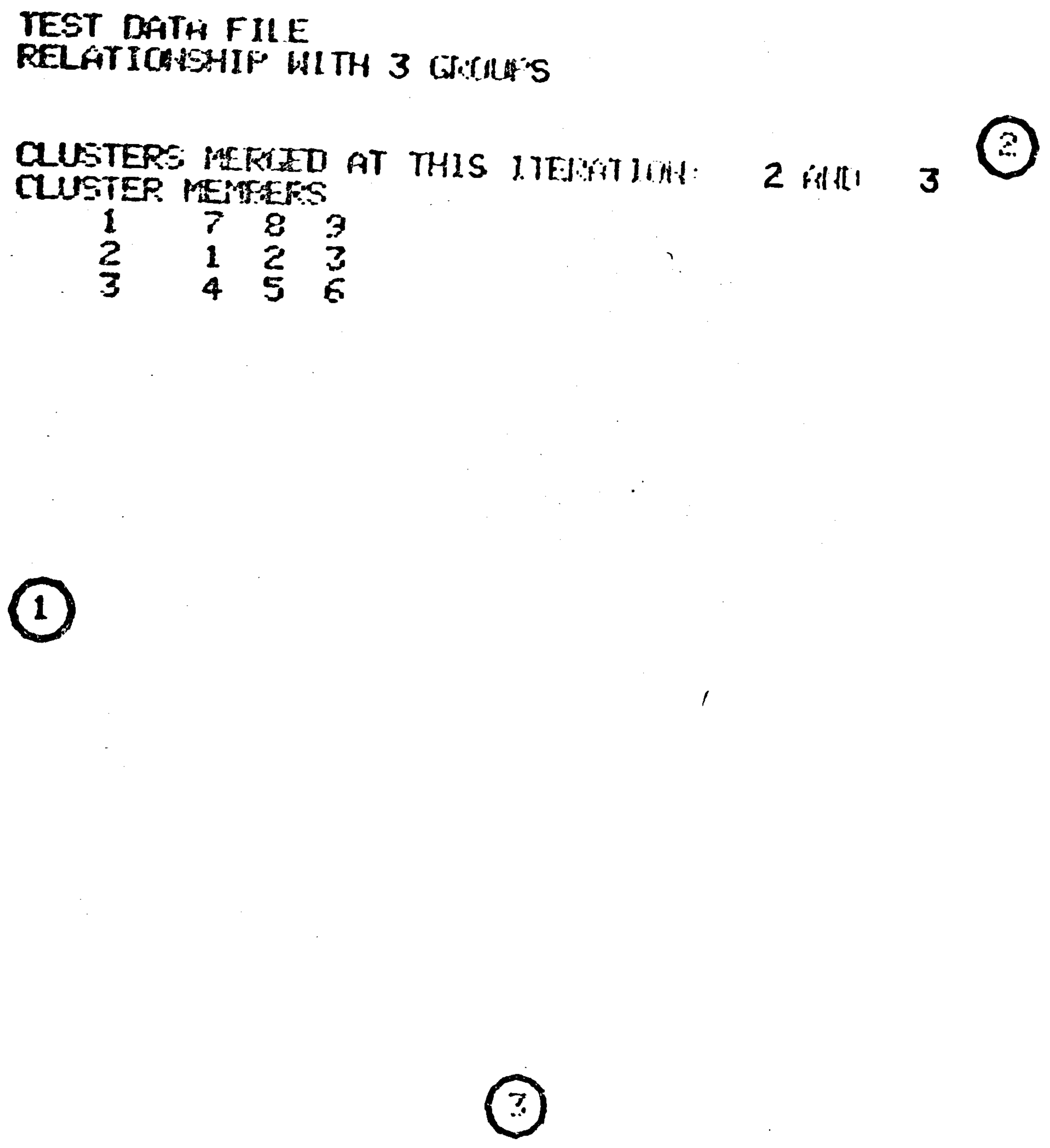
Figure 10.3 - Typical Euclidean cluster analysis output (continued)



```
TEST DATA FILE
RELATIONSHIP WITH 4 GROUPS

CLUSTER MEMBERS
   1        7  8  9
   2        2  3
   3        1
   4        4  5  6
```

Figure 10.4

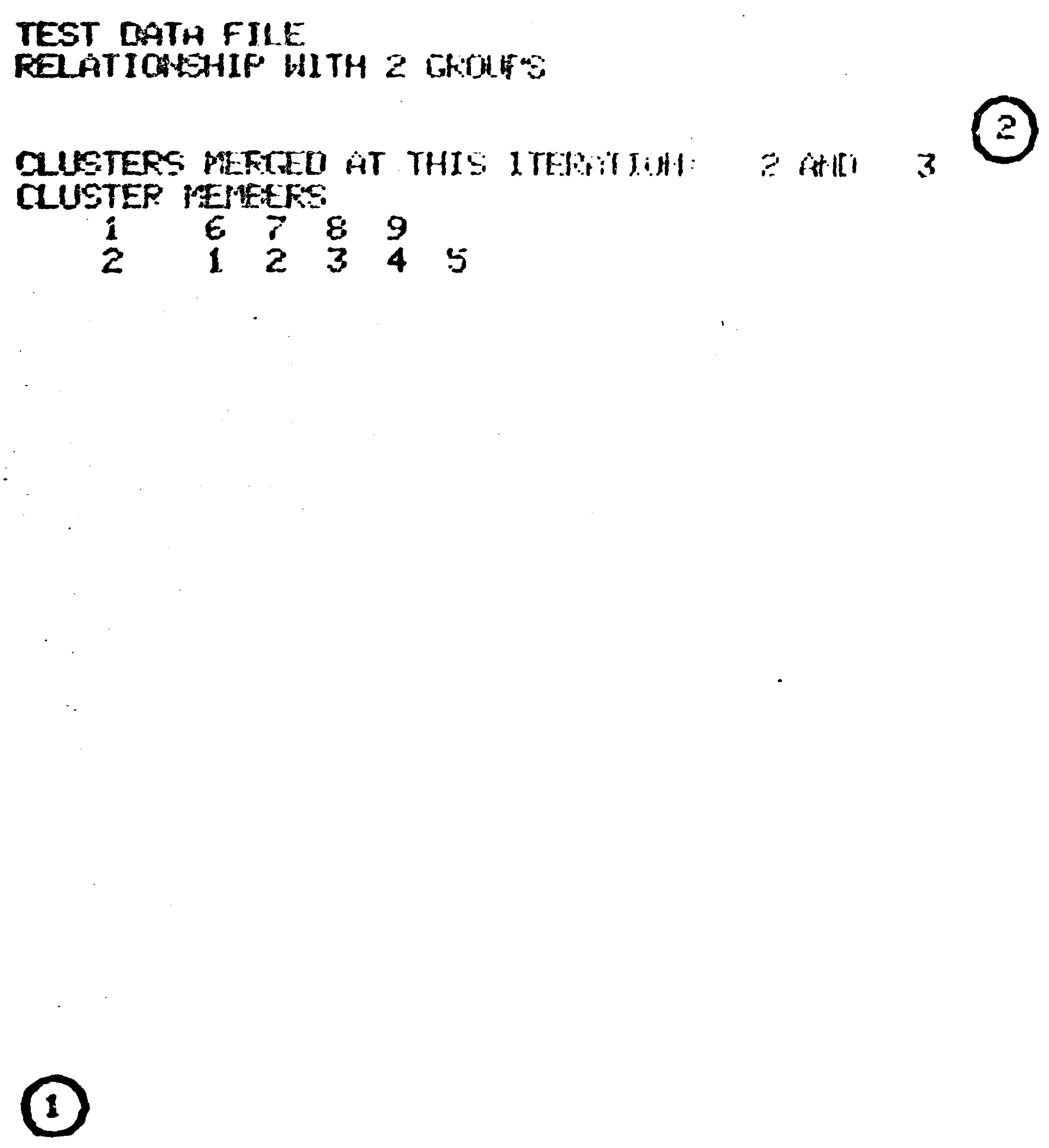
MAGIC Euclidean cluster analysis output



```
TEST DATA FILE
RELATIONSHIP WITH 3 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   2 AND   3
CLUSTER MEMBERS
   1     7  8  9
   2     1  2  3
   3     4  5  6
```

Figure 10.4

MAGIC Euclidean cluster analysis output (continued)



```
TEST DATA FILE
RELATIONSHIP WITH 2 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   2 AND   3
CLUSTER MEMBERS
    1    6  7  8  9
    2    1  2  3  4  5
```

Figure 10.4

MAGIC Euclidean cluster analysis output (continued)

10.4 ORDINATION PLOTS

These are the only displays in MAGIC which may be interactively modified by the program user. The facilities are provided for two reasons. Firstly they provide a convenient method of dealing with the problem of local minima. Secondly as these displays are not fixed clusters but an overall picture of relationships users may wish to rotate or otherwise modify the display to agree more closely with their own mental picture or tentative layout. Figure 10.5 shows a typical display and the controlling menu for the nonlinear mapping analysis. Rather than simply describe the effect of each menu command the principles of "point displays" are dealt with in a more general manner to develop the link between the analysis method and the display of results. The simple translations, rotations, reflections and scalings are discussed first and then their more complex combined use in configuration comparison described in section 10.5.

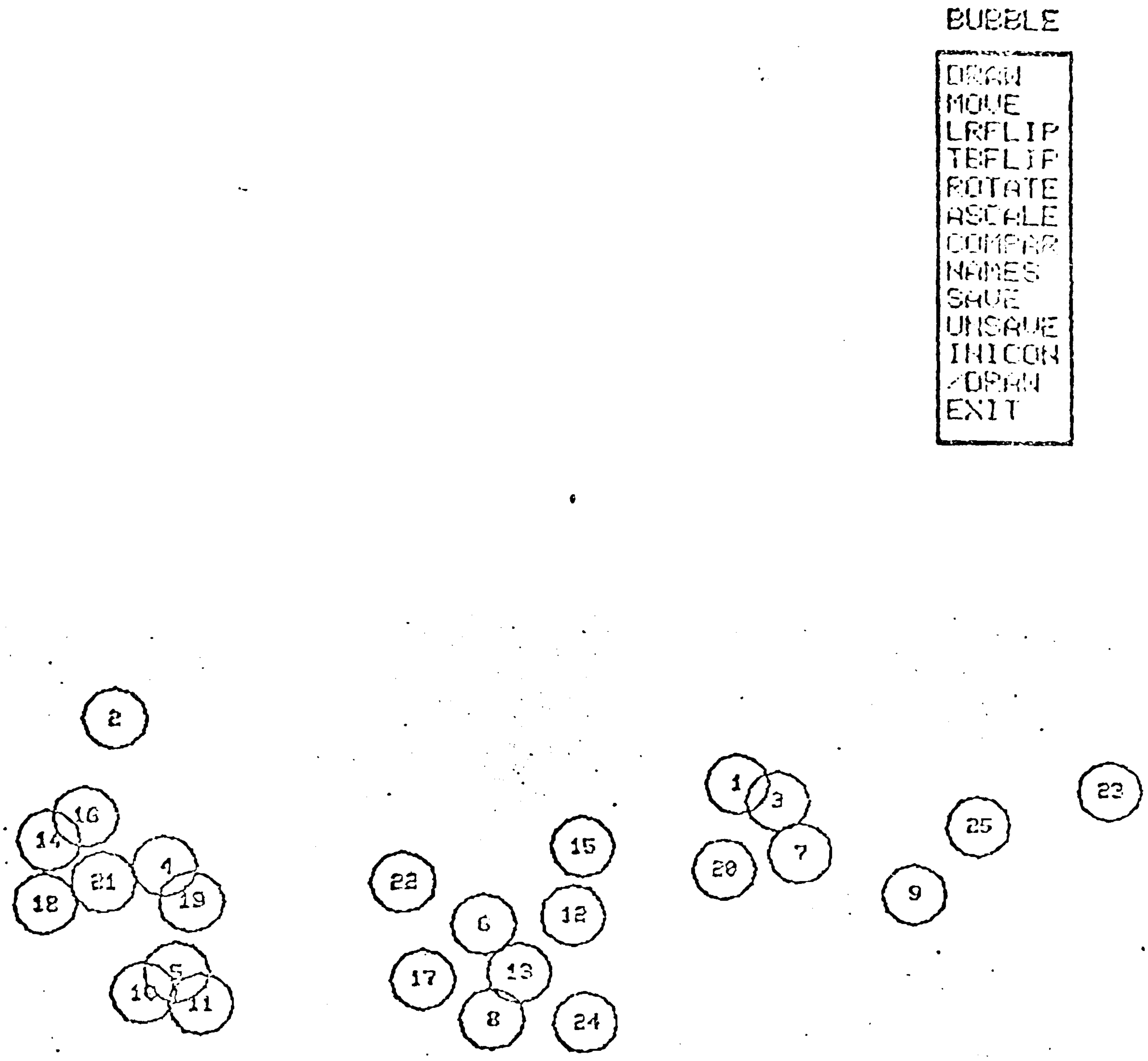


Figure 10.5

BUBBLE menu and display

10.4.1 2D Transformations

Points in the xy plane may be translated to new positions by adding translation amounts to the coordinates of the points. For each point P(x,y) which is to be moved by Dx units parallel to the x-axis and by Dy units parallel to the y-axis to the new point P'(x',y'), we may write

$$x' = x + Dx, \quad y' = y + Dy$$

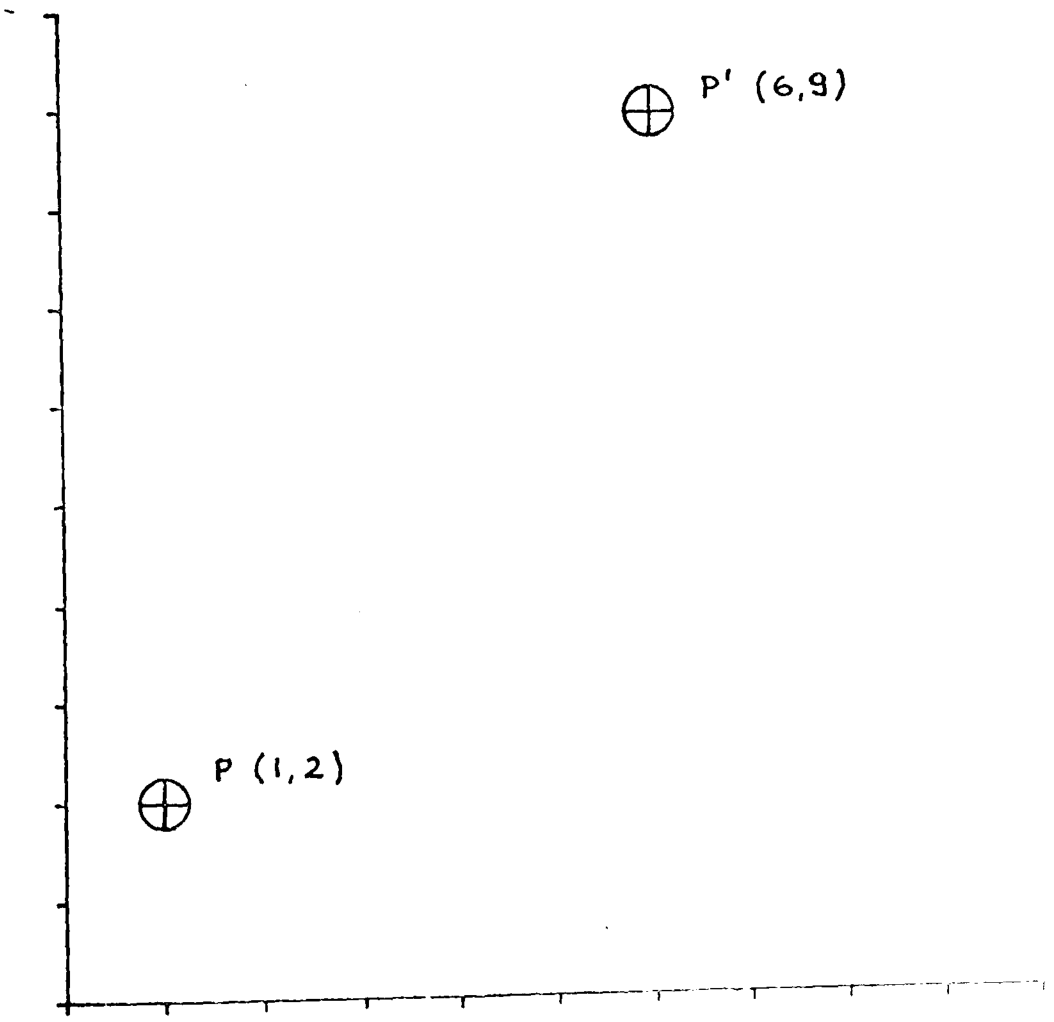


Figure 10.6

Translation of a point

This is illustrated in figure 10.6, in which the point (1,2) is translated by (5,7) to become point (6,9). Defining the following row vectors as

$$P = (x\ y), \quad P' = (x'\ y'), \quad T = (Dx\ Dy)$$

the translation equation may be rewritten

$$(x'\ y') = (x\ y) + (Dx\ Dy)$$

and, even more concisely,

$$P' = P + T$$

Points can be scaled (stretched) by Sx along the x-axis and by Sy along the y-axis into new points by the multiplications:

$$x' = x \cdot Sx, \qquad y' = y \cdot Sy$$

Defining S as $\begin{bmatrix} Sx & 0 \\ 0 & Sy \end{bmatrix}$ we can write in matrix form

$$(x'\ \ y') = (x\ \ y) \begin{bmatrix} Sx & 0 \\ 0 & Sy \end{bmatrix}$$

or

$$P' = P \cdot S$$

In figure 10.7 the single point (6,6) is scaled by 1/2 in x and 1/3 in y. Scaling is about the origin, the point moving closer to the origin. If the scale factors were greater than one the point would move away from the origin. It is also possible for scaling to occur about some point other than the origin. The example shows differential scaling, for which $Sx \neq Sy$, has been used. With a uniform scaling $Sx=Sy$.

Points may be rotated through an angle of θ about the origin, as illustrated in figure 10.8 for the point P(6,1) and angle $\theta = 30°$. The rotation is defined mathematically as:

$$x' = x \cdot \cos\theta - y \cdot \sin\theta$$

$$y' = x \cdot \sin\theta + y \cdot \cos\theta$$

In matrix form we have

$$(x' \quad y') = (x \quad y) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

or

$$P' = P \cdot R$$

where R represents the rotation matrix as defined above. As with scaling the rotation is about the origin, but rotation about an arbitary point is also possible.

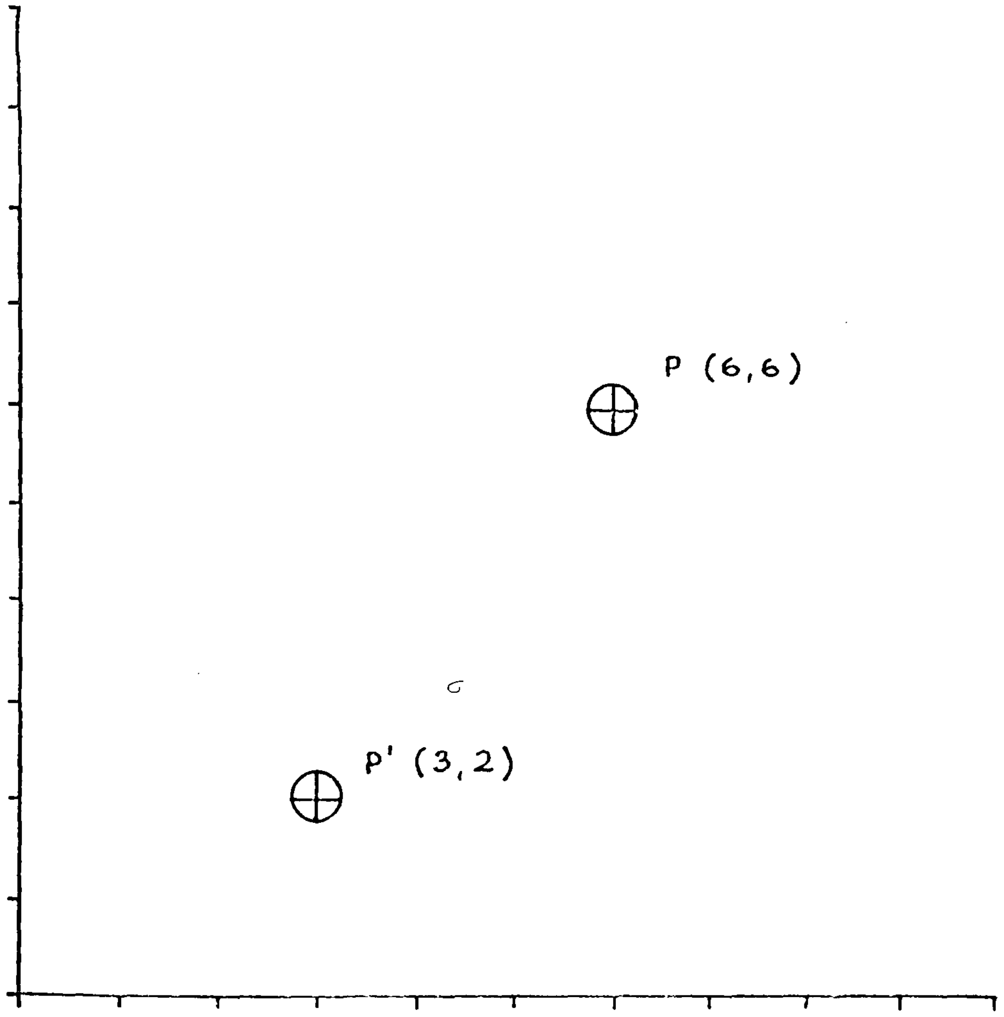


Figure 10.7

Scaling of a point

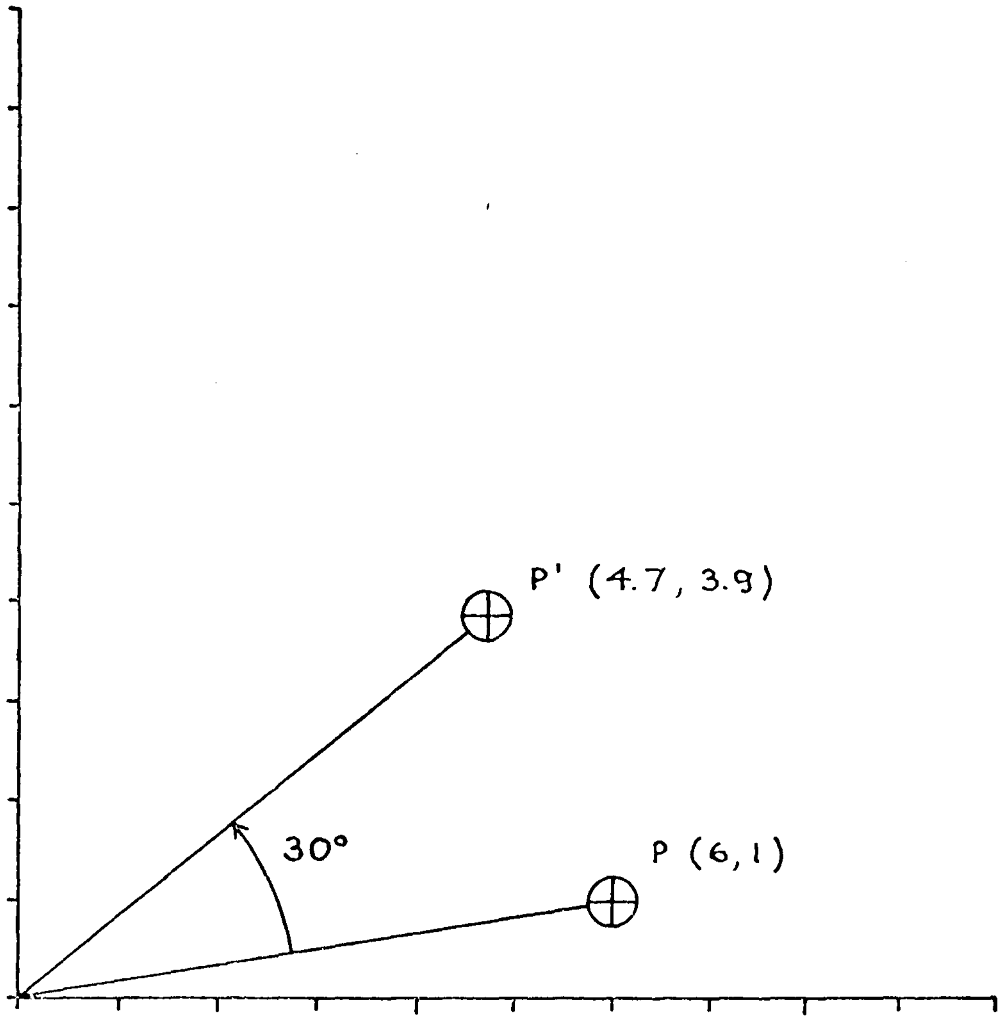


Figure 10.8

Rotation of a point

Positive angles are measured counterclockwise from x towards y. For negative (clockwise) angles the identities $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$ can be used to modify the above equations. The derivation of the equations is easily seen by reference to figure 10.9 in which a rotation by θ transforms P(x,y) into P (x',y'). Because the rotation is about the origin, the distances from the origin to P and P' are equal and labelled r in the figure. By simple trigonometry we note that

$$x = r\cos\phi, \qquad y = r\sin\phi$$

and

$$x' = r\cos(\theta + \phi) = r\cos\phi\cos\theta - r\sin\phi\sin\theta$$

$$y' = r\sin(\theta + \phi) = r\cos\phi\sin\theta + r\sin\phi\cos\theta$$

Then by substitution the basic equations are easily derived.

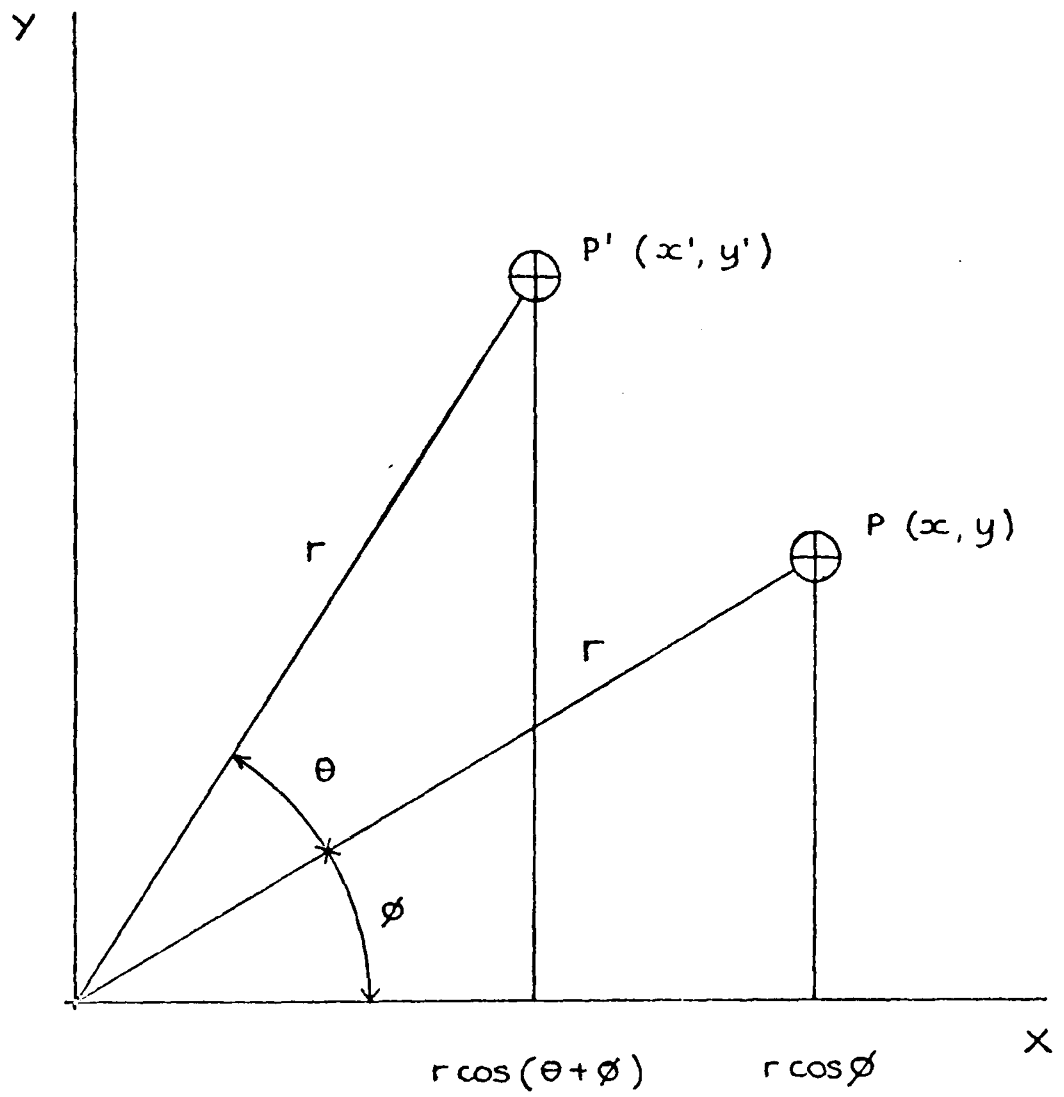


Figure 10.9

Derivation of the rotation equation

10.4.2 Homogenous Coordinates

The matrix equations for translation, scaling, and rotation are, respectively,

$$P' = P + T$$

$$P' = P \cdot S$$

$$P' = P \cdot R$$

Unfortunately, translation is treated differently (as an addition) to scaling and rotation (multiplications). It is advantageous to be able to treat all three in a consistent or homogenous way, so that all three basic transformations may be combined together.

If the points are expressed in homogenous coordinates all three transformations may be treated as multiplications. Homogenous coordinates were developed in geometry (Maxwell 1946,1951) and have subsequently been adopted in computer graphics (Blinn 1977). In homogenous coordinates, point $P(x,y)$ is represented as $P(W \cdot x, W \cdot y, W)$ for any scale factor $W \neq 0$. Then, given a homogenous coordinate representation for a point $P(X,Y,W)$, the two-dimensional cartesian coordinate representation for the point is $x=X/W$ and $y=Y/W$. In MAGIC W is always 1, so the division is never required. Homogenous coordinates may be considered as embedding the two-dimensional plane, scaled by W, in the z=W (here z=1) plane in three-space.

Points are now three-element row vectors, so transformation matrices, which multiply a point vector to produce another point vector, must be 3x3. In the 3x3 matrix form for homogenous coordinates the translation equation is represented as:

$$(x' \quad y' \quad 1) = (x \quad y \quad 1)\cdot\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ Dx & Dy & 1 \end{bmatrix}$$

or, expressed differently,

$$P' = P\cdot T(Dx,Dy),$$

where

$$T(Dx,Dy) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ Dx & Dy & 1 \end{bmatrix}$$

Similarly the scaling equations become

$$(x' \quad y' \quad 1) = (x \quad y \quad 1)\cdot\begin{bmatrix} Sx & 0 & 0 \\ 0 & Sy & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Defining

$$S(Sx,Sy) = \begin{bmatrix} Sx & 0 & 0 \\ 0 & Sy & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

we obtain

$$P' = P \cdot S(Sx, Sy)$$

Finally, the rotation equations become

$$(x' \quad y' \quad 1) = (x \quad y \quad 1) \cdot \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

letting

$$R(\theta) = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

we have

$$P' = P \cdot R(\theta)$$

10.5 PROCRUSTES ANALYSIS

Chapters 7, 8 and 9 have described a number of ways of obtaining a two-dimensional representation of the data set. The objective of Procrustes Analysis is to enable a comparison of the configurations. The procedure is referred to as Procrustes analysis after the mythological Greek innkeeper Procrustes who ensured his clients fitted his beds by either stretching them or cutting off their limbs. This section describes a method for the comparison of two geometrical configurations; a generalisation to comparing more than two configurations is given by Gower (1975) and ten Berge (1977). The idea of seeking to transform one matrix into another was first proposed by Mosier (1939). A solution with transformations restricted to orthogonal rotations/reflections was given by Green (1952) and later generalised by Schonemann (1966, 1968), Gruvaeus (1970), and Schonemann and Carroll (1970). This description is roughly based on Schonemann and Carroll's solution where four basic geometric transformations are included:

(i) translation of the origin

(ii) rotation of points

(iii) reflection of points

(iv) uniform dilation of points

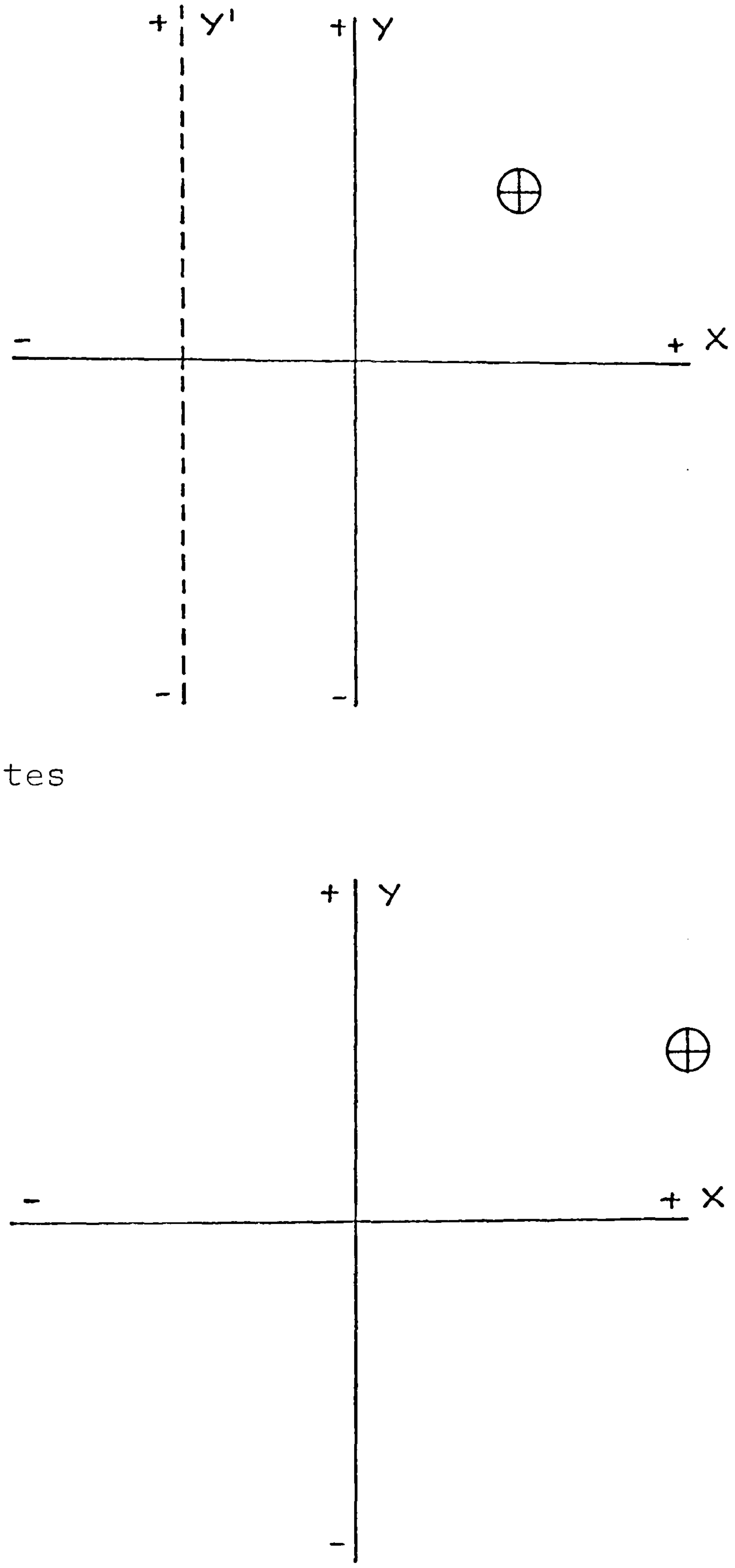
It will be assumed that the configurations to be compared have coordinates given by the (nxp) matrices X = (x_{ik}) and Y = (y_{ik}) where the activities are specified in the same order in the rows of the two matrices. The measure of fit used to assess the resemblance of the two configurations is the sum of the squared distances between corresponding points in the two configurations:

$$\Delta^2(X,Y) = \sum_{i=1}^{n}\sum_{k=1}^{p}(x_{ik} - y_{ik})^2$$
$$= \mathrm{trace}\{(X - Y)'(X - Y)\}$$

This measure as it stands is not used directly as it is usual for one of the configurations to be held fixed and the other transformed to fit as closely as possible. The four geometric transformations used to map the "to-be-fitted" matrix to the "target" matrix are considered below.

10.5.1 Matching Under Translation

There are two ways of viewing the elementary geometric transformations: (a) as an alteration of the coordinate system leaving the space element undisturbed, or (b) as a displacement of the space element itself while the coordinates remain fixed. In both cases the end result is the same. The alternative approaches may be seen in figure 10.10.



(a) co-ordinates

(b) point

Figure 10.10

Translation of points and co-ordinates

The distance measure above may be rewritten as

$$\Delta^2(X,Y) = \sum_{i=1}^{n}\sum_{K=1}^{p}\{(x_{ik} - \bar{x}_{.k}) - (y_{ik} - \bar{y}_{.k})\}^2 + n\sum_{k=1}^{p}(\bar{x}_{.k} - \bar{y}_{.k})^2$$

where

$$\bar{x}_{.k} = 1/n\sum_{i=1}^{n} x_{ik}$$

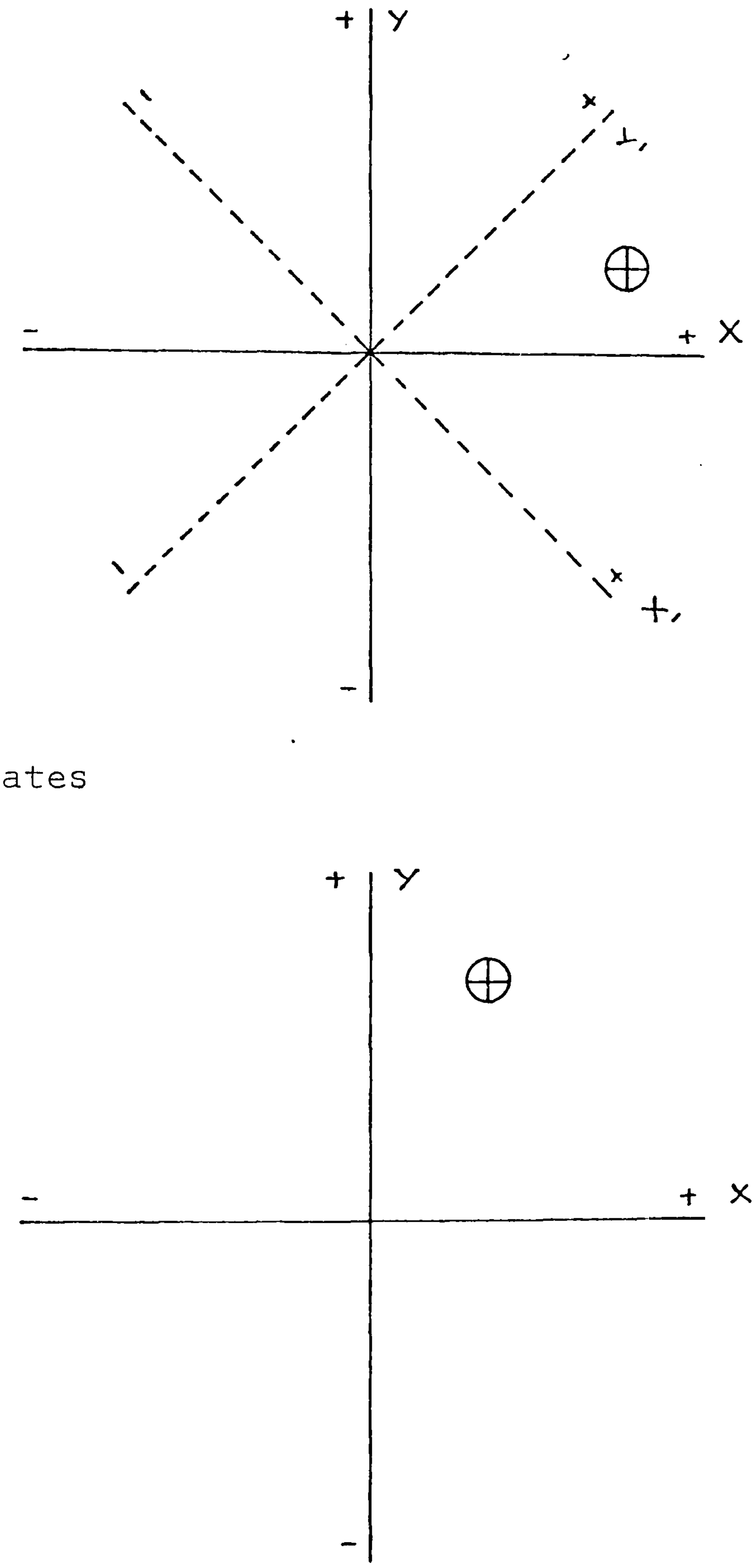
and

$$\bar{y}_{.k} = 1/n\sum_{i=1}^{n} y_{ik} \qquad (K = 1, \ldots, p)$$

Hence optimal matching under translation of origins is attained uniquely by ensuring that the centroids of the two configurations coincide. By placing this common centroid at the origin of coordinates this standardisation will be undisturbed by subsequent rotation, reflection and dilation. Thus it is possible to assume throughout the following descriptions that all configurations are centre-at-origin standardised, and this shall be done to simplify the presentation without altering its content.

10.5.2 Matching Under Rotation And Reflection

The geometric motions are shown in figures 10.11 and 10.12. After the centroid-at-origin standardisation has been made, the matching problem reduces to finding an orthogonal matrix R which minimises $\Delta^2(X,YR)$.

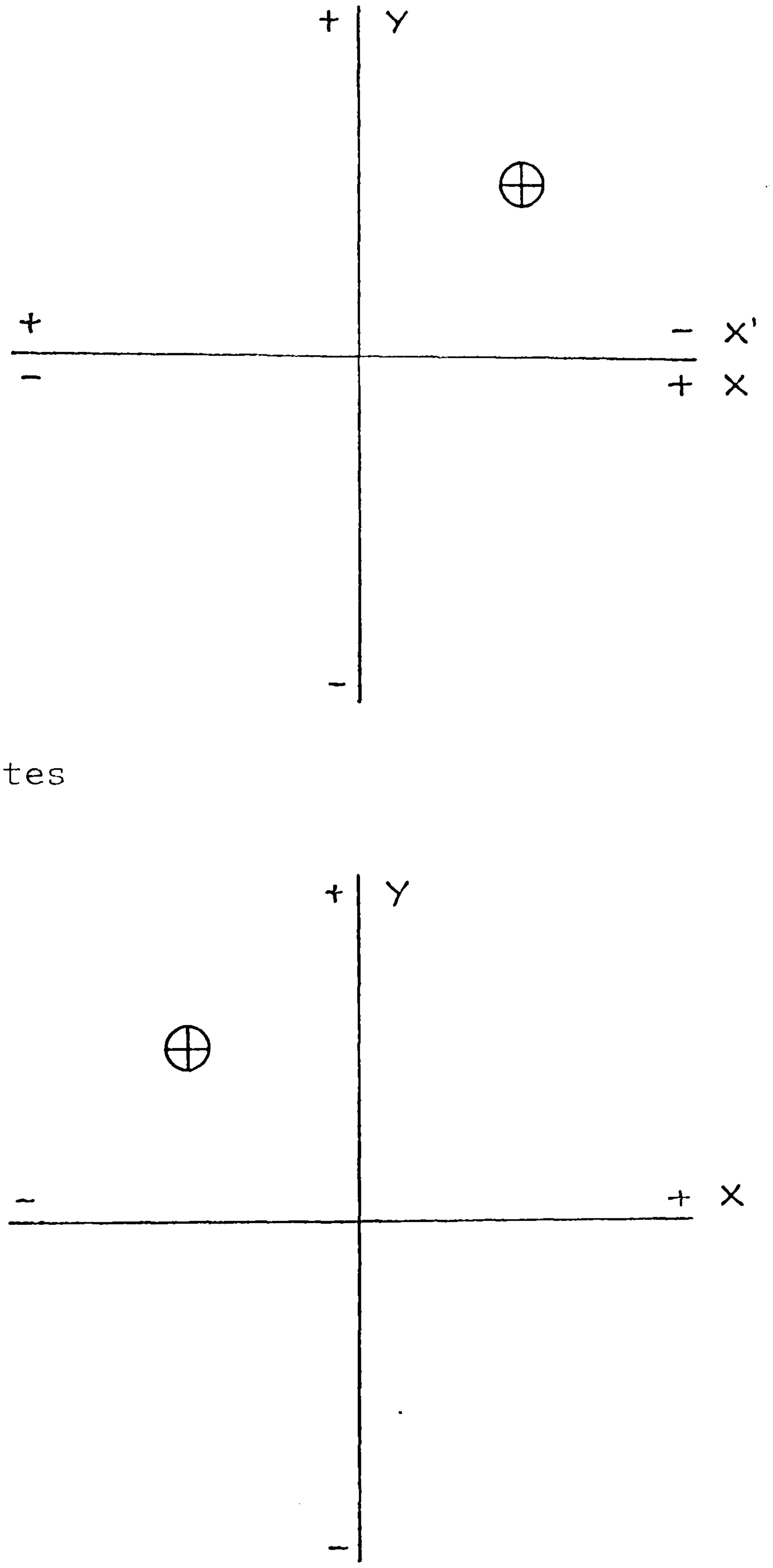


(a) co-ordinates

(b) point

Figure 10.11

Rotation of points and co-ordinates



(a) co-ordinates

(b) point

Figure 10.12

Reflection of points and co-ordinates

Expanding this expression,

$$\begin{aligned}\Delta^2(X,YR) &= \text{trace}\{(X - YR)'(X - YR)\}\\ &= \text{trace}(X'X) + \text{trace}(Y'Y) - 2\text{trace}(R'Y'X)\end{aligned}$$

Sibson (1978) shows that given a square matrix, A, and an orthogonal matrix R, of the same size, then

$$\text{trace}(R'A) \leqslant \text{trace}\{(A'A)^{\frac{1}{2}}\}$$

with equality if and only if R satisfies. $R'A = (A'A)^{\frac{1}{2}}$ ($M^{\frac{1}{2}}$ denotes the non-negative definite symmetric square root of the non-negative definite symmetric matrix M). This equation always has an orthogonal solution R, and if A is non-singular the solution is uniquely $R = A(A'A)^{-\frac{1}{2}}$.

Using this result, the following theorem is obtained by substituting A = (Y'X) in the equation above.

Theorem

If X and Y are configurations which have been centred at the origin, $\Delta^2(X,Y)$ is minimised by transforming Y to YR, where R is an orthogonal solution of $R'Y'X = (X'YY'X)^{\frac{1}{2}}$. If Y'X is non-singular, $R = Y'X(X'YY'X)^{-\frac{1}{2}}$. The minimum value is

$$\Delta^2(X,YR) = \text{trace}(X'X) + \text{trace}(Y'Y) - 2\text{trace}\{(X'YY'X)^{1/2}\}$$

10.5.3 Matching Under Dilation

Finally the transformation of scale may be applied as defined in figure 10.13. The transformation of uniform dilation involves multiplying all the coordinate values of Y by a positive constant σ. For a given value of σ the theorem defined above shows that the minimum value of $\Delta^2(X,\sigma YR)$ is

$$\sigma^2\text{trace}(Y'Y) - 2\sigma\text{trace}\{(X'YY'X)^{1/2}\} + \text{trace}(X'X)$$

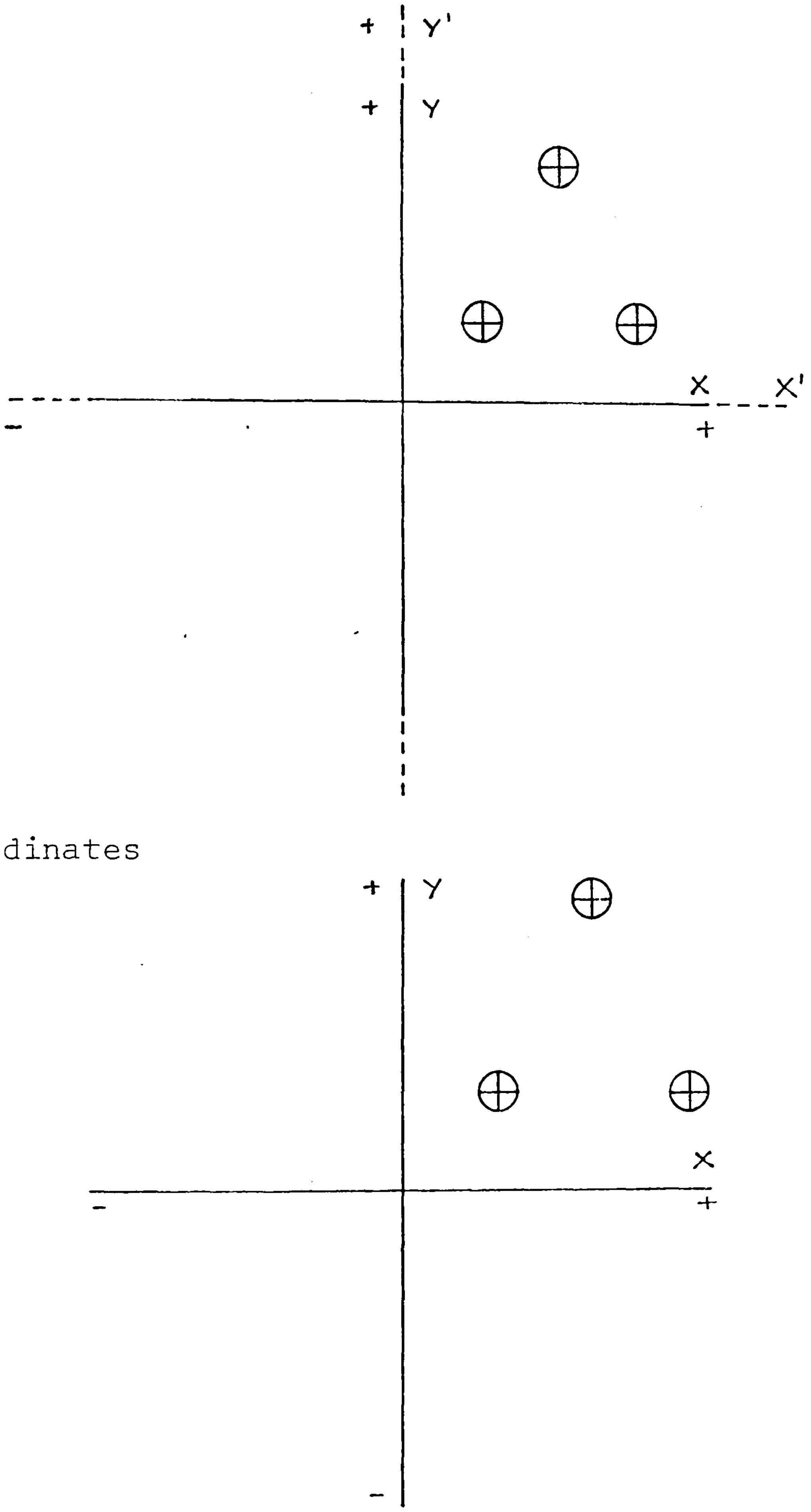
For given X and Y, this quadratic expression in σ^2 is reduced to its minimum value of

$$\text{trace}(X'X) - [\text{trace}\{(X'YY'X)^{1/2}\}]^2/\text{trace}(Y'Y)$$

by choosing

$$\sigma = \text{trace}\{(X'YY'X)^{1/2}/\text{trace}(Y'Y)$$

As the procedure is independent of any scaling factor σ, optimal fit is obtained within the class of transformations considered by carrying them out in the order described.



(a) co-ordinates

(b) points

Figure 10.13

Dilation of points and co-ordinates

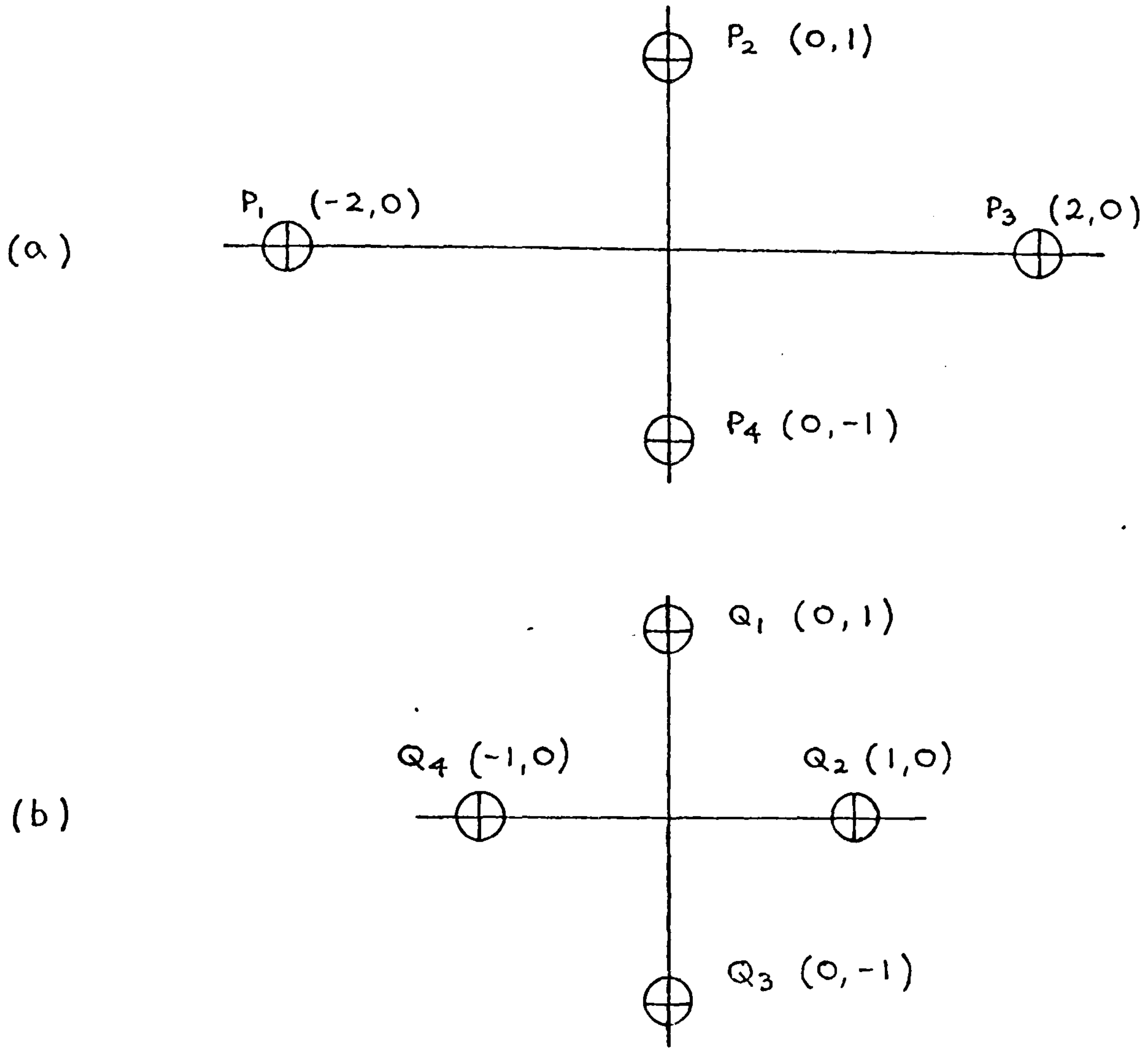


Figure 10.14

Two geometrical representations of the same set of four objects, the ith object (i=1,...,4) being represented by point P_i in (a) and by point Q_i in (b). The configurations are defined by the following matrices:

$$\text{(a)} \quad X = \begin{bmatrix} -2 & 0 \\ 0 & 1 \\ 2 & 0 \\ 0 & -1 \end{bmatrix} \qquad \text{(b)} \quad Y = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}$$

10.5.4 Computational Procedure

The theory presented in this section is illustrated by a simple numerical example, summarised in figure 10.14, which shows two different geometrical representations of a set of four objects. The Y-configuration will be transformed to fit the X-configuration as closely as possible. Both configurations are already centred at the origin, and so the next step is to find the optimal rotation and/or reflection for the Y-configuration. For these data,

$$A = Y'X = \begin{bmatrix} 0 & 2 \\ -4 & 0 \end{bmatrix}$$

which is non-singular

$$(A'A)^{-\frac{1}{2}} = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Hence the optimal rotation/reflection is

$$R = A(A'A)^{-\frac{1}{2}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

a counterclockwise rotation through ninety degrees

Similarly, the dilation factor is

$$\sigma = \text{trace}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \Big/ \text{trace}\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 6/4$$

Thus, the optimal transformed Y-configuration is given by

$$Y^* = \sigma YR = 6/4 \begin{bmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}$$

The transformed points $\{Q^*_i\ (i = 1, \ldots, 4)\}$ whose coordinates are given by Y* are plotted in figure 10.15(a), together with the X-configuration, to which they are the optimal approximation under the given set of transformations. The sum of squared distances between the two configurations is 1.

In order to demonstrate that fitting X to Y need not give the inverse scaling to fitting Y to X, the results are summarised in figure 10.15(b) of the transformation of $\{P_i\ (i = 1, \ldots, 4)\}$ to $\{P^*_i (i = 1, \ldots, 4)\}$ so as to obtain an optimal approximation to the set of points $\{Q_i\ (i = 1, \ldots, 4)\}$; in this case, the sum of squared distances between the corresponding points in the two configurations takes the value 0.4. There is no rotation and dilation which will exactly match the configuration of eight points in figure 10.15.

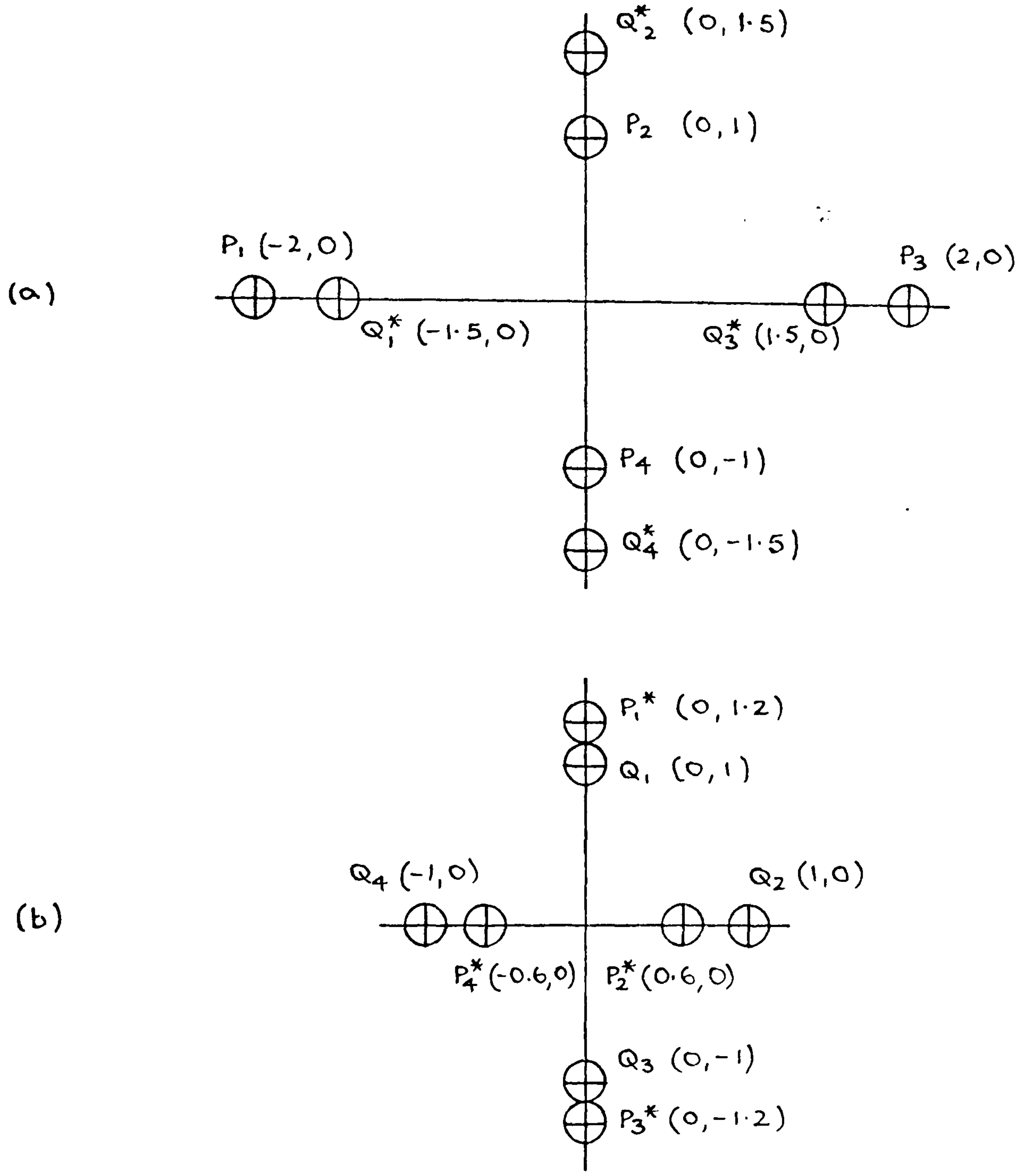


Figure 10.15

Demonstration of the transformation of one of the configurations defined in figure 10.14 in order to fit the other configuration: (a) Y transformed to fit X; (b) X transformed to fit Y

10.5.5 Example From MAGIC

Figures 10.16 and 10.17 show two test matrices roughly based about a square in two-dimensional space. Figure 10.18 shows the tabulated output of the Procrustes analysis with two statistical measures of goodness of fit developed by Lingoes(1973) and Schonemann (1970). Schonemann's symmetric coefficient (s) is probably the most understandable as it varies $0 \leqslant s \leqslant 1$ with 0 being a perfect fit. In the graphical implementation in MAGIC (figure 10.19) neither measure is shown - it being considered irrelevent to the conceptual rather than theoretical fitting being carried out in this context.

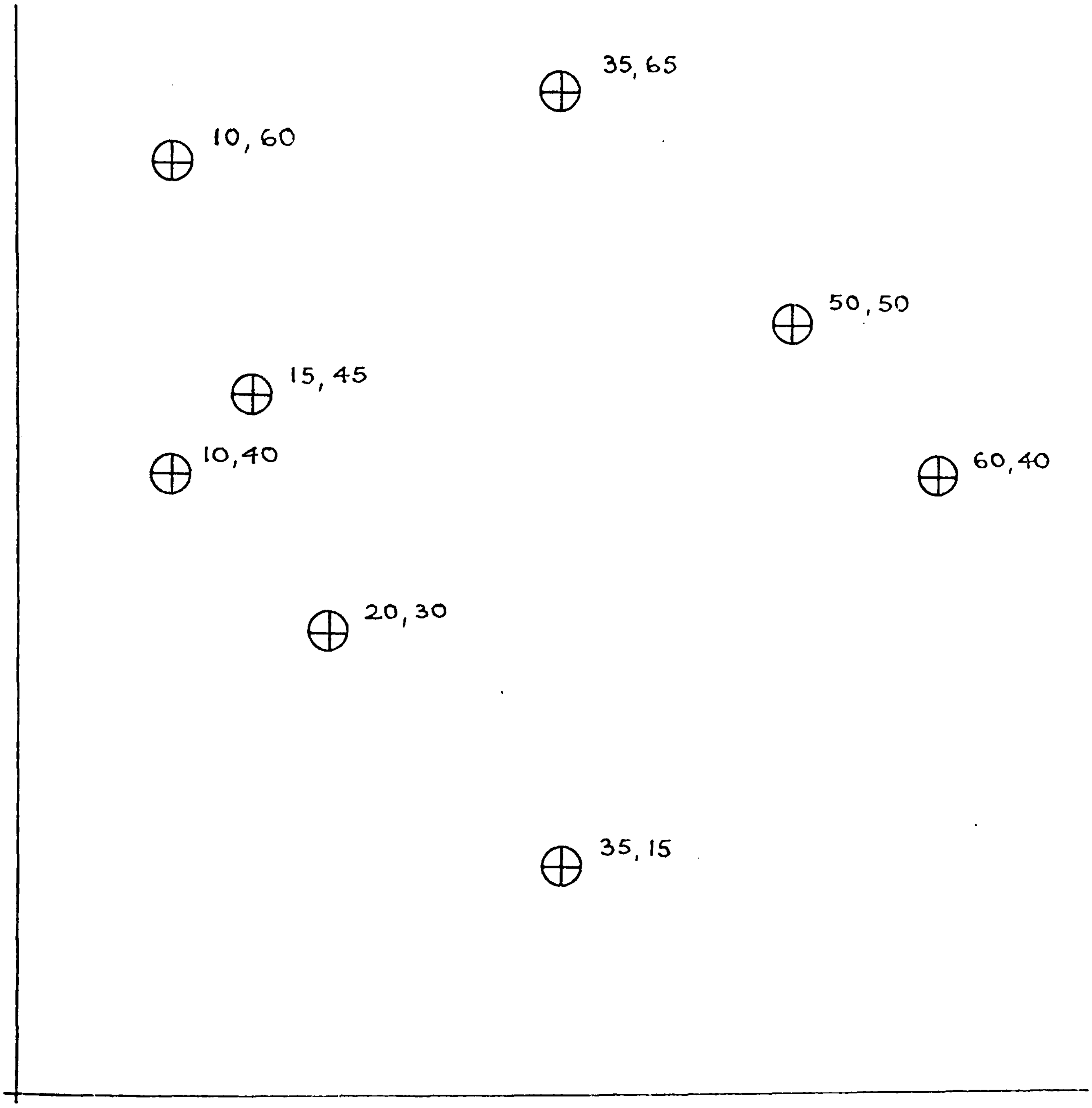


Figure 10.16

Test data matrix 1 - "Matrix A"

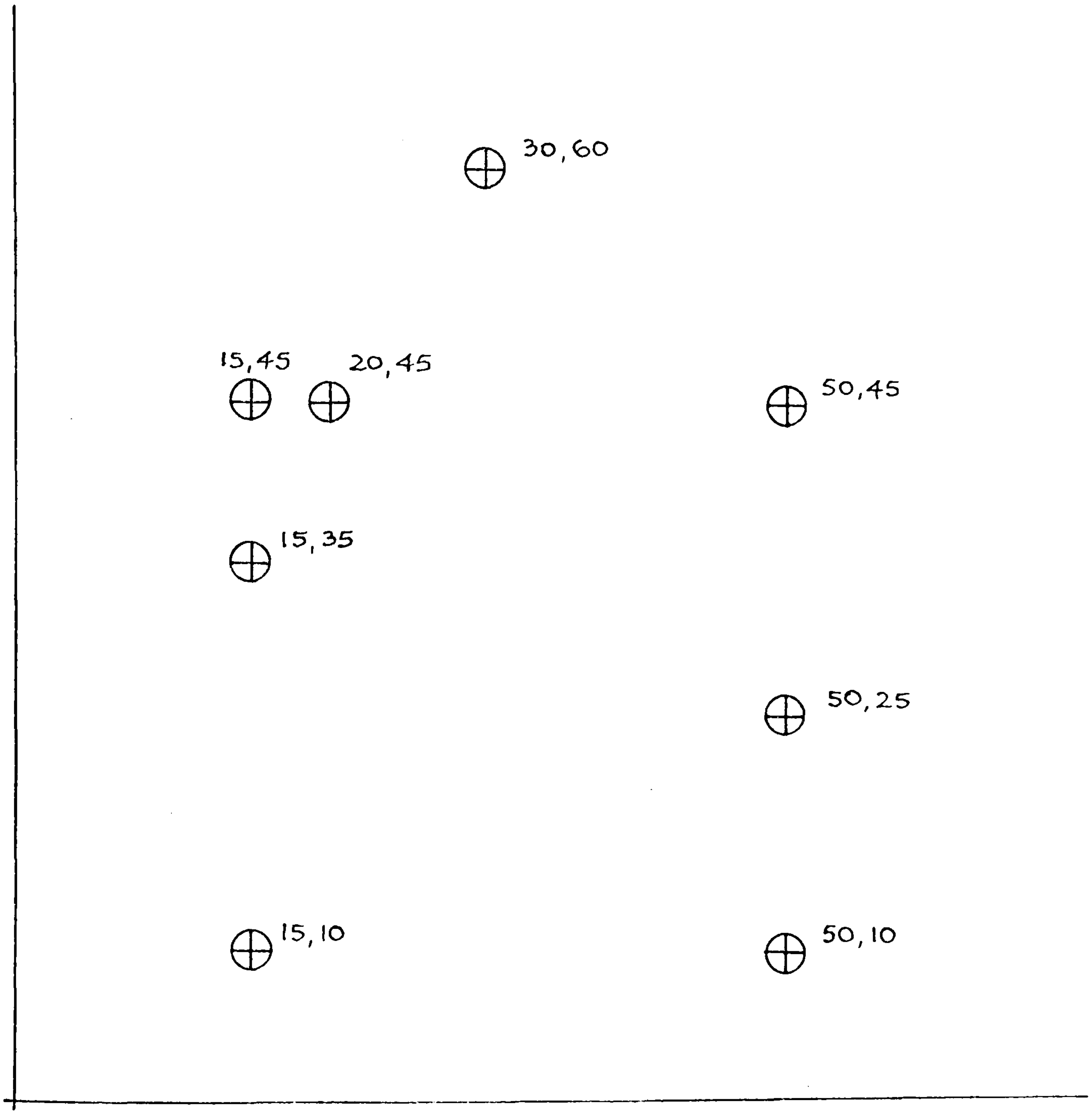
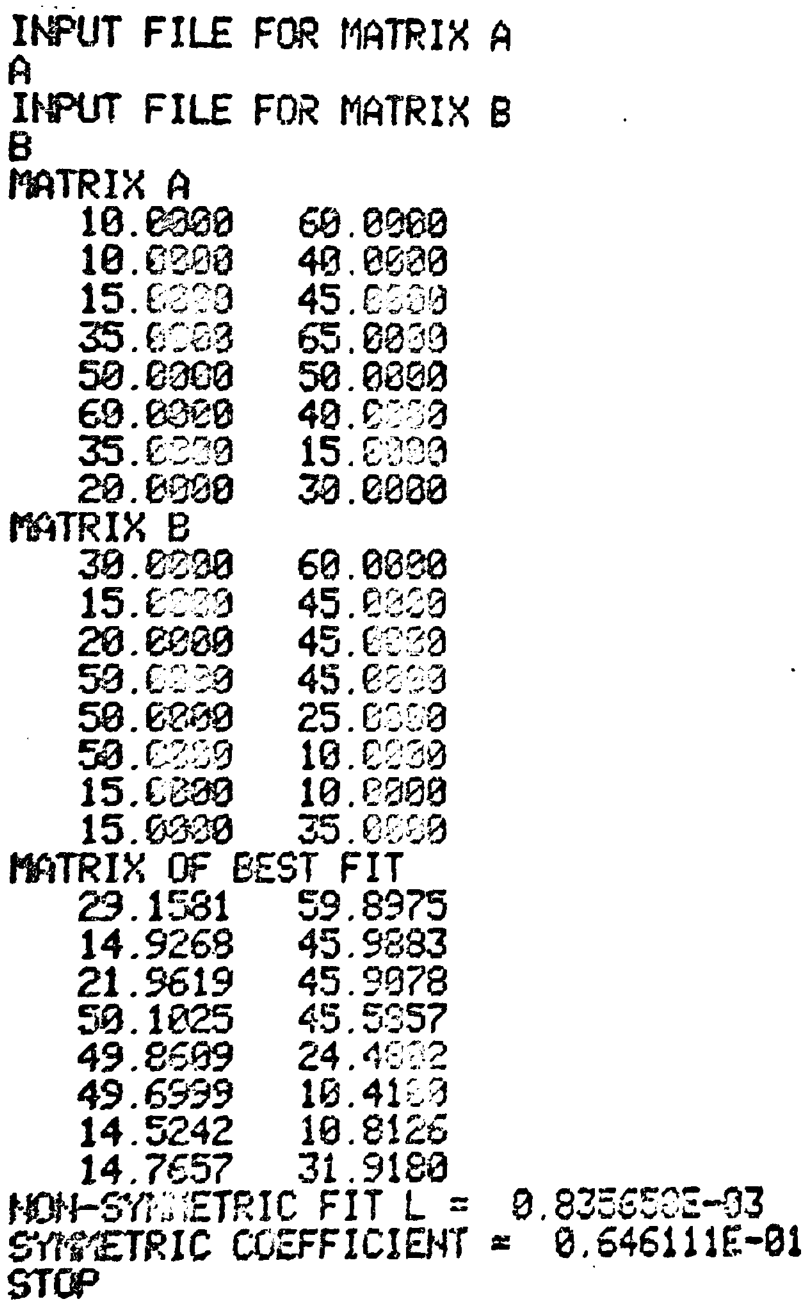


Figure 10.17

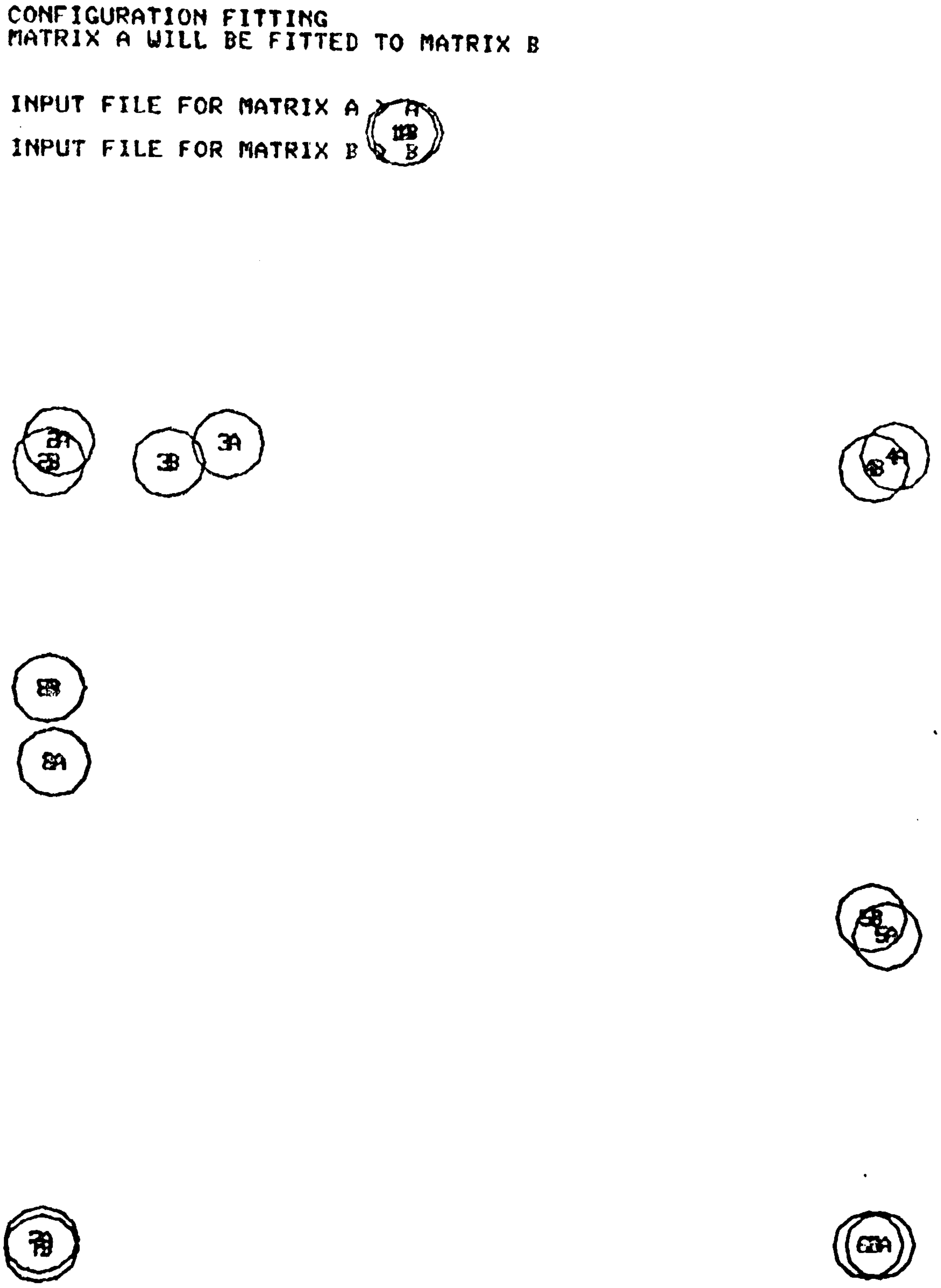
Test data matrix 2 - "Matrix B"



```
INPUT FILE FOR MATRIX A
A
INPUT FILE FOR MATRIX B
B
MATRIX A
    10.0000    60.0000
    10.0000    40.0000
    15.0000    45.0000
    35.0000    65.0000
    50.0000    50.0000
    60.0000    40.0000
    35.0000    15.0000
    20.0000    30.0000
MATRIX B
    30.0000    60.0000
    15.0000    45.0000
    20.0000    45.0000
    50.0000    45.0000
    50.0000    25.0000
    50.0000    10.0000
    15.0000    10.0000
    15.0000    35.0000
MATRIX OF BEST FIT
    29.1581    59.8975
    14.9268    45.9883
    21.9619    45.9078
    50.1025    45.5957
    49.8609    24.4882
    49.6999    10.4180
    14.5242    10.8126
    14.7657    31.9180
NON-SYMMETRIC FIT L =  0.835658E-03
SYMMETRIC COEFFICIENT =  0.646111E-01
STOP
```

Figure 10.18

Tabulated output from Procrustes analysis



```
CONFIGURATION FITTING
MATRIX A WILL BE FITTED TO MATRIX B

INPUT FILE FOR MATRIX A > A
INPUT FILE FOR MATRIX B > B
```

Figure 10.19

MAGIC Procrustes analysis output

CHAPTER 11

DATA CLUSTERING

11.1 INTRODUCTION

This chapter presents a technique for identifying and displaying natural groups and clusters that may occur in complex data arrays. The method adopted is the well-known technique of permuting the rows and columns of the data matrix in such a way as to group the larger array elements together. This option is included in MAGIC for two reasons. Firstly, although an established formal design method, the task is very difficult to accomplish manually. Secondly all the other clustering and ordination techniques operate on modified data with some inevitable loss of information. It is therefore valuable to have available an option which operates on the "raw" data matrix and represents it in a way which enables clusters to be identified. The technique thus provides a useful check when used in conjunction with the other techniques, but also supplies additional information insofar as it is possible to identify from the rearranged matrix not only the clusters of activities, but also those variables upon which the clusters are based.

11.2 THE TECHNIQUE USED

The problem may be formulated as a "travelling salesman problem". This classical operations research problem notionally concerns a salesman who wishes to find the shortest route through a number of cities and back home again. Stated more formally, given a finite set of N cities and a distance matrix (d_{ij}) $(i,j \in N)$, determine

$$\min_{*} \sum_{i \in n} d_{i*(i)}$$

where * runs over all cyclic permutations of N; *k(i) is the kth city reached by the salesman from city i. If N = (1,.....,n), then an equivalent formulation is

$$\min_{v}(\sum_{i=1}^{n-1} d_{v(i)v(i+1)} + d_{v(n)v(1)})$$

where v runs over all permutations of N; here v(k) is the kth city in a salesman's tour. If $d_{ij} = d_{ji}$ for all (i,j) the problem is called symmetric, otherwise it is assymetric. If $d_{ik} \leqslant d_{ij} + d_{jk}$ for all (i,j,k) the problem is Euclidean.

Bellmore and Nemhauser (1968), Eilon et al (1971), Bellmore and Malone (1971) and Christofides (1975) all contain surveys of well-known solution techniques. The solution technique adopted in MAGIC is a modification of the suboptimal method which constructs a tour by successively inserting cities.

11.3 PROBLEM FORMULATION

Suppose that a data array (a_{ij}) $(i \in R,\ j \in S)$ is given, where a_{ij} measures the strength of the relationship between the elements $i \in R$ and $j \in S$. A clustering of the array is obtained by permuting its rows and columns and should identify subsets of R that are strongly related to subsets of S.

To convert this problem into an optimisation problem some criterion must be defined. By defining a "clumping factor", CF, as a criterion to optimise, the problem may be formulated in terms of the travelling salesman problem. The CF used is the sum of all products of horizontally or vertically adjacent elements in the matrix. Figure 11.1 shows how this criterion relates to various permutations of a 4 x 4 array.

$$\begin{array}{ccccc} & s_1 & s_2 & s_3 & s_4 \\ r_1 & 1 & 0 & 1 & 0 \\ r_2 & 0 & 1 & 0 & 1 \\ r_3 & 1 & 0 & 1 & 0 \\ r_4 & 0 & 1 & 0 & 1 \end{array} \qquad CF = 0$$

$$\begin{array}{ccccc} & s_1 & s_2 & s_3 & s_4 \\ r_1 & 1 & 0 & 1 & 0 \\ r_2 & 0 & 1 & 0 & 1 \\ r_3 & 0 & 1 & 0 & 1 \\ r_4 & 1 & 0 & 1 & 0 \end{array} \qquad CF = 2$$

$$\begin{array}{ccccc} & s_1 & s_2 & s_3 & s_4 \\ r_1 & 1 & 0 & 1 & 0 \\ r_2 & 1 & 0 & 1 & 0 \\ r_3 & 0 & 1 & 0 & 1 \\ r_4 & 0 & 1 & 0 & 1 \end{array} \qquad CF = 4$$

$$\begin{array}{ccccc} & s_1 & s_2 & s_3 & s_4 \\ r_1 & 1 & 1 & 0 & 0 \\ r_2 & 0 & 0 & 1 & 1 \\ r_3 & 0 & 0 & 1 & 1 \\ r_4 & 1 & 1 & 0 & 0 \end{array} \qquad CF = 6$$

$$\begin{array}{ccccc} & s_1 & s_2 & s_3 & s_4 \\ r_1 & 1 & 1 & 0 & 0 \\ r_2 & 1 & 1 & 0 & 0 \\ r_3 & 0 & 0 & 1 & 1 \\ r_4 & 0 & 0 & 1 & 1 \end{array} \qquad CF = 8$$

Figure 11.1

Clumping Factors for various permutations
of a 4 x 4 array.

The problem is then to find the permutation of rows and columns of (a_{ij}) which maximises CF.

Let $R = (1, . . ., r)$

and $S = (1, . . ., s)$

with the conventions

$\rho(0) = \rho(r+1) = \sigma(0) = \sigma(s+1) = *$

$a_{i*} = a_{*j} = 0$ for $i \in R$, $j \in S$

Then CF, corresponding to permutations ρ of R and σ of S, is given by

$$\begin{aligned} CF(\rho,\sigma) &= 0.5 \sum_{i \in R} \sum_{j \in S} a_{\rho(i)\sigma(j)}(a_{\rho(i)\sigma(j-1)} + a_{\rho(i)\sigma(j+1)} + \\ &\qquad a_{\rho(i-1)\sigma(j)} + a_{\rho(i+1)\sigma(j)}) \\ &= \sum_{j=0}^{s} \sum_{i \in R} a_{i\sigma(j)}\, a_{i\sigma(j+1)} + \sum_{i=0}^{r} \sum_{j \in S} a_{\rho(i)j}\, a_{\rho(i+1)j} \\ &= CF(\sigma) + CF(\rho) \end{aligned}$$

Thus $CF(\rho,\sigma)$ decomposes into two parts, and its maximisation reduces to two seperate and similar optimisations, one of $CF(\sigma)$ for the columns and the other of $CF(\rho)$ for the rows.

11 4 SOLUTION ALGORITHM

A sequential selection suboptimal algorithm is used to determine array orderings corresponding to local optima of CF. The algorithm operates as follows

1. Place one of the columns arbitrarily. Set i=1.

2. Try placing each of the remaining N-i columns in

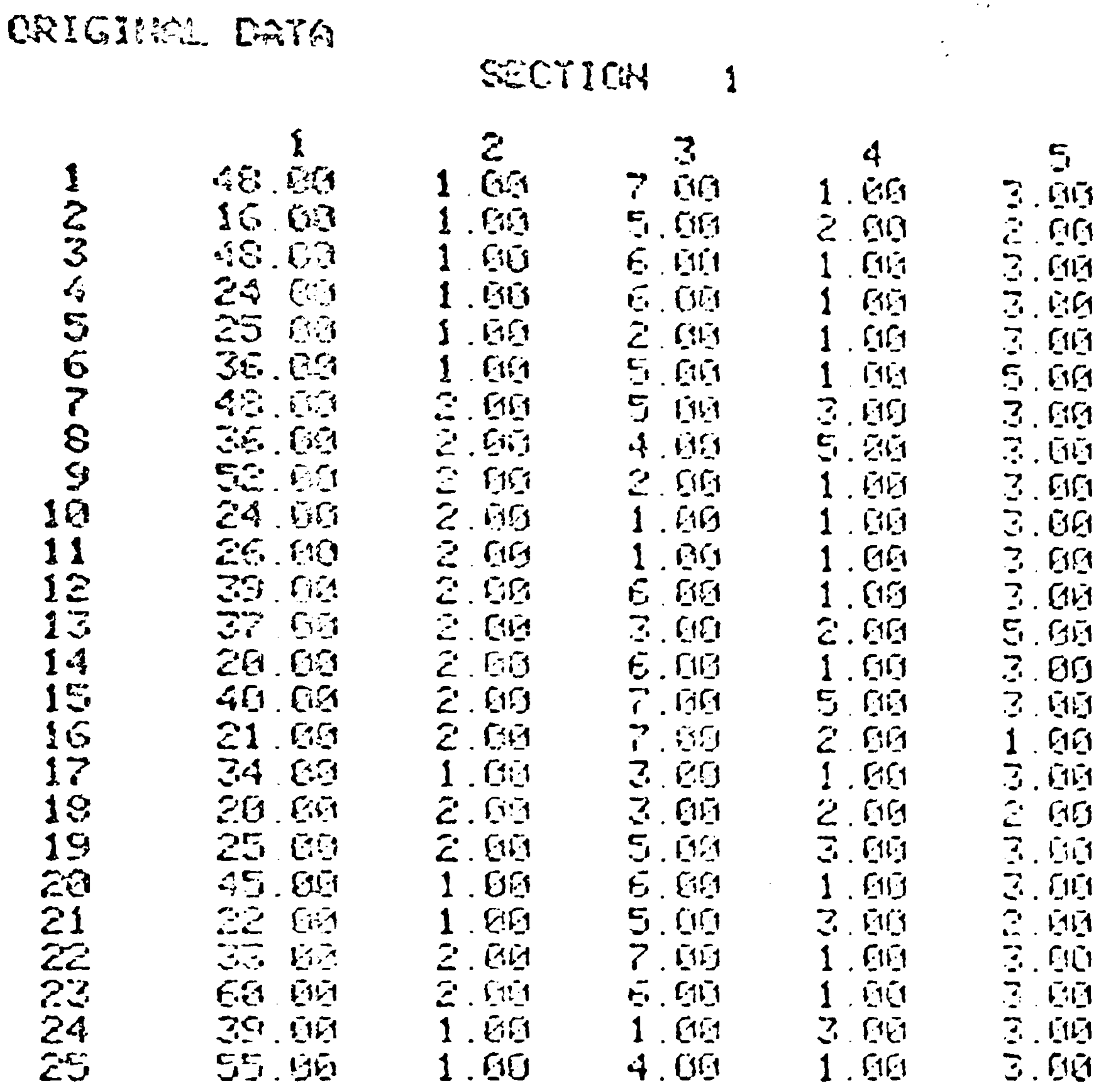
each of the possible positions (to the left and right of the i columns already placed) and compute each columns contribution to the CF. Place the column that gives the largest incremental contribution to CF in its best position. Increment i by 1 and repeat until i = N.

3. When all the columns have been placed, repeat the procedure on the rows. This step is unnecessary if the matrix is symmetrical as the row and column reorderings will be identical.

This algorithm has several important characteristics. It is quick and effective in operation. Storage requirements are only linearly related to the size of the data array. It is also applicable to matrices of any size or shape, the only restriction on the array elements being that they should not be negative.

11.5 EXAMPLE

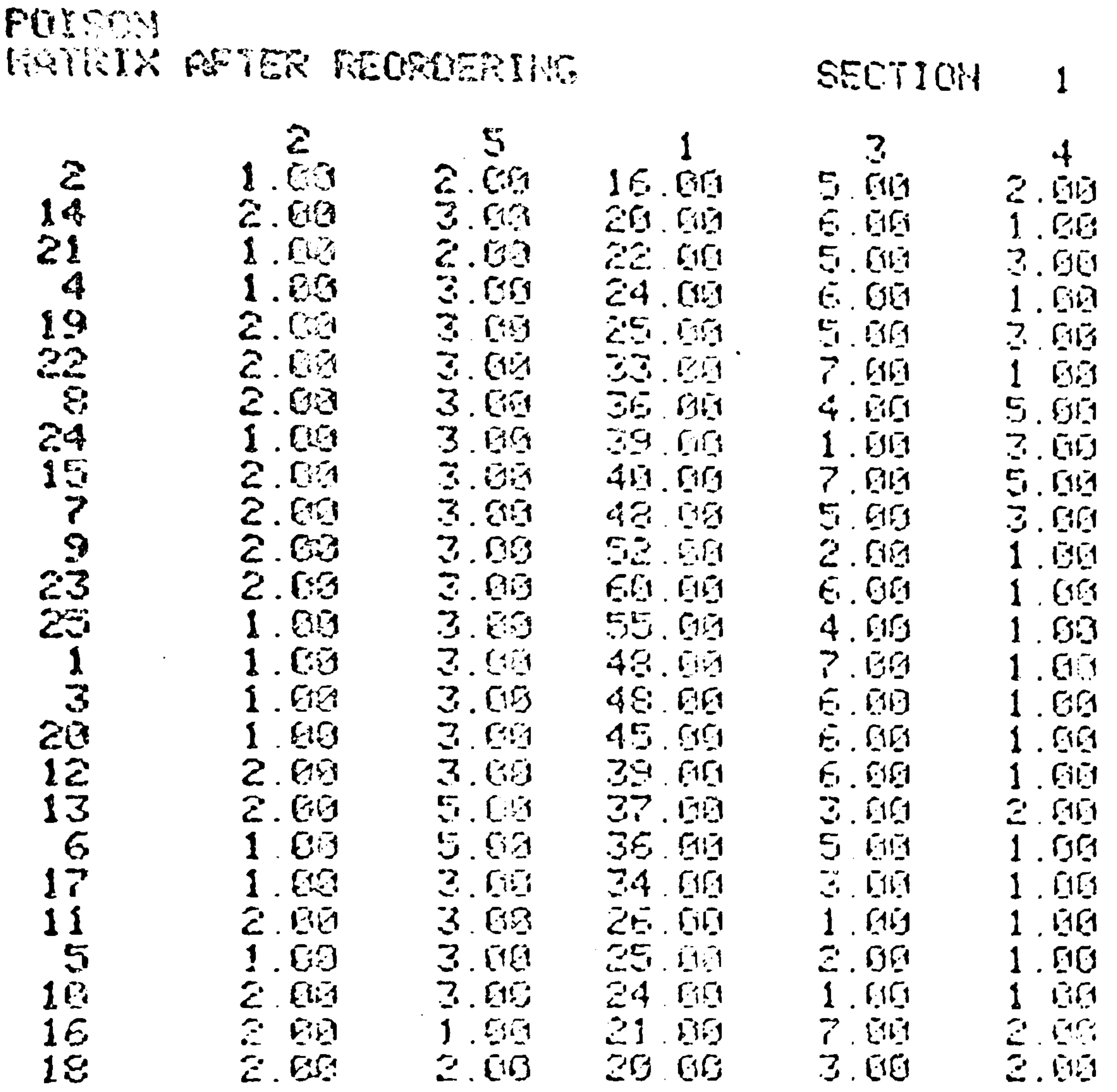
Figures 11.1 and 11.2 show the original data matrix and the transformed matrix.



```
ORIGINAL DATA
                           SECTION   1

               1          2          3          4          5
   1         48.00      1.00       7.00       1.00       3.00
   2         16.00      1.00       5.00       2.00       2.00
   3         48.00      1.00       6.00       1.00       3.00
   4         24.00      1.00       6.00       1.00       3.00
   5         25.00      1.00       2.00       1.00       3.00
   6         36.00      1.00       5.00       1.00       5.00
   7         48.00      2.00       5.00       3.00       3.00
   8         36.00      2.00       4.00       5.00       3.00
   9         52.00      2.00       2.00       1.00       3.00
  10         24.00      2.00       1.00       1.00       3.00
  11         26.00      2.00       1.00       1.00       3.00
  12         39.00      2.00       6.00       1.00       3.00
  13         37.00      2.00       3.00       2.00       5.00
  14         20.00      2.00       6.00       1.00       3.00
  15         40.00      2.00       7.00       5.00       3.00
  16         21.00      2.00       7.00       2.00       1.00
  17         34.00      1.00       3.00       1.00       3.00
  18         20.00      2.00       3.00       2.00       2.00
  19         25.00      2.00       5.00       3.00       3.00
  20         45.00      1.00       6.00       1.00       3.00
  21         22.00      1.00       5.00       3.00       2.00
  22         33.00      2.00       7.00       1.00       3.00
  23         60.00      2.00       6.00       1.00       3.00
  24         39.00      1.00       1.00       3.00       3.00
  25         55.00      1.00       4.00       1.00       3.00
```

Figure 11.1

Original data matrix



```
POISON
MATRIX AFTER REORDERING                     SECTION    1

                 2        5         1         3         4
    2          1.00     2.00     16.00      5.00      2.00
   14          2.00     3.00     20.00      6.00      1.00
   21          1.00     2.00     22.00      5.00      3.00
    4          1.00     3.00     24.00      6.00      1.00
   19          2.00     3.00     25.00      5.00      3.00
   22          2.00     3.00     33.00      7.00      1.00
    8          2.00     3.00     36.00      4.00      5.00
   24          1.00     3.00     39.00      1.00      3.00
   15          2.00     3.00     40.00      7.00      5.00
    7          2.00     3.00     42.00      5.00      3.00
    9          2.00     3.00     52.00      2.00      1.00
   23          2.00     3.00     60.00      6.00      1.00
   25          1.00     3.00     55.00      4.00      1.00
    1          1.00     3.00     48.00      7.00      1.00
    3          1.00     3.00     48.00      6.00      1.00
   20          1.00     3.00     45.00      6.00      1.00
   12          2.00     3.00     39.00      6.00      1.00
   13          2.00     5.00     37.00      3.00      2.00
    6          1.00     5.00     36.00      5.00      1.00
   17          1.00     3.00     34.00      3.00      1.00
   11          2.00     3.00     26.00      1.00      1.00
    5          1.00     3.00     25.00      2.00      1.00
   10          2.00     3.00     24.00      1.00      1.00
   16          2.00     1.00     21.00      7.00      2.00
   18          2.00     2.00     20.00      3.00      2.00
```

Figure 11.2

Reordered data matrix

CHAPTER 12

EXAMPLES OF USE

12.1 INTRODUCTION

Two examples of the use of MAGIC are presented. The first is in the context of replanning a traditional cellular office organisation into a restructured open plan layout within the existing but upgraded office building The second shows a variety of applications in the post-occupancy evaluation of another building.

The replanned office is the Central Accounting Office (CAO) of the Eastern Electricity Board, in Ipswich. The post-occupancy evaluation was carried out on the MRC Genetics Building at the Western General Hospital, Edinburgh: the use of MAGIC in this context forms part of a more complete evaluation reported by Markus and Aylward (1980).

12.2 REPLANNING OF EASTERN ELECTRICITY CAO

The Eastern Electricity Board is the largest of the English area boards, and covers the area shown in figure 12.1. The range of services provided and the Boards organisation is outlined in figure 12.2. This example of MAGIC concerns the replanning of office facilities within the Central Accounting Office, and is a straightforward example of layout planning.

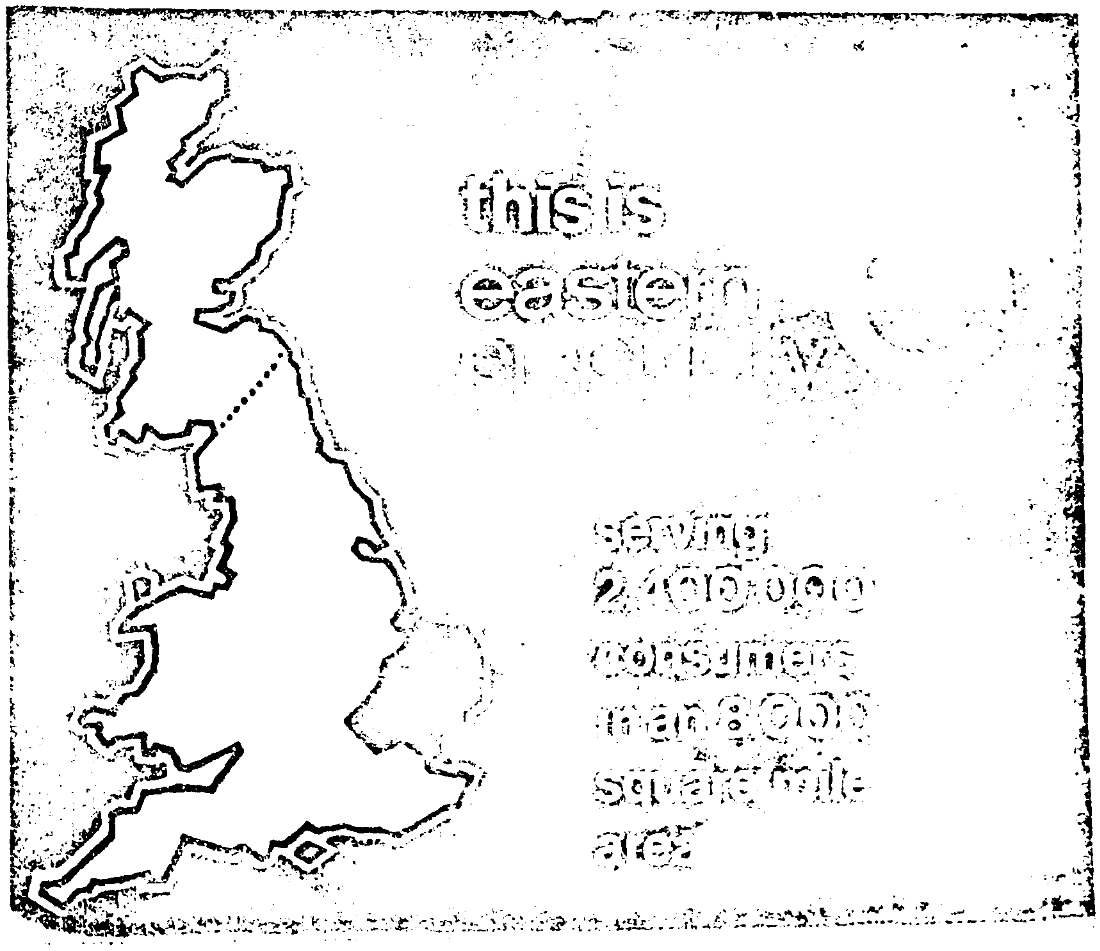
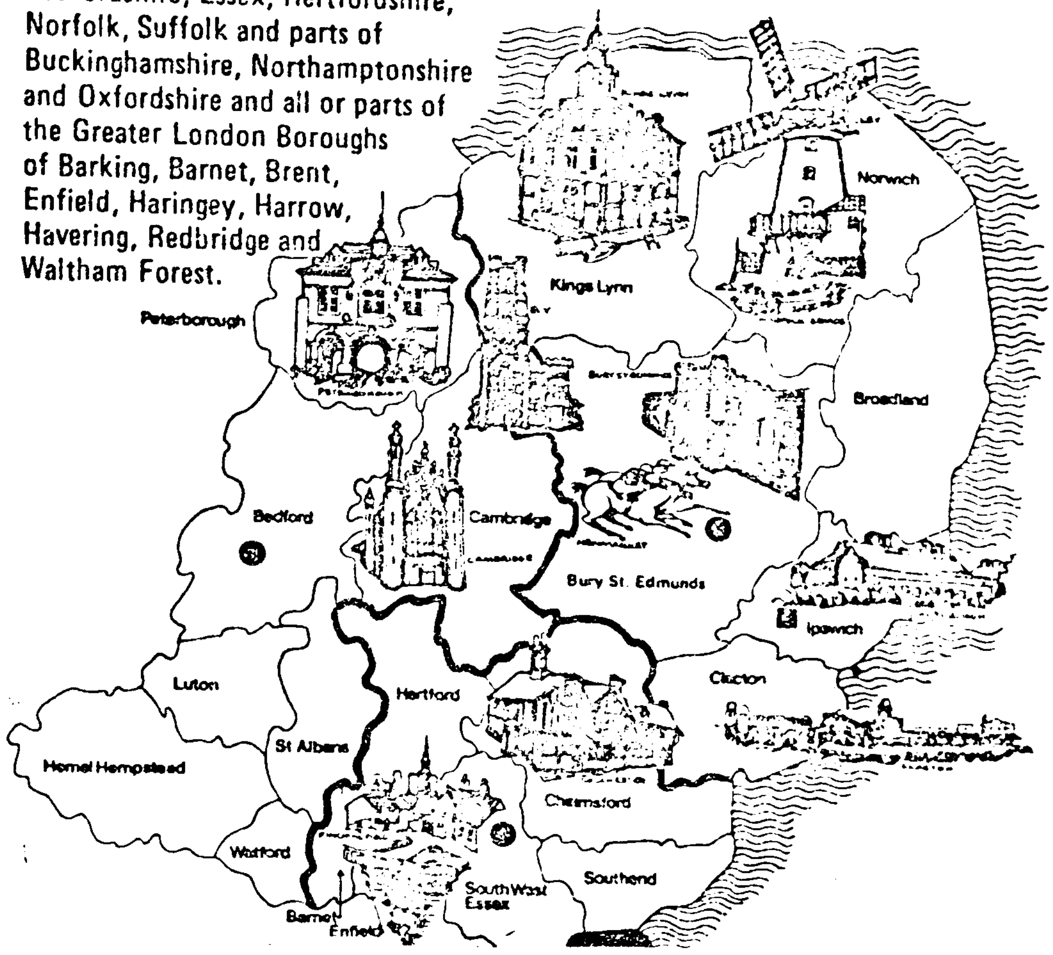


Figure 12.1

Eastern Electricity Board location map

The area covers Cambridgeshire, Bedfordshire, Essex, Hertfordshire, Norfolk, Suffolk and parts of Buckinghamshire, Northamptonshire and Oxfordshire and all or parts of the Greater London Boroughs of Barking, Barnet, Brent, Enfield, Haringey, Harrow, Havering, Redbridge and Waltham Forest.



Eastern Electricity's organisation includes

1	HEADQUARTERS
2	CENTRAL SERVICE UNITS
1	CENTRAL ACCOUNTING OFFICE
3	GROUP OFFICES
19	DISTRICT OFFICES
120	SHOPS

Staff	INDUSTRIAL	5500
	NON-INDUSTRIAL	5400
	TRAINEES & APPRENTICES	280
	MANAGEMENT	70
		11250

The electricity supply network comprises 76,000 km of overhead lines and underground cables, together with 50,000 transforming points

Where we fit in

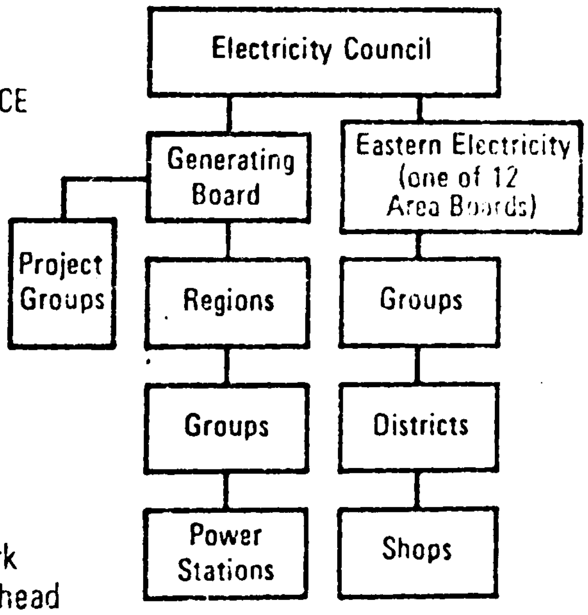


Figure 12.2

Eastern Electricity Board organisation

Layout planning may be divided into four basic stages - data collection; analysis of information; diagrammatic representation of relationships; and the translation of the diagrams into the final layout.

Data collection usually entails some form of survey. The items covered include:

(i) the main organisation and its departmental functions

(ii) activities within departments

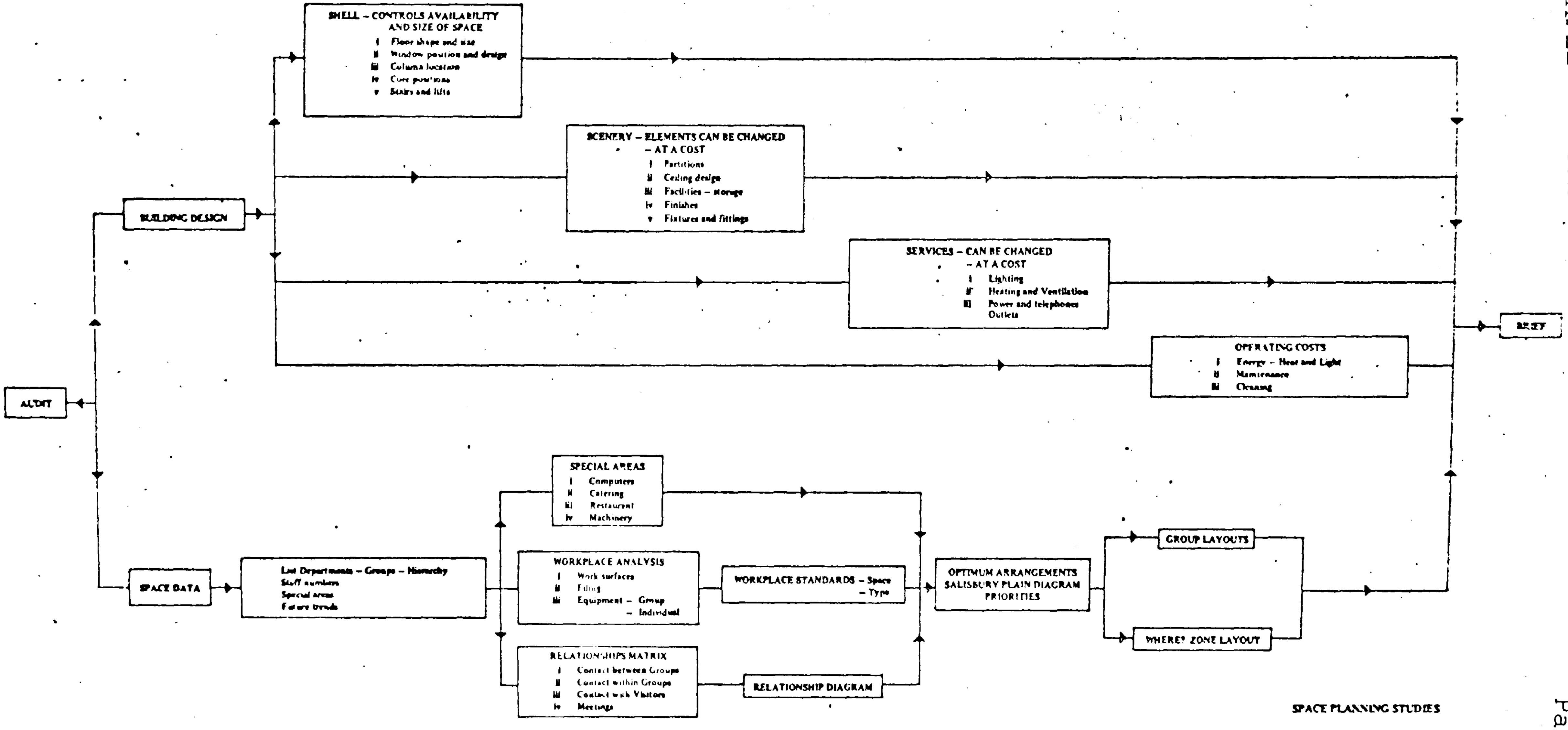
(iii) group working

(iv) each individuals (or groups) activities, including basic space allocation, equipment and furniture requirements, etc.

(v) communications - personal, telephone, and paper flow patterns

(vi) storage requirements - personal, group, central file, archives.

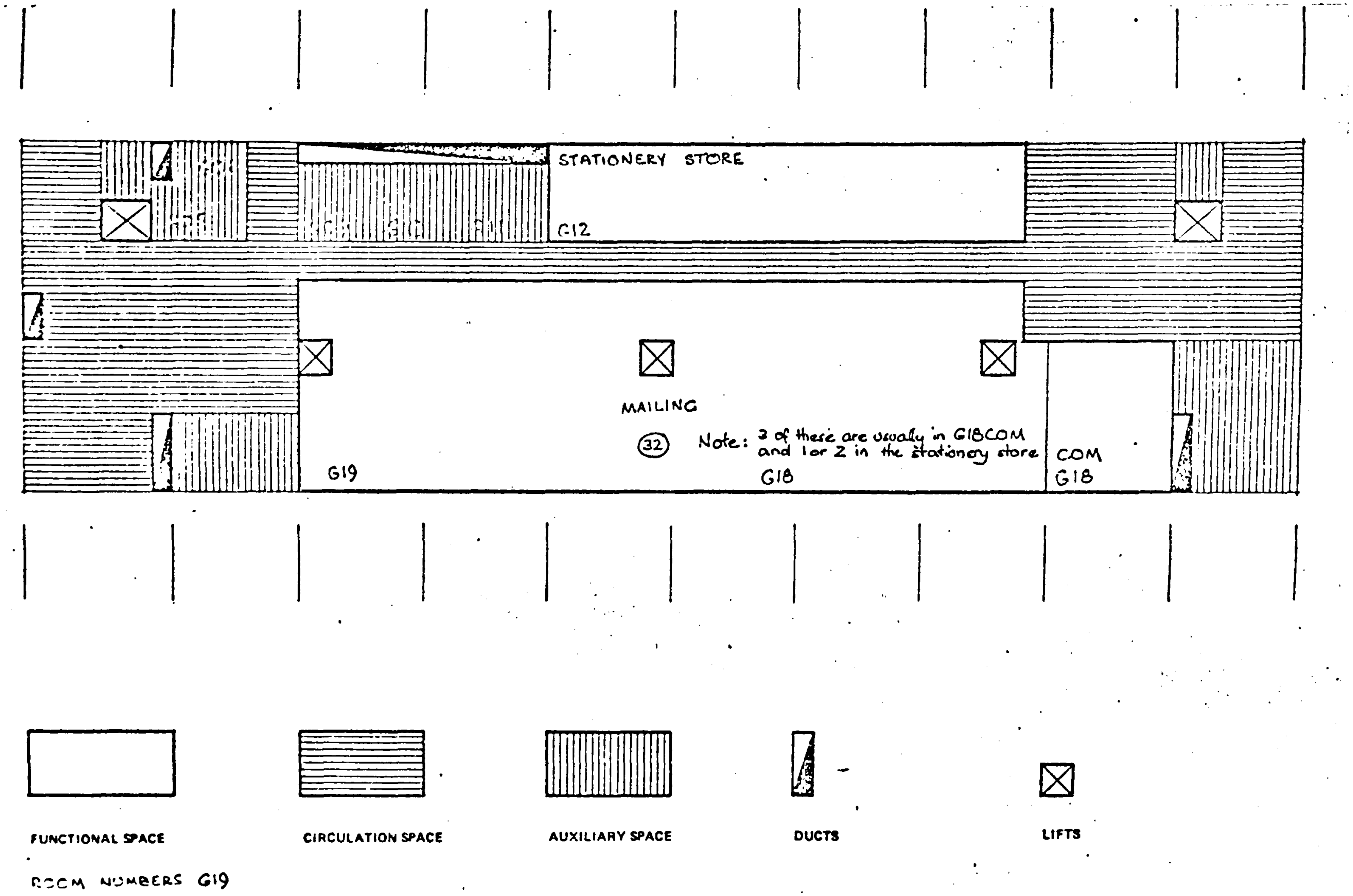
Because of very strict time constraints and important physical restrictions a number of a priori decisions were made. The "reanimated" office was to be open-plan and the major departments were to be left as the main functional divisions. This last constraint is quite reasonable when the physical problems of reorganisation are considered. The Chief Accountant's Department handles the billing of 2.5 million consumers and so utilises much heavy mailing equipment. The Management and Computer Services Department is obviously affected by the position of the computer machine room, which is fixed. The other major decision was that the building design (survey of space, services, etc.) should be carried out at the same time, and in parallel to, the space data collection and planning. The flow chart of the space planning study is shown in figure 12.3. It was agreed that the use of MAGIC should be to determine optimum arrangements in the form of "Salisbury Plain" diagrams and the detailed group and zone layouts would be carried out manually. The outline zone layouts of the building as existing are shown in figures 12.4 to 12.9.



SPACE PLANNING STUDIES

Figure 12.3

Space planning flowchart



Note: 3 of these are usually in G18COM and 1or 2 in the stationery store

Figure 12.4

Existing ground floor plan

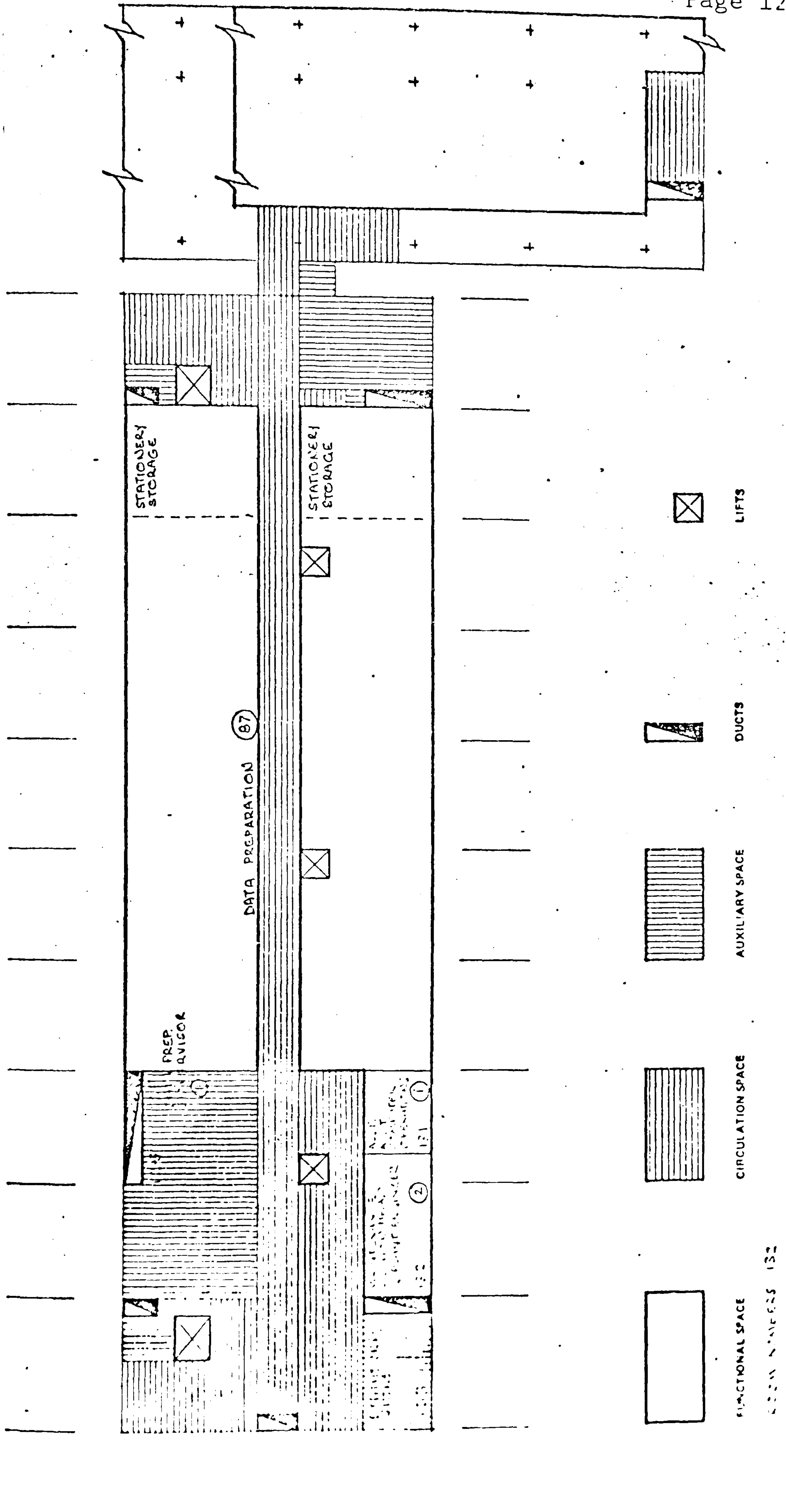


Figure 12.5

Existing first floor plan

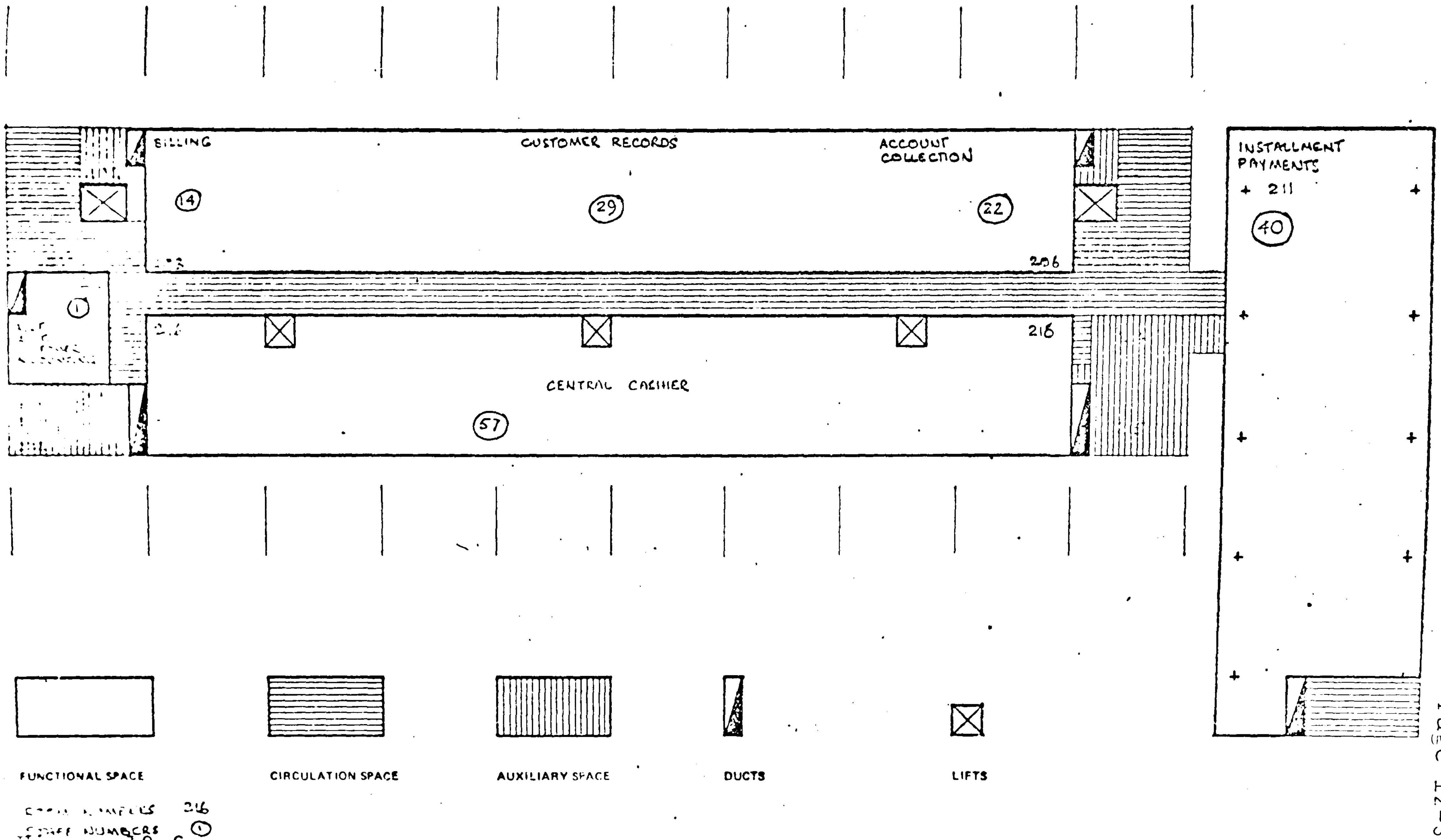


Figure 12.6

Existing second floor plan

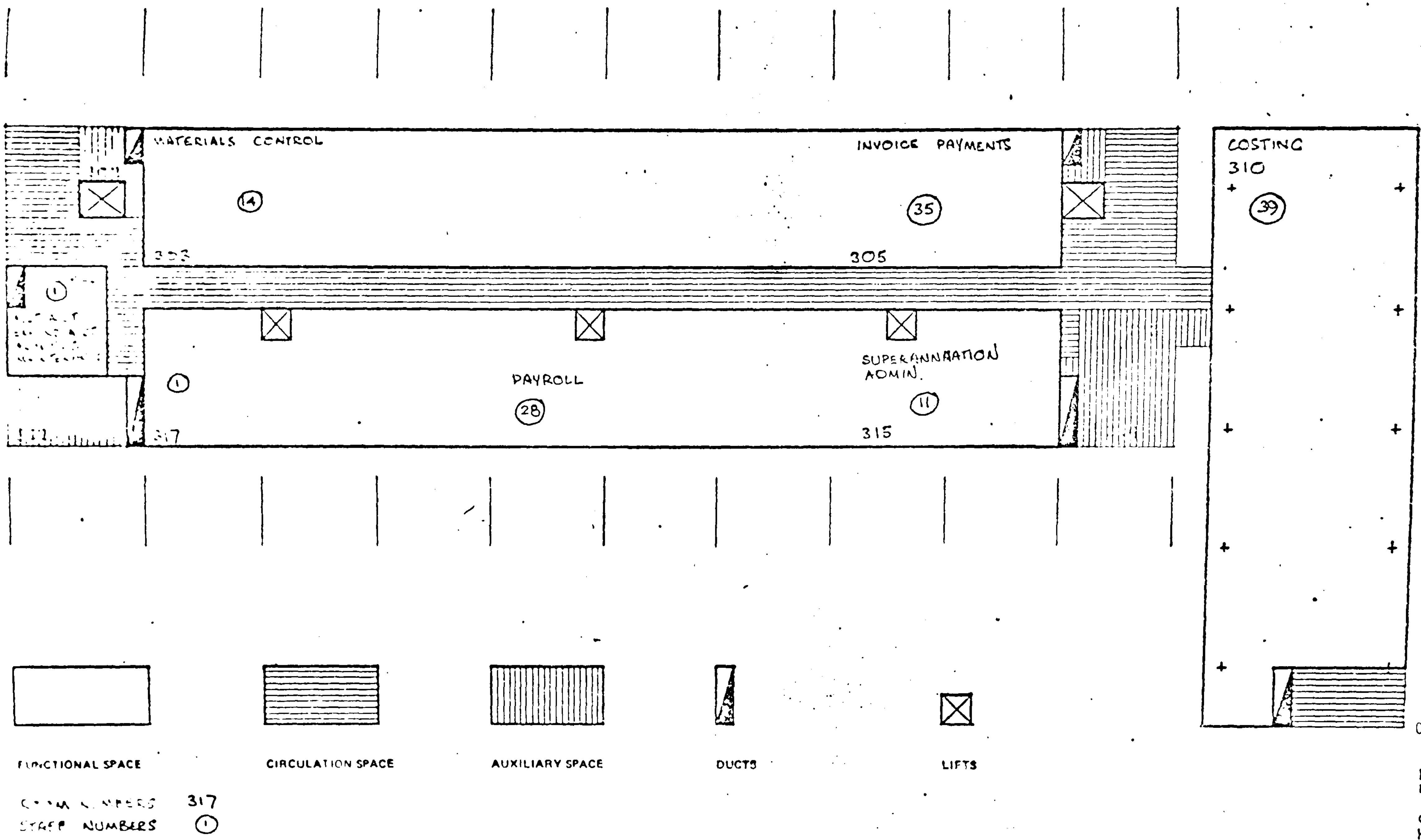


Figure 12.7

Existing third floor plan

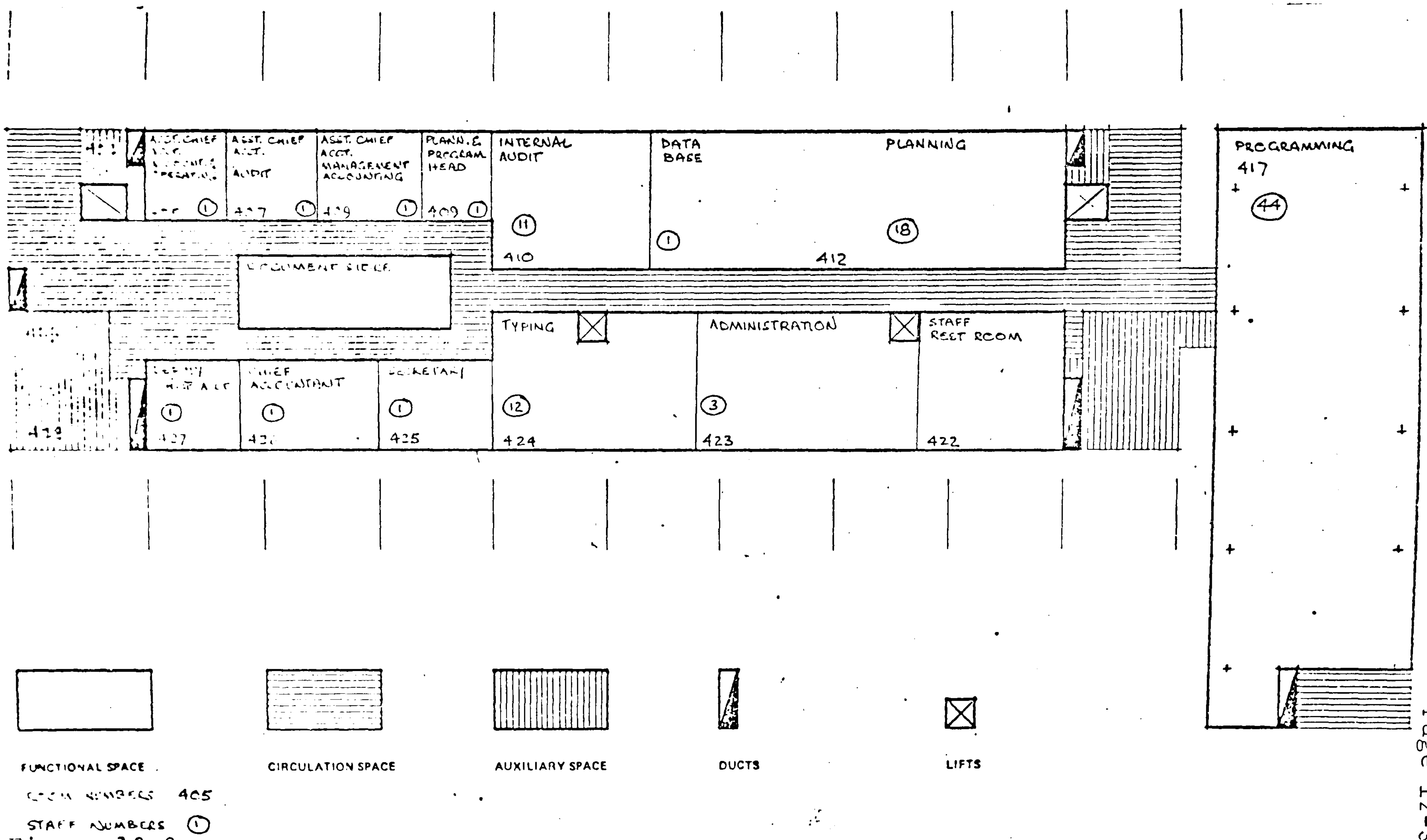


Figure 12.8

Existing fourth floor plan

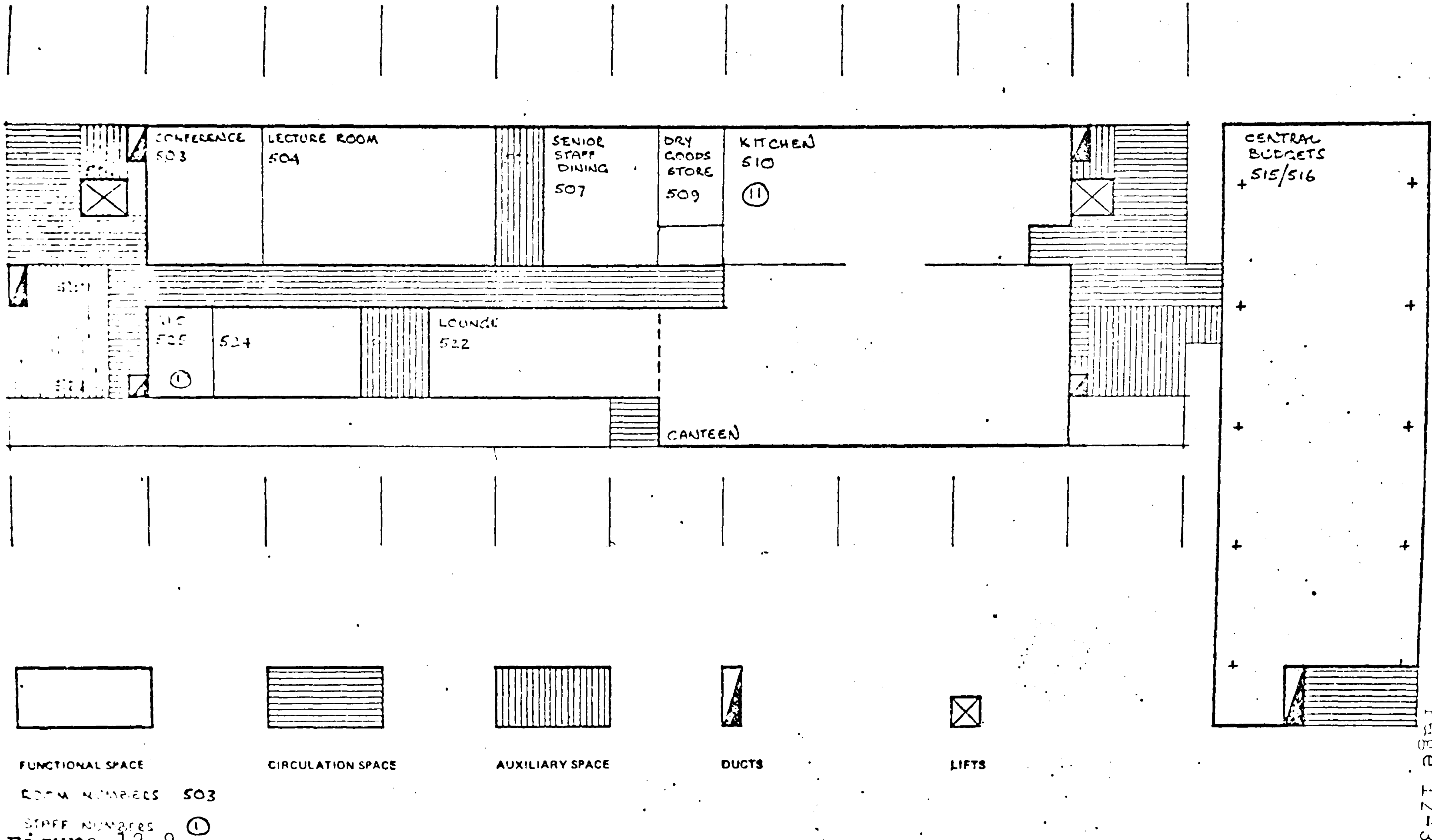


Figure 12.9

Existing fifth floor plan

12.2.1 The Activity Survey

The Chief Accounts Department and the Management and Computer Services Department were each surveyed seperately. A matrix of interrelationships for each department was drawn up, and individual activity information on space requirements, storage, equipment, etc., collected. This information is summarised here: full details are to be found in Bridges (1978). A master activity matrix was prepared for each department and activity data sheets compiled. Figure 12.10 shows the matrix for the Computer Services Department and figures 12.11 and 12.12 show typical data sheets.

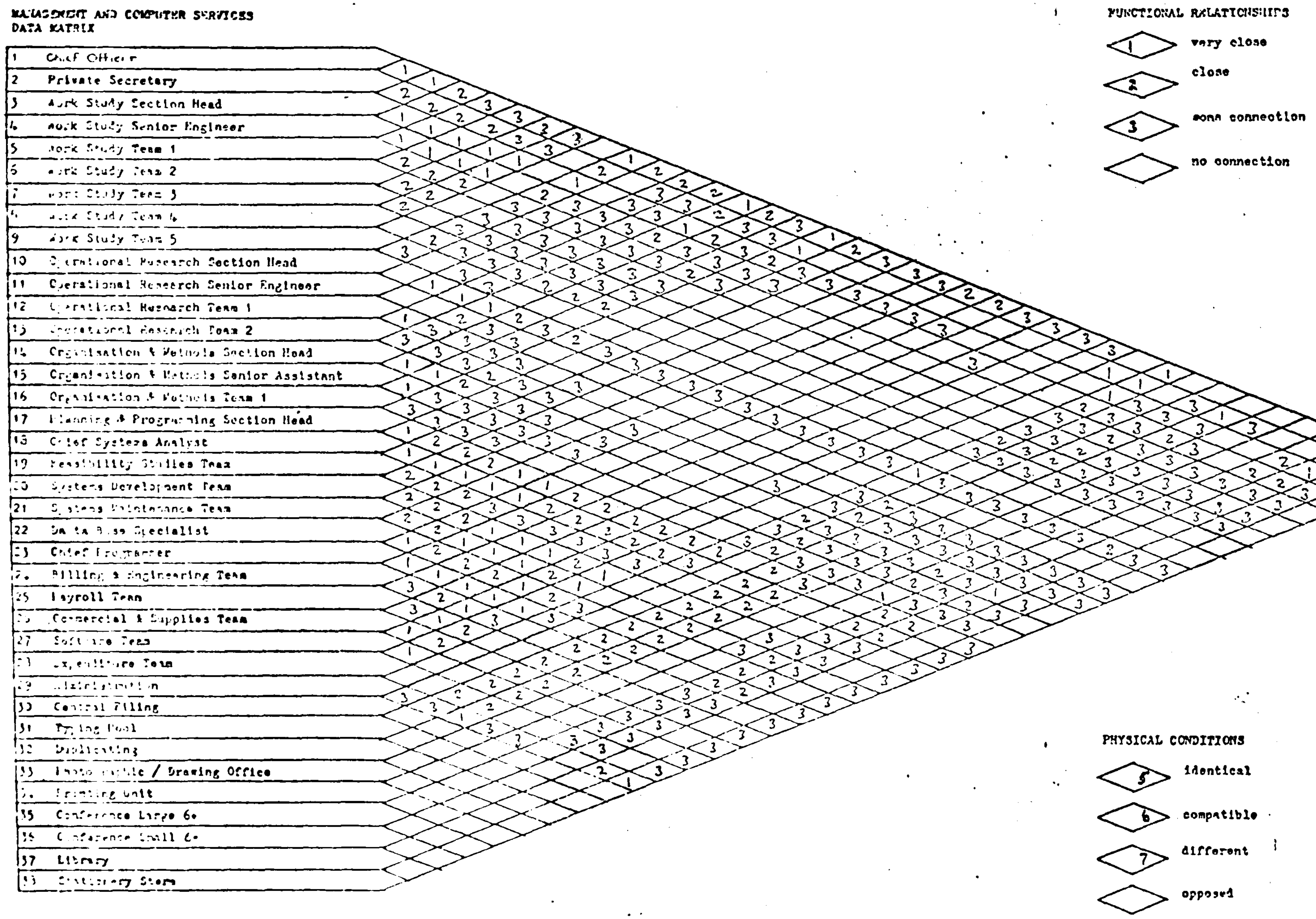
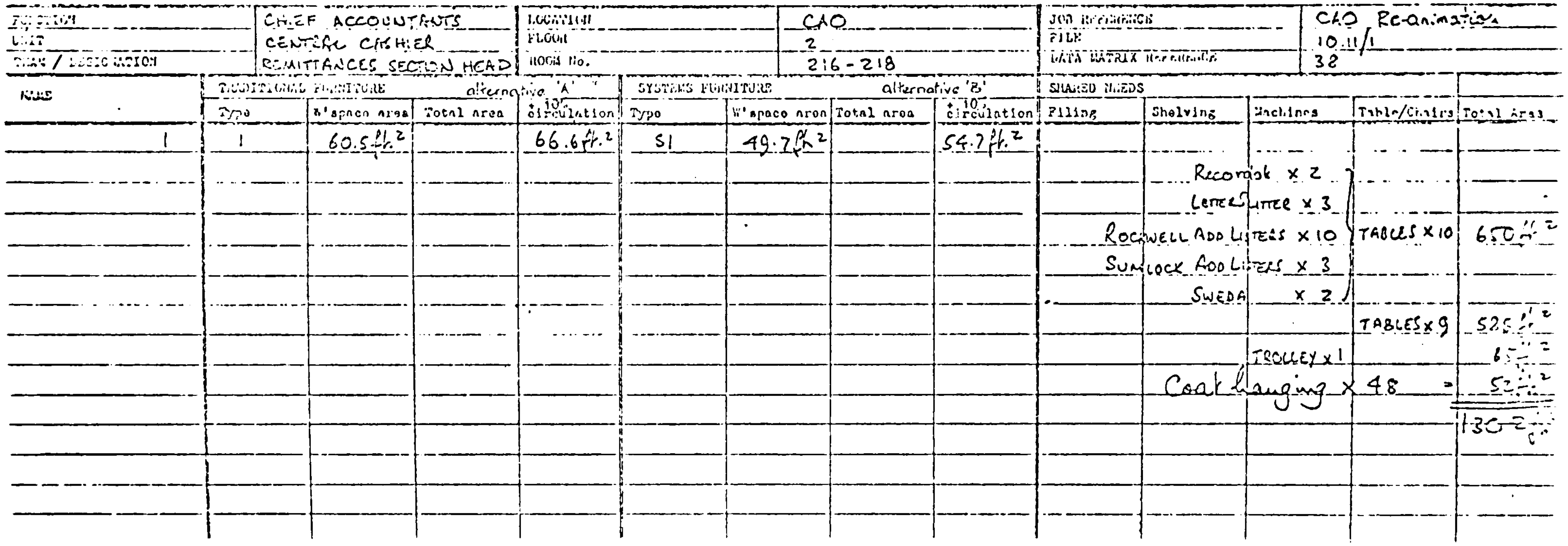

MANAGEMENT AND COMPUTER SERVICES
DATA MATRIX
1 Chief Officer
2 Private Secretary
3 Work Study Section Head
4 Work Study Senior Engineer
5 Work Study Team 1
6 Work Study Team 2
7 Work Study Team 3
8 Work Study Team 4
9 Work Study Team 5
10 Operational Research Section Head
11 Operational Research Senior Engineer
12 Operational Research Team 1
13 Operational Research Team 2
14 Organisation & Methods Section Head
15 Organisation & Methods Senior Assistant
16 Organisation & Methods Team 1
17 Planning & Programming Section Head
18 Chief Systems Analyst
19 Feasibility Studies Team
20 Systems Development Team
21 Systems Maintenance Team
22 Data Base Specialist
23 Chief Programmer
24 Billing & Engineering Team
25 Payroll Team
26 Commercial & Supplies Team
27 Software Team
28 Expenditure Team
29 Administration
30 Central Filing
31 Typing Pool
32 Duplicating
33 Photographic / Drawing Office
34 Printing Unit
35 Conference Large 6+
36 Conference Small 6+
37 Library
38 Stationery Store
FUNCTIONAL RELATIONSHIPS
1 very close
2 close
3 some connection
no connection
PHYSICAL CONDITIONS
5 identical
6 compatible
7 different
opposed

Figure 12.10

Management and Computer Services Department, data matrix

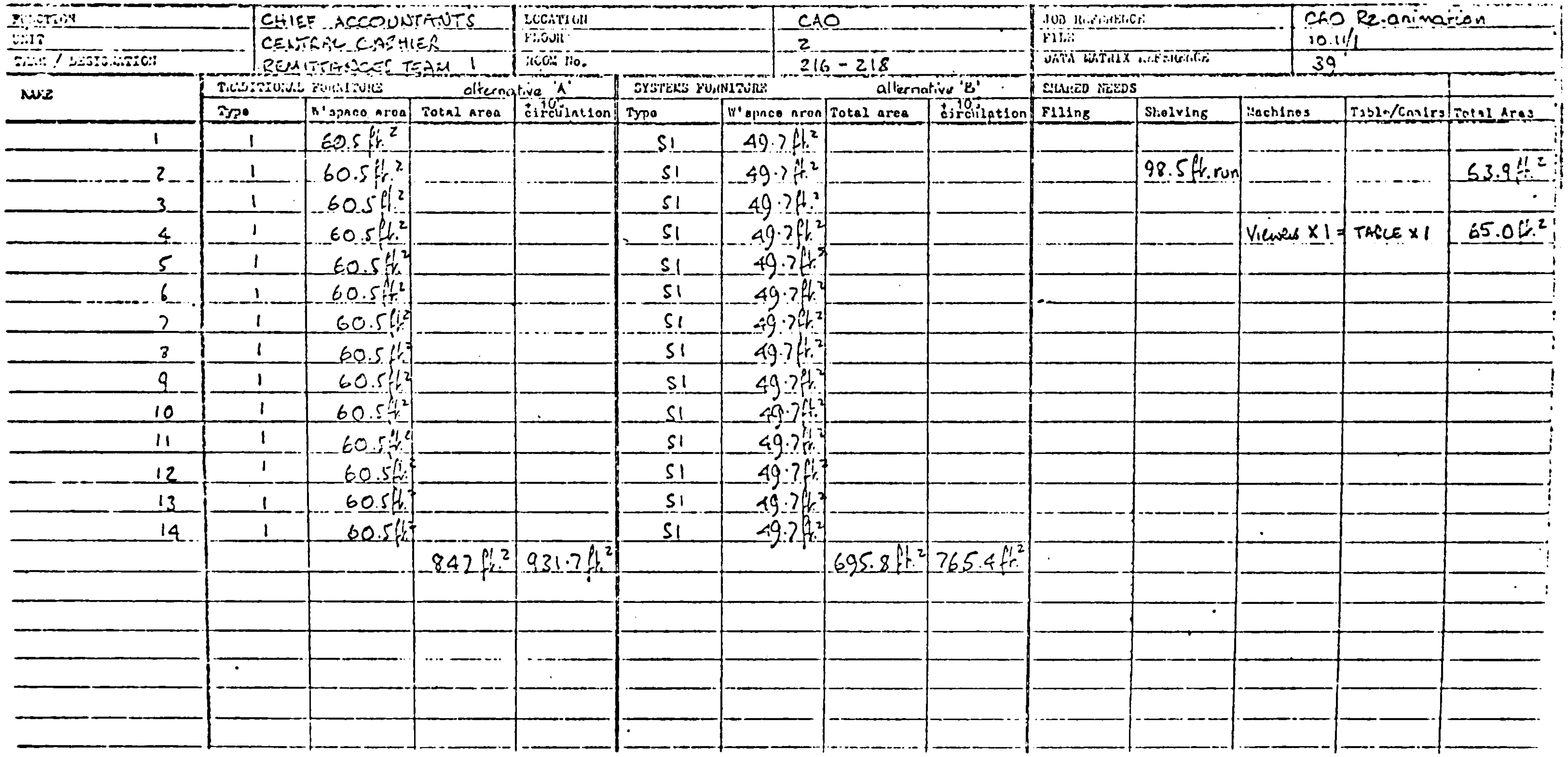


FUNCTION	CHIEF ACCOUNTANTS	LOCATION	CAO	JOB REFERENCE	CAO Re-organisation
UNIT	CENTRAL CASHIER	FLOOR	2	FILE	10.11/1
TEAM / DESIGNATION	REMITTANCES SECTION HEAD	ROOM No.	216 - 218	DATA MATRIX REFERENCE	32

NAME	TRADITIONAL FURNITURE		alternative 'A'		SYSTEMS FURNITURE		alternative 'B'		SHARED NEEDS				
	Type	W'space area	Total area	+ 10% circulation	Type	W'space area	Total area	+ 10% circulation	Filing	Shelving	Machines	Table/Chairs	Total Area
1	1	60.5 ft^2		66.6 $ft.^2$	S1	49.7 ft^2		54.7 $ft.^2$					
											Recordak x 2		
											Letter Sitter x 3		
											Rockwell Add Listers x 10	TABLES x 10	650 ft^2
											Sumlock Add Listers x 3		
											SWEDA x 2		
												TABLES x 9	525 ft^2
											TROLLEY x 1		65 ft^2
										Coat hanging x 48 =			52 ft^2
													1302 ft^2

Figure 12.11

Example room data sheet 1



FUNCTION	CHIEF ACCOUNTANTS	LOCATION	CAO
UNIT	CENTRAL CASHIER	FLOOR	2
TASK / DESIGNATION	REMITTANCES TEAM 1	ROOM No.	216 - 218

JOB REFERENCE	CAO Re-animation
FILE	10.11/1
DATA MATRIX REFERENCE	39

NAME	TRADITIONAL FURNITURE	alternative 'A'			SYSTEMS FURNITURE	alternative 'B'			SHARED NEEDS				
	Type	W'space area	Total Area	+ 10% circulation	Type	W'space area	Total area	+ 10% circulation	Filing	Shelving	Machines	Table/Chairs	Total Area
1	1	60.5ft.^{2}			S1	49.7ft.^{2}							
2	1	60.5ft.^{2}			S1	49.7ft.^{2}				98.5ft.run			53.9ft.^{2}
3	1	60.5ft.^{2}			S1	49.7ft.^{2}							
4	1	60.5ft.^{2}			S1	49.7ft.^{2}					Viewer X 1 =	TABLE x 1	65.0ft.^{2}
5	1	60.5ft.^{2}			S1	49.7ft.^{2}							
6	1	60.5ft.^{2}			S1	49.7ft.^{2}							
7	1	60.5ft.^{2}			S1	49.7ft.^{2}							
8	1	60.5ft.^{2}			S1	49.7ft.^{2}							
9	1	60.5ft.^{2}			S1	49.7ft.^{2}							
10	1	60.5ft.^{2}			S1	49.7ft.^{2}							
11	1	60.5ft.^{2}			S1	49.7ft.^{2}							
12	1	60.5ft.^{2}			S1	49.7ft.^{2}							
13	1	60.5ft.^{2}			S1	49.7ft.^{2}							
14	1	60.5ft.^{2}			S1	49.7ft.^{2}							
			847ft.^{2}	931.7ft.^{2}			695.8ft.^{2}	765.4ft.^{2}					

Figure 12.12

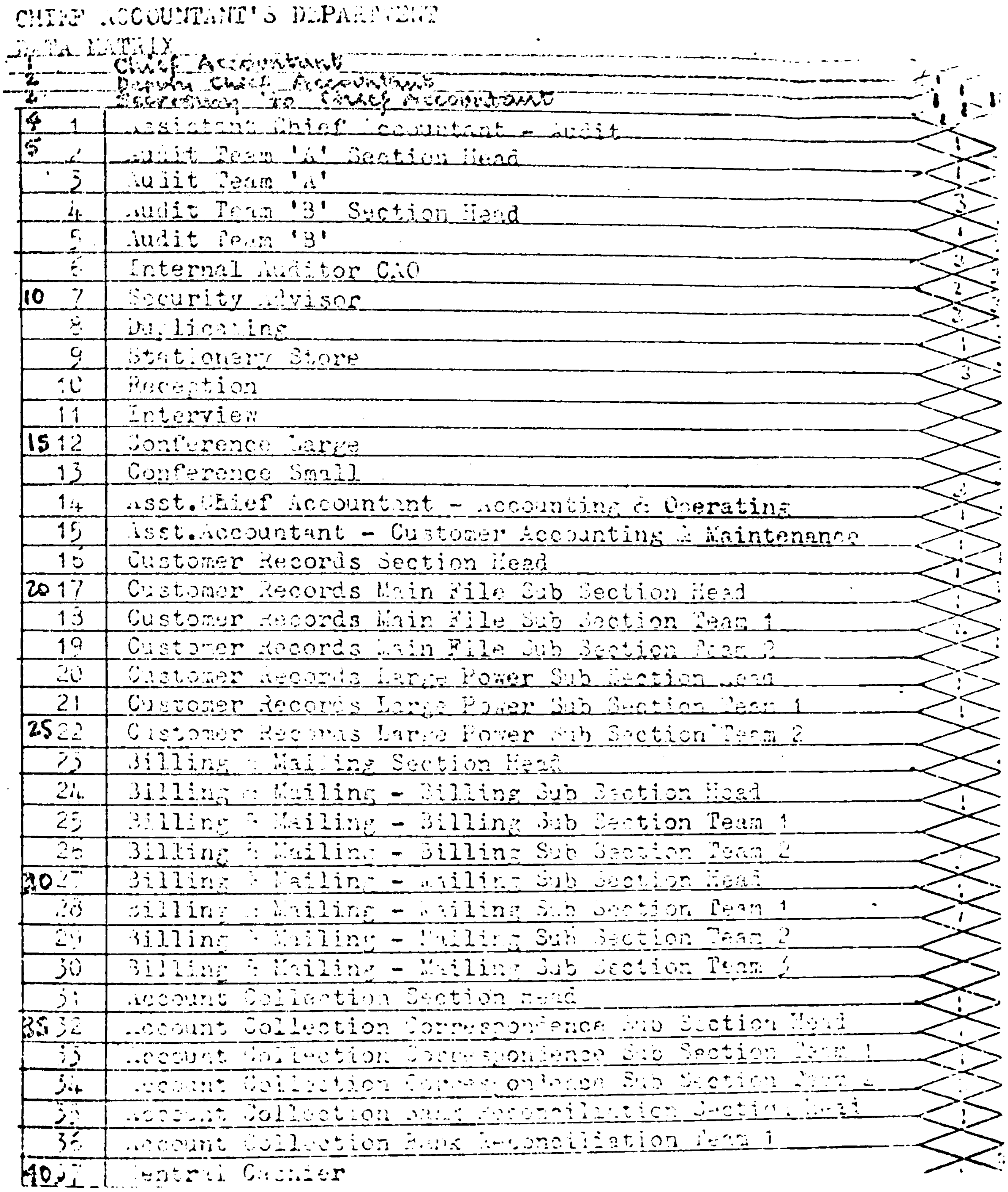
Example room data sheet 2

12.2.2 Analysis Of Chief Accountant's Department

The data matrix for the Chief Accountant's Department was analysed by Hierarchical Cluster Analysis and Nonlinear Mapping.

The "best fit" dendrogram was produced by the Group Average method and is shown here as a straightforward dendrogram plus a marked up version identifying the major administrative divisions (figures 12.15 and 12.16). The activities are numbered 1 to 100. This refers to the data matrix and are all as on that matrix up to and including activity 81; activities 82 to 90 are all grouped together as 82 - Computer Room. Activities 91 to 108 on the matrix are consequently renumbered 83 - 100. The matrix key is reproduced here, together with the grouping order list (figures 12.13 and 12.14).

The bubble diagram is drawn once without area scaling (figure 12.17) for the 100 activities used in the clustering analysis and once with area scaling (figure 12.18) for the complete 108 activities as shown on the data matrix.



CHIEF ACCOUNTANT'S DEPARTMENT

DATA MATRIX

	1	Chief Accountant
	2	Deputy Chief Accountant
	2	Secretary to Chief Accountant
4	1	Assistant Chief Accountant - Audit
5	2	Audit Team 'A' Section Head
	3	Audit Team 'A'
	4	Audit Team 'B' Section Head
	5	Audit Team 'B'
	6	Internal Auditor CAO
10	7	Security Advisor
	8	Duplicating
	9	Stationery Store
	10	Reception
	11	Interview
15	12	Conference Large
	13	Conference Small
	14	Asst.Chief Accountant - Accounting & Operating
	15	Asst.Accountant - Customer Accounting & Maintenance
	16	Customer Records Section Head
20	17	Customer Records Main File Sub Section Head
	18	Customer Records Main File Sub Section Team 1
	19	Customer Records Main File Sub Section Team 2
	20	Customer Records Large Power Sub Section Head
	21	Customer Records Large Power Sub Section Team 1
25	22	Customer Records Large Power Sub Section Team 2
	23	Billing & Mailing Section Head
	24	Billing & Mailing - Billing Sub Section Head
	25	Billing & Mailing - Billing Sub Section Team 1
	26	Billing & Mailing - Billing Sub Section Team 2
30	27	Billing & Mailing - Mailing Sub Section Head
	28	Billing & Mailing - Mailing Sub Section Team 1
	29	Billing & Mailing - Mailing Sub Section Team 2
	30	Billing & Mailing - Mailing Sub Section Team 3
	31	Account Collection Section Head
35	32	Account Collection Correspondence Sub Section Head
	33	Account Collection Correspondence Sub Section Team 1
	34	Account Collection Correspondence Sub Section Team 2
	35	Account Collection Bank Reconciliation Section Head
	36	Account Collection Bank Reconciliation Team 1
40	37	Central Cashier

Figure 12.13

Chief Accountant's Department, activity list

	38	Central Cashier Remittances Sub Section Head
	39	Central Cashier Remittances Sub Section Team 1
	40	Central Cashier Remittances Sub Section Team 2
	41	Central Cashier Remittances Sub Section Team 3
45	42	Central Cashier Cash Balancing Sub Section Head
	43	Central Cashier Cash Balancing Sub Section Team 1
	44	Central Cashier Cash Balancing Sub Section Team 2
	45	Installment Payments Section Head
	46	Installment Payments 'A' Sub Section Head
50	47	Installment Payments 'A' Sub Section Team 1
	48	Installment Payments 'A' Sub Section Team 2
	49	Installment Payments 'B' Sub Section Head
	50	Installment Payments 'B' Sub Section Team 1
	51	Installment Payments 'B' Sub Section Team 2
55	52	Installment Payments 'C' Sub Section Head
	53	Installment Payments 'C' Sub Section Team 1
	54	Installment Payments 'C' Sub Section Team 2
	55	Maintenance Section Head
	56	Maintenance Section Team
60	57	Assistant Accountant- Expenditure Accounting & Admin.
	58	Payroll Section Head
	59	Payroll 'A' Sub Section Head
	60	Payroll 'A' Sub Section Team 1
	61	Payroll 'A' Sub Section Team 2
65	62	Payroll 'B' Sub Section Head
	63	Payroll 'B' Sub Section Team 1
	64	Payroll 'B' Sub Section Team 2
	65	Invoice Payments Section Head
	66	Invoice Payments Certification 'A' Sub Section Head
70	67	Invoice Payments Certification 'A' Sub Section Team 1
	68	Invoice Payments Certification 'A' Sub Section Team 2
	69	Invoice Payments Certification 'B' Sub Section Head
	70	Invoice Payments Certification 'B' Sub Section Team 1
	71	Invoice Payments Certification 'B' Sub Section Team 2
75	72	Payments Sub Section Head
	73	Payments Sub Section
	74	Superannuation Administration Section Head
	75	Superannuation Administration Section

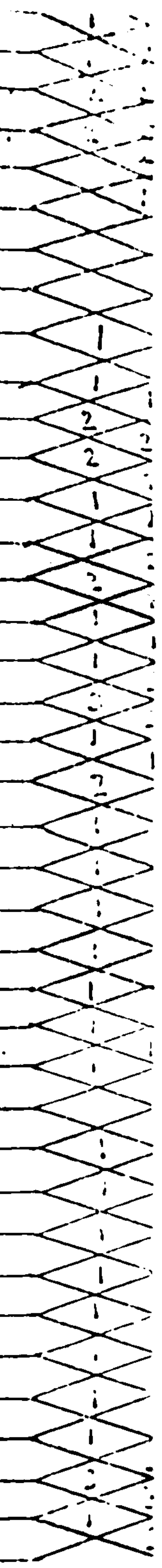


Figure 12.13

Chief Accountant's Department, activity list (continued)

	76	Administration Section Head
	80 77	Administration Section
	78	Staff Restaurant
	79	Materials Control Section Head
	80	Materials Control Section Team 1
	81	Materials Control Section Team 2
	85 82	Assistant Accountant - Computer Operating
	83	Computer Room Section Head
	84	Computer Room Shift 1
	85	Computer Room Shift 2
82	86	Computer Room Shift 3
	90 87	Data Preparation Section Head
	88	Data Preparation Keyflex 1
	89	Data Preparation Keyflex 2
	90	Data Preparation Assembly Sub Section
83	91	Data Preparation Typing Centre
84	95 92	Assistant Chief Accountant Management Accounting
85	93	Assistant Accountant Management Accounting
86	94	Central Budget Section Head 'A'
87	95	Central Budget Accounting Control Sub Section Head
88	96	Central Budget Accounting Control Sub Section
89	100 97	Central Budgetary Control - Districts
90	98	Central Budget Section Head 'B'
91	99	Central Budget - Budgetary Control Group 1
92	100	Central Budget - Budgetary Control Group 2
93	101	Costing Section Head
94	105 102	Rechargeable Section Head
95	103	Rechargeable Section Team 1
96	104	Major Costing Section Head
97	105	Major Costing Team 1
98	106	Major Costing Team 2
99	110 107	Capital Section Head
100	111 108	Capital Section Team 1

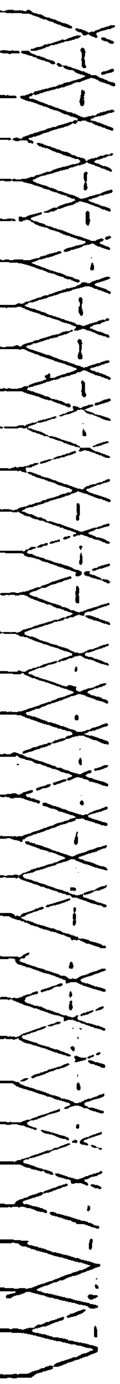
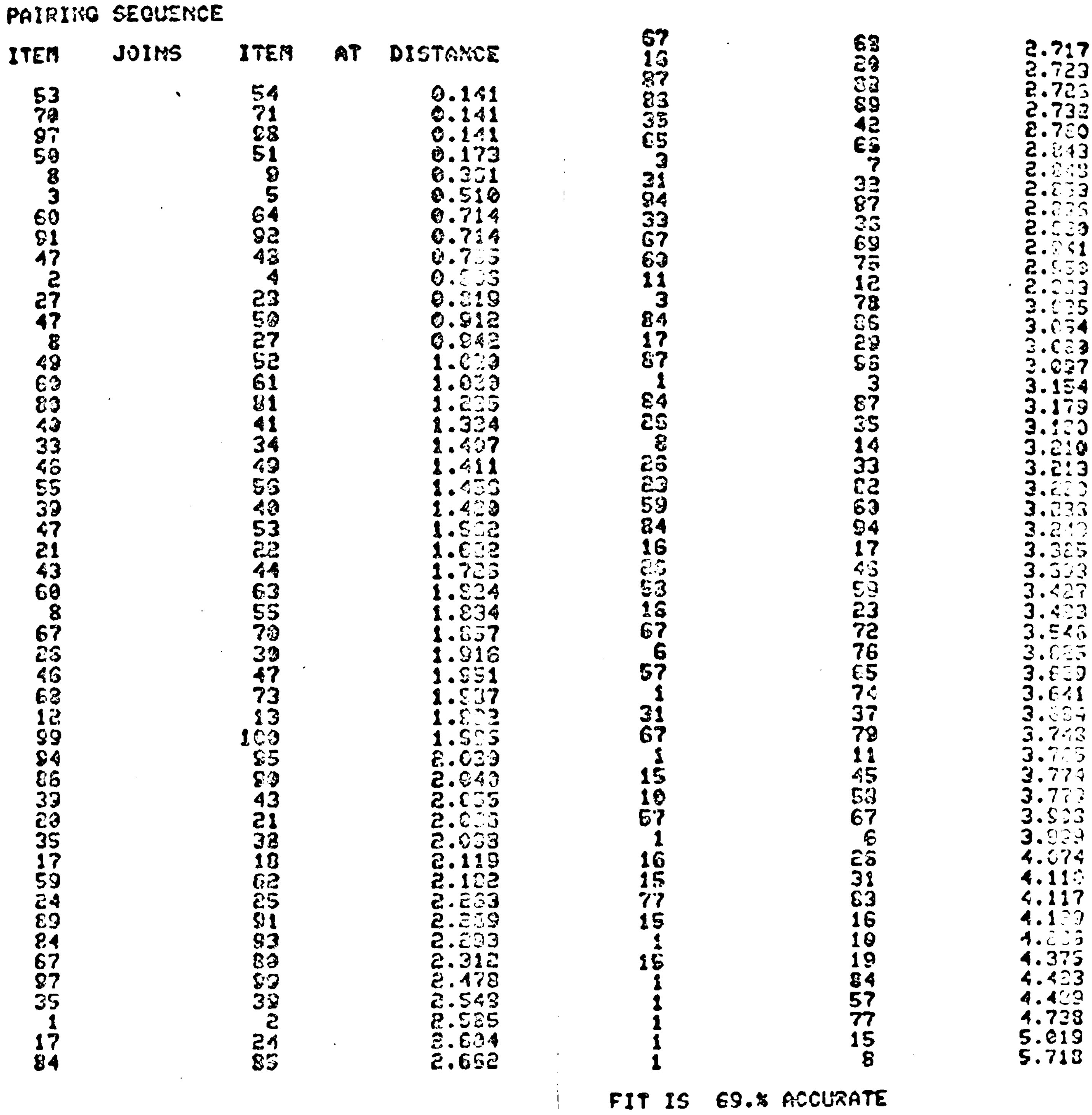


Figure 12.13

Chief Accountant's Department, activity list (continued)



```
PAIRING SEQUENCE

ITEM     JOINS     ITEM    AT   DISTANCE

 53                 54           0.141
 70                 71           0.141
 97                 98           0.141
 50                 51           0.173
  8                  9           0.301
  3                  5           0.510
 60                 64           0.714
 91                 92           0.714
 47                 48           0.755
  2                  4           0.806
 27                 23           0.819
 47                 50           0.912
  8                 27           0.942
 49                 52           1.020
 69                 61           1.030
 80                 81           1.225
 40                 41           1.334
 33                 34           1.407
 46                 49           1.411
 55                 56           1.415
 39                 40           1.420
 47                 53           1.502
 21                 22           1.632
 43                 44           1.726
 60                 63           1.824
  8                 55           1.834
 67                 70           1.857
 26                 30           1.916
 46                 47           1.951
 68                 73           1.957
 12                 13           1.972
 99                100           1.995
 94                 95           2.030
 86                 89           2.040
 30                 43           2.055
 20                 21           2.066
 35                 38           2.098
 17                 18           2.119
 59                 62           2.182
 24                 25           2.263
 89                 91           2.289
 84                 93           2.293
 67                 80           2.312
 97                 99           2.478
 35                 39           2.543
  1                  2           2.585
 17                 24           2.604
 84                 85           2.662

 67                 68           2.717
 16                 20           2.723
 87                 88           2.725
 83                 89           2.732
 35                 42           2.780
 65                 66           2.843
  3                  7           2.848
 31                 32           2.853
 94                 87           2.856
 33                 36           2.859
 67                 69           2.941
 69                 75           2.950
 11                 12           2.983
  3                 78           3.025
 84                 86           3.054
 17                 29           3.083
 87                 96           3.097
  1                  3           3.154
 84                 87           3.179
 26                 35           3.190
  8                 14           3.210
 26                 33           3.213
 23                 62           3.220
 59                 63           3.236
 84                 94           3.240
 16                 17           3.325
 26                 46           3.393
 53                 59           3.427
 16                 23           3.483
 67                 72           3.546
  6                 76           3.625
 57                 65           3.639
  1                 74           3.641
 31                 37           3.684
 67                 79           3.748
  1                 11           3.765
 15                 45           3.774
 10                 58           3.779
 57                 67           3.903
  1                  6           3.939
 16                 26           4.074
 15                 31           4.110
 77                 83           4.117
 15                 16           4.179
  1                 10           4.236
 16                 19           4.375
  1                 84           4.423
  1                 57           4.429
  1                 77           4.738
  1                 15           5.019
  1                  8           5.718

FIT IS  69.% ACCURATE
```

Figure 12.14

Chief Accountant's Department, pairing sequence

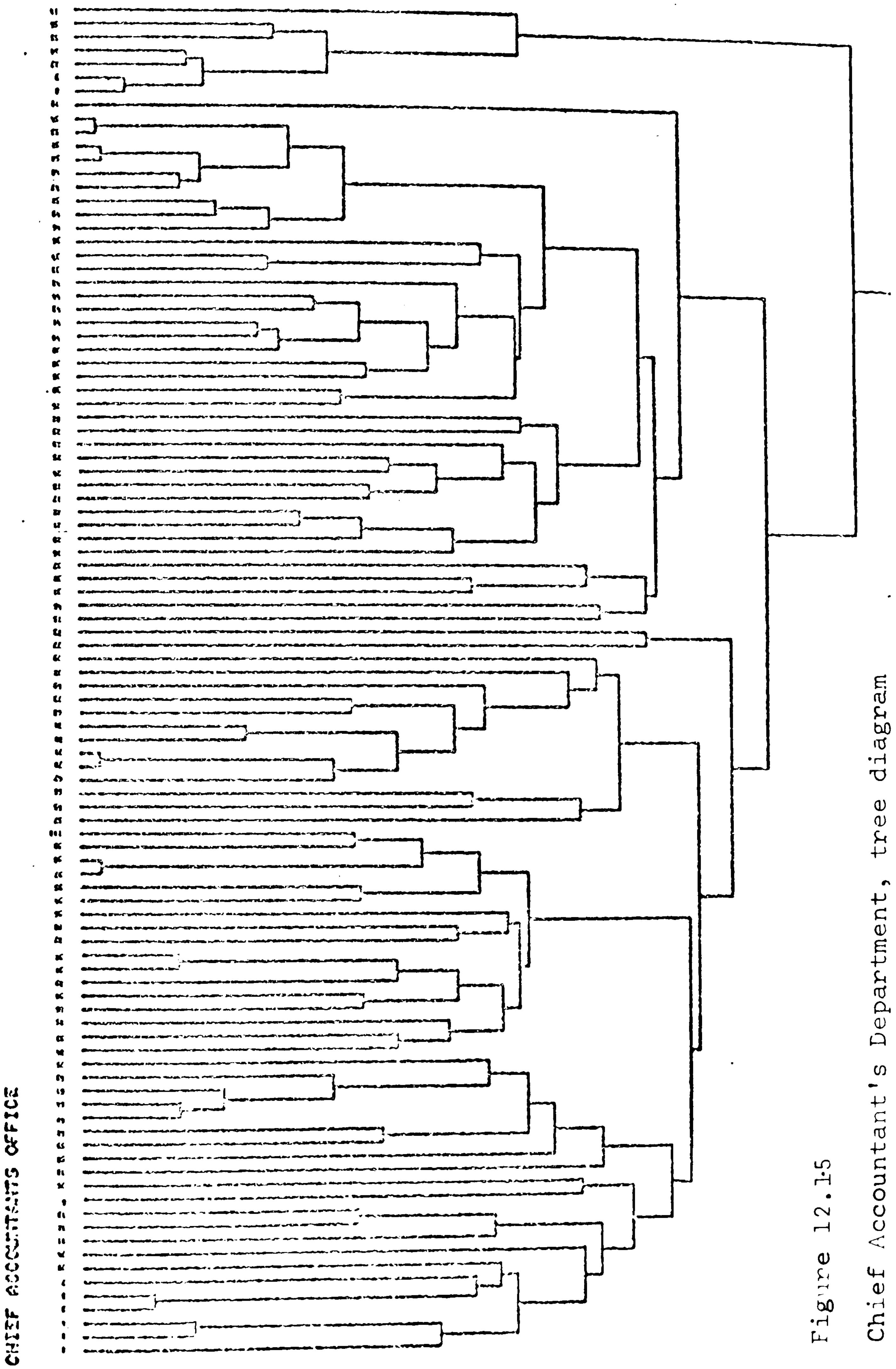


Figure 12.15

Chief Accountant's Department, tree diagram

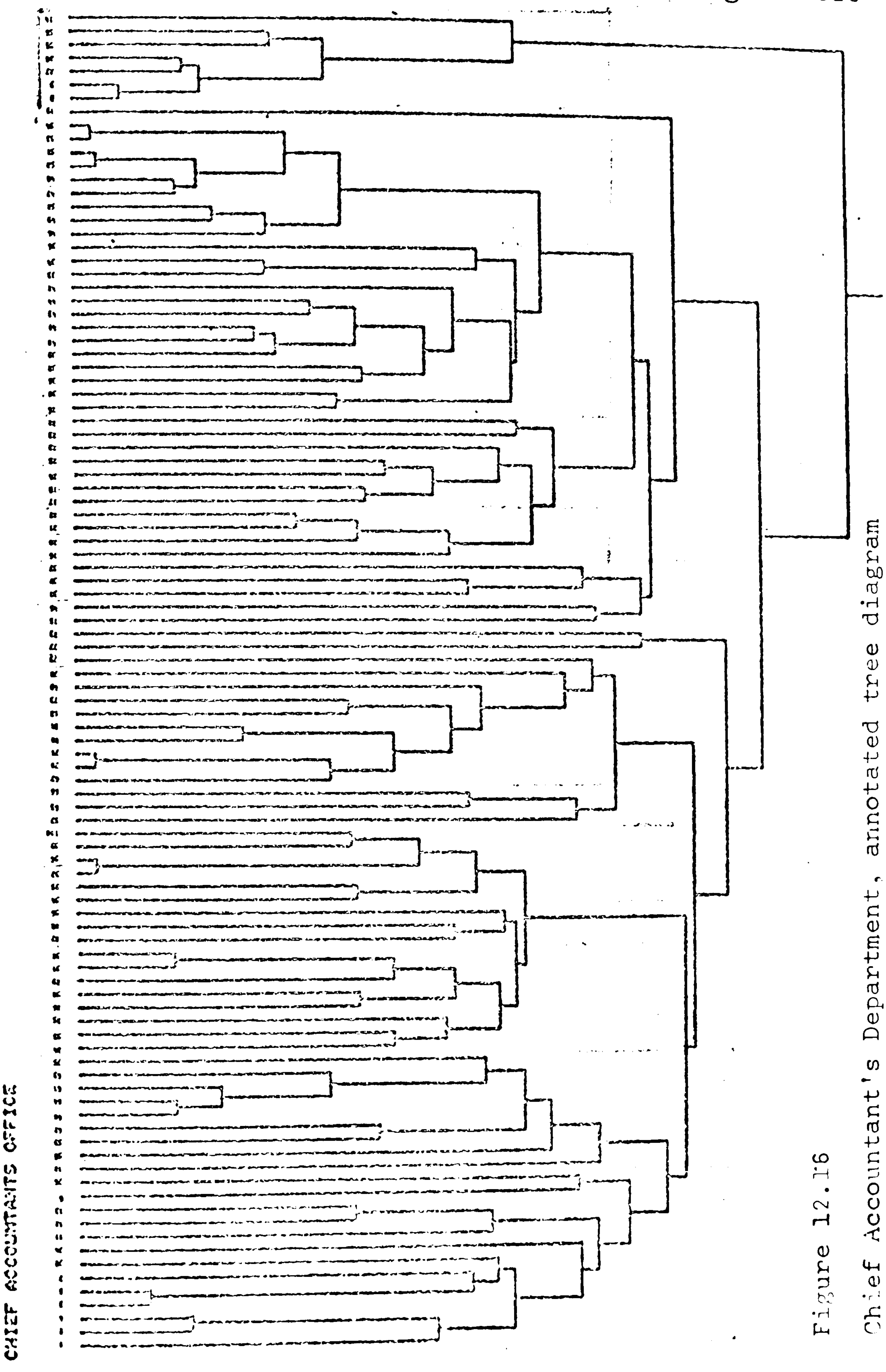


Figure 12.16

Chief Accountant's Department, annotated tree diagram

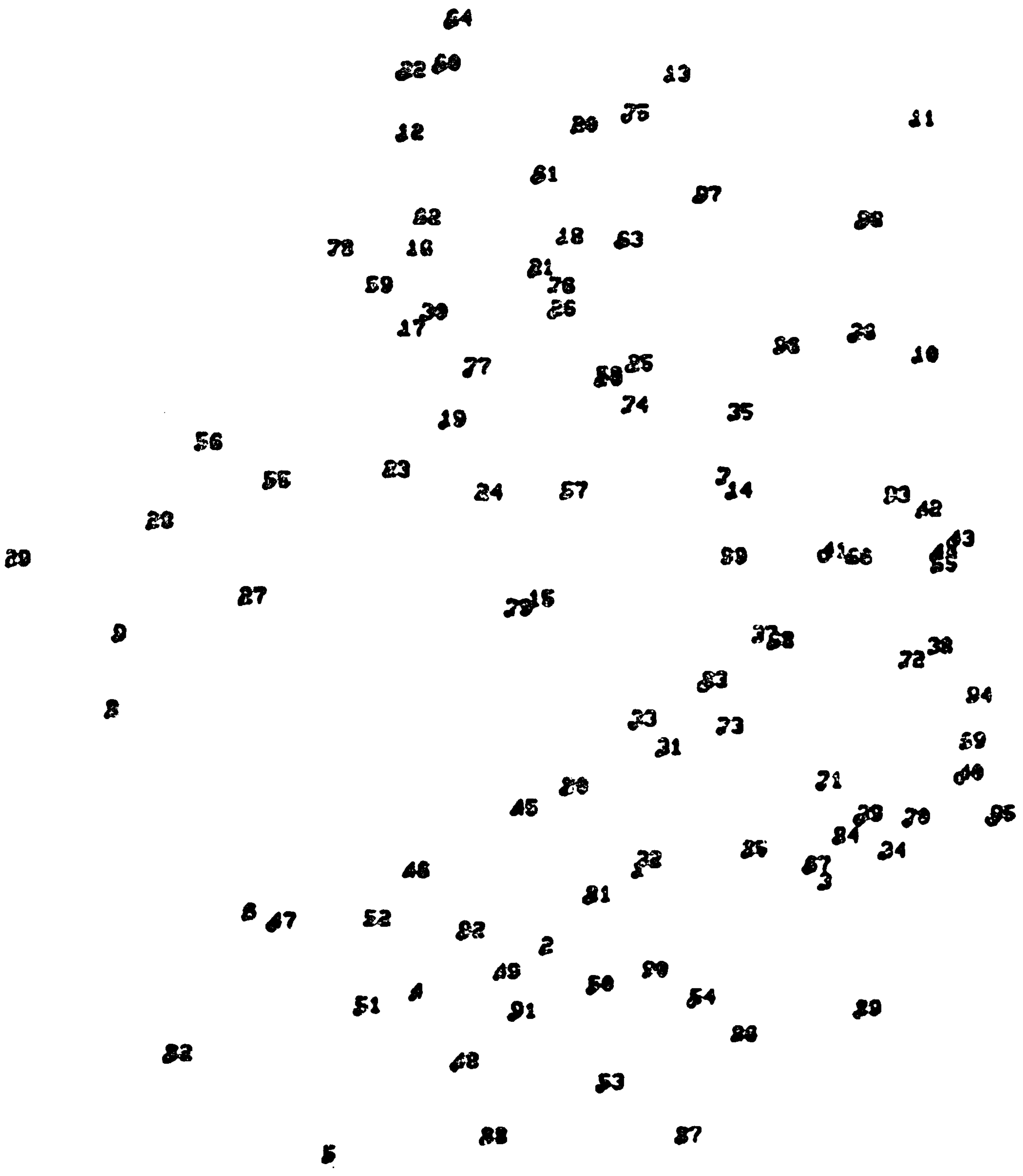


Figure 12.17

Chief Accountant's Department, bubble diagram

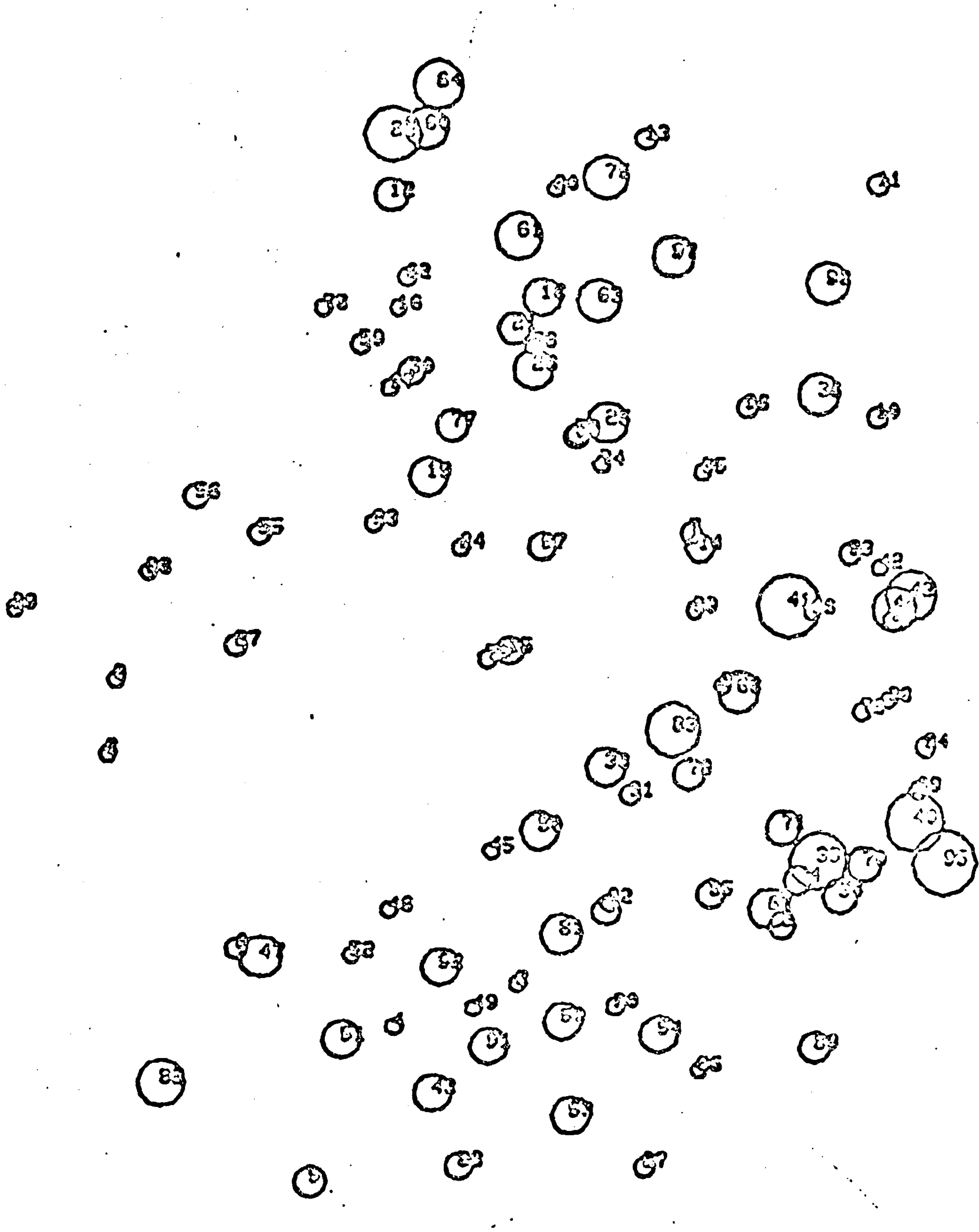


Figure 12.18

Chief Accountant's Department, bubble diagram (with areas)

12.2.3 Analysis Of Management And Computer Services Dept.

There are four different analyses of the organisational structure. Although all four are calculated differently there is a large amount of agreement between them, thus indicating some underlying structure in the data. In each case (except for the analysis into seperate groups which is self-explanatory) a possible interpretation is marked up on the printout. These interpretations have been made solely on the computed data without reference to the job title list, and thus represent the actual relationships hidden in the data and not necessarily the ones which are believed to exist.

```
MATRIX AFTER REORDERING           SECTION   1

          11     34     30     37     33      2      8      4      7      6      5     12     13
 11     0.00   0.00   0.00   0.00   0.00   0.00   3.00   3.00   3.00   3.00   3.00   0.00   0.00
 34     0.00   0.00   0.00   0.00   0.00   0.00   3.00   0.00   3.00   3.00   3.00   3.00   3.00
 30     0.00   0.00   0.00   0.00   0.00   0.00   3.00   3.00   3.00   3.00   3.00   2.00   2.00
 37     0.00   0.00   0.00   0.00   0.00   3.00   0.00   3.00   3.00   3.00   3.00   3.00   3.00
 33     0.00   0.00   0.00   0.00   0.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
  2     0.00   0.00   0.00   3.00   3.00   0.00   3.00   2.00   3.00   2.00   2.00   3.00   3.00
  8     3.00   3.00   3.00   0.00   3.00   3.00   0.00   1.00   2.00   2.00   2.00   3.00   3.00
  4     3.00   0.00   3.00   3.00   3.00   2.00   1.00   0.00   1.00   1.00   1.00   3.00   3.00
  7     3.00   3.00   3.00   3.00   3.00   3.00   2.00   1.00   0.00   2.00   2.00   3.00   3.00
  6     3.00   3.00   3.00   3.00   3.00   2.00   2.00   1.00   2.00   0.00   2.00   3.00   3.00
  5     3.00   3.00   3.00   3.00   3.00   2.00   2.00   1.00   2.00   2.00   0.00   3.00   3.00
 12     0.00   3.00   2.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   0.00   1.00
 13     0.00   3.00   2.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   1.00   0.00
 10     0.00   3.00   3.00   3.00   3.00   2.00   2.00   2.00   3.00   3.00   3.00   1.00   1.00
 15     3.00   2.00   2.00   3.00   3.00   3.00   2.00   3.00   3.00   3.00   3.00   3.00   3.00
 16     3.00   1.00   2.00   3.00   3.00   3.00   2.00   3.00   3.00   3.00   2.00   3.00   3.00
 14     2.00   2.00   3.00   3.00   3.00   2.00   2.00   2.00   3.00   3.00   3.00   3.00   3.00
 29     0.00   3.00   3.00   2.00   3.00   1.00   3.00   2.00   2.00   3.00   3.00   3.00   3.00
 32     0.00   0.00   0.00   0.00   0.00   1.00   3.00   2.00   3.00   3.00   3.00   3.00   3.00
 35     0.00   0.00   0.00   0.00   0.00   0.00   0.00   2.00   3.00   3.00   3.00   3.00   3.00
 36     0.00   0.00   0.00   0.00   0.00   0.00   0.00   2.00   3.00   3.00   3.00   3.00   3.00
  1     2.00   0.00   0.00   0.00   0.00   1.00   3.00   2.00   3.00   3.00   3.00   2.00   2.00
 31     0.00   0.00   0.00   0.00   0.00   3.00   3.00   3.00   2.00   2.00   2.00   3.00   3.00
 19     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00   3.00
 18     0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00   0.00   0.00   0.00   0.00   3.00
 17     0.00   0.00   0.00   0.00   0.00   1.00   0.00   3.00   0.00   0.00   3.00   3.00   2.00
 22     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00
 20     0.00   3.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
 21     0.00   3.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
 24     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
 26     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
 25     0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00   0.00   0.00   0.00   0.00   0.00
 27     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00
 28     0.00   0.00   0.00   0.00   0.00   1.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
 38     0.00   0.00   0.00   0.00   0.00   2.00   0.00   0.00   0.00   0.00   0.00   3.00   3.00
  3     0.00   0.00   3.00   0.00   0.00   2.00   1.00   1.00   1.00   1.00   1.00   3.00   3.00
 23     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
  9     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   3.00   3.00
```

Figure 12.19 - Management and Computer Services Department, reordered matrix

MATRIX AFTER REORDERING SECTION 2

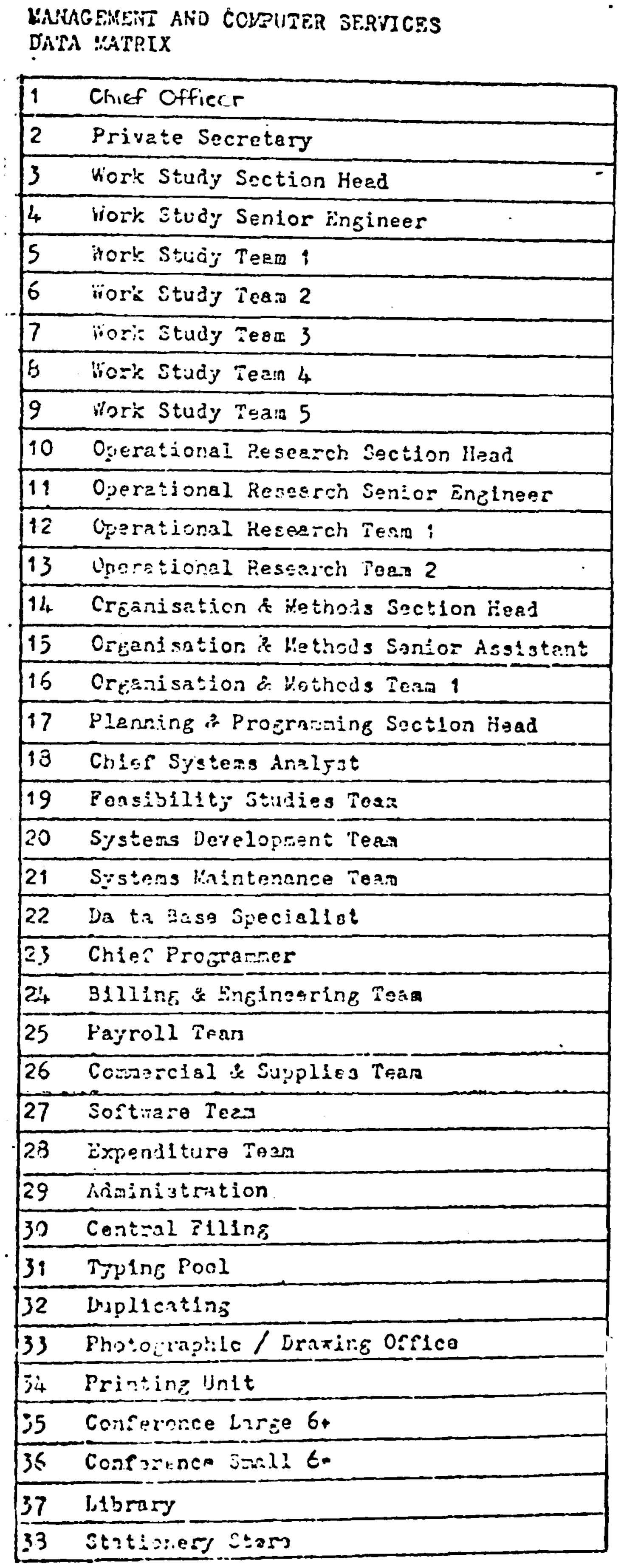
	10	15	16	14	29	32	35	36	1	31	19	18	17
11	0.00	3.00	3.00	2.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00
34	3.00	2.00	1.00	2.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	3.00	2.00	2.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
37	3.00	3.00	3.00	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	3.00	3.00	3.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2.00	3.00	3.00	2.00	1.00	1.00	0.00	0.00	1.00	3.00	0.00	0.00	1.00
8	2.00	2.00	2.00	2.00	3.00	3.00	0.00	0.00	3.00	3.00	0.00	0.00	0.00
4	2.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	0.00	3.00	3.00
7	3.00	3.00	3.00	3.00	2.00	3.00	3.00	3.00	3.00	2.00	0.00	0.00	0.00
6	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	0.00	0.00	0.00
5	3.00	3.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	0.00	0.00	3.00
12	1.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	3.00	3.00	0.00	3.00
13	1.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	3.00	3.00	3.00	2.00
10	0.00	2.00	3.00	1.00	1.00	3.00	3.00	2.00	1.00	3.00	3.00	3.00	2.00
15	2.00	0.00	1.00	1.00	3.00	3.00	3.00	2.00	2.00	3.00	3.00	3.00	3.00
16	3.00	1.00	0.00	1.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
14	1.00	1.00	1.00	0.00	2.00	3.00	3.00	1.00	1.00	3.00	3.00	3.00	2.00
29	1.00	3.00	3.00	2.00	0.00	1.00	0.00	0.00	1.00	3.00	0.00	3.00	2.00
32	3.00	3.00	3.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	2.00	2.00	3.00
35	3.00	3.00	3.00	3.00	0.00	0.00	0.00	0.00	1.00	0.00	3.00	2.00	2.00
36	2.00	2.00	3.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	3.00	2.00	2.00
1	1.00	2.00	3.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	3.00	2.00	1.00
31	3.00	3.00	3.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	2.00	2.00	2.00
19	3.00	3.00	3.00	3.00	0.00	2.00	3.00	3.00	3.00	2.00	0.00	1.00	2.00
18	3.00	3.00	3.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00	1.00	0.00	1.00
17	2.00	3.00	3.00	2.00	2.00	3.00	2.00	2.00	1.00	2.00	2.00	1.00	0.00
22	3.00	3.00	0.00	3.00	3.00	2.00	2.00	3.00	2.00	2.00	2.00	1.00	1.00
20	3.00	3.00	3.00	3.00	0.00	2.00	3.00	3.00	3.00	2.00	2.00	1.00	2.00
21	3.00	3.00	0.00	0.00	0.00	2.00	2.00	3.00	3.00	2.00	2.00	1.00	2.00
24	3.00	0.00	0.00	0.00	0.00	2.00	3.00	3.00	3.00	2.00	3.00	2.00	2.00
26	3.00	0.00	0.00	0.00	0.00	2.00	3.00	3.00	3.00	2.00	3.00	2.00	2.00
25	3.00	0.00	0.00	0.00	0.00	2.00	3.00	3.00	3.00	2.00	3.00	2.00	2.00
27	3.00	0.00	0.00	0.00	0.00	2.00	3.00	3.00	3.00	2.00	3.00	2.00	2.00
28	3.00	0.00	0.00	0.00	0.00	2.00	3.00	3.00	0.00	2.00	3.00	2.00	2.00
38	3.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	3.00	3.00	3.00
3	1.00	2.00	3.00	1.00	1.00	3.00	2.00	1.00	1.00	3.00	3.00	3.00	2.00
23	3.00	0.00	0.00	0.00	3.00	2.00	2.00	2.00	2.00	2.00	3.00	1.00	1.00
9	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 12.19 - Management and Computer Services Department, reordered matrix (continued)

MATRIX AFTER REORDERING SECTION 3

	22	20	21	24	26	25	27	28	38	3	23	9
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
34	0.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00	0.00
37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.00	2.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	3.00	0.00	0.00	0.00	1.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	3.00	0.00	3.00
13	3.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00	3.00	3.00	0.00	3.00
10	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	3.00	3.00
15	3.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00
16	0.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00	0.00
14	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
29	3.00	0.00	0.00	0.00	0.00	0.00	0..00	0.00	1.00	1.00	3.00	0.00
32	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0.00	3.00	2.00	0.00
35	2.00	3.00	2.00	3.00	3.00	3.00	3.00	3.00	0.00	2.00	2.00	0.00
36	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	1.00	2.00	0.00
1	2.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	0.00	1.00	2.00	0.00
31	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0.00	3.00	2.00	0.00
19	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00
18	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	1.00	0.00
17	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	2.00	1.00	0.00
22	0.00	2.00	2.00	2.00	2.00	2.00	1.00	2.00	3.00	3.00	1.00	0.00
20	2.00	0.00	2.00	1.00	1.00	1.00	2.00	1.00	3.00	3.00	2.00	0.00
21	2.00	2.00	0.00	1.00	1.00	1.00	2.00	1.00	3.00	3.00	2.00	0.00
24	2.00	1.00	1.00	0.00	2.00	3.00	1.00	3.00	3.00	0.00	1.00	0.00
26	2.00	1.00	1.00	2.00	0.00	3.00	1.00	2.00	3.00	0.00	1.00	0.00
25	2.00	1.00	1.00	3.00	3.00	0.00	1.00	2.00	3.00	0.00	1.00	0.00
27	1.00	2.00	2.00	1.00	1.00	1.00	0.00	1.00	3.00	0.00	1.00	0.00
28	2.00	1.00	1.00	3.00	2.00	2.00	1.00	0.00	3.00	0.00	1.00	0.00
38	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	0.00	3.00	0.00
3	3.00	3.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00
23	1.00	2.00	2.00	1.00	1.00	1.00	1.00	1.00	3.00	3.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

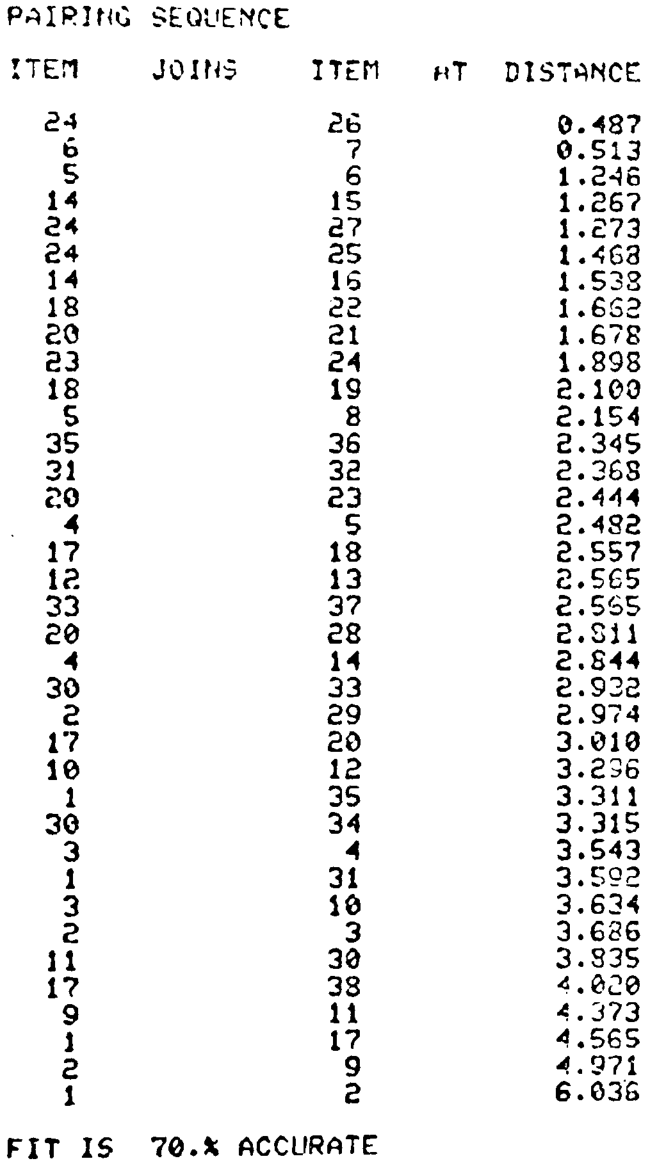
Figure 12.19 - Management and Computer Services Department, reordered matrix (continued)



MANAGEMENT AND COMPUTER SERVICES
DATA MATRIX

1	Chief Officer
2	Private Secretary
3	Work Study Section Head
4	Work Study Senior Engineer
5	Work Study Team 1
6	Work Study Team 2
7	Work Study Team 3
8	Work Study Team 4
9	Work Study Team 5
10	Operational Research Section Head
11	Operational Research Senior Engineer
12	Operational Research Team 1
13	Operational Research Team 2
14	Organisation & Methods Section Head
15	Organisation & Methods Senior Assistant
16	Organisation & Methods Team 1
17	Planning & Programming Section Head
18	Chief Systems Analyst
19	Feasibility Studies Team
20	Systems Development Team
21	Systems Maintenance Team
22	Da ta Base Specialist
23	Chief Programmer
24	Billing & Engineering Team
25	Payroll Team
26	Commercial & Supplies Team
27	Software Team
28	Expenditure Team
29	Administration
30	Central Filing
31	Typing Pool
32	Duplicating
33	Photographic / Drawing Office
34	Printing Unit
35	Conference Large 6+
36	Conference Small 6-
37	Library
38	Stationery Store

Figure 12.20
Management and Computer Services Dept., activity list



```
PAIRING SEQUENCE

ITEM     JOINS     ITEM    HT  DISTANCE

  24                 26           0.487
   6                  7           0.513
   5                  6           1.246
  14                 15           1.267
  24                 27           1.273
  24                 25           1.468
  14                 16           1.538
  18                 22           1.662
  20                 21           1.678
  23                 24           1.898
  18                 19           2.100
   5                  8           2.154
  35                 36           2.345
  31                 32           2.368
  20                 23           2.444
   4                  5           2.482
  17                 18           2.557
  12                 13           2.565
  33                 37           2.565
  20                 28           2.811
   4                 14           2.844
  30                 33           2.932
   2                 29           2.974
  17                 20           3.010
  10                 12           3.296
   1                 35           3.311
  30                 34           3.315
   3                  4           3.543
   1                 31           3.592
   3                 10           3.634
   2                  3           3.686
  11                 30           3.835
  17                 38           4.020
   9                 11           4.373
   1                 17           4.565
   2                  9           4.971
   1                  2           6.036

FIT IS  70.% ACCURATE
```

Figure 12.21

Management and Computer Services Dept., pairing sequence

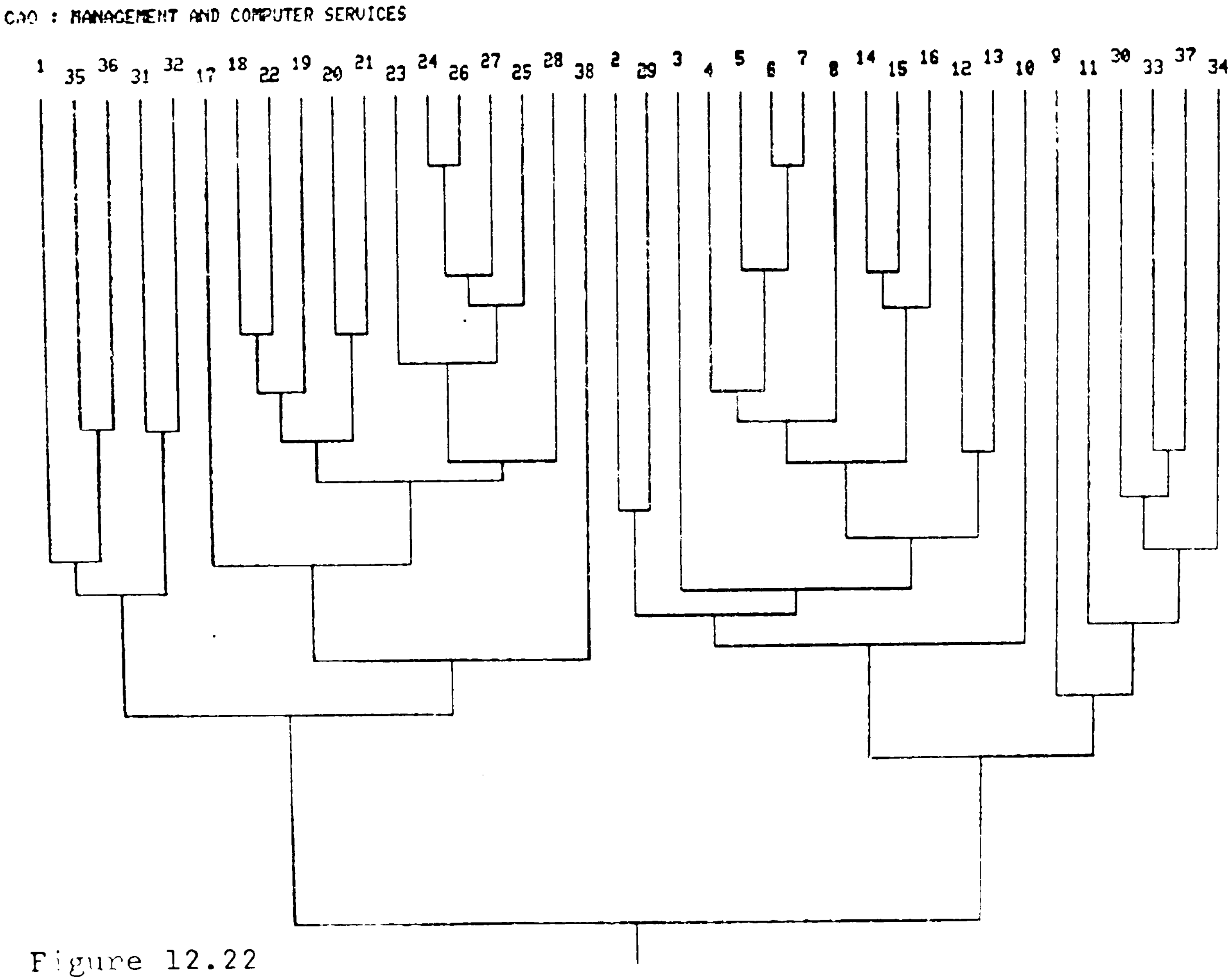


Figure 12.22

Management and Computer Services Department, tree diagram

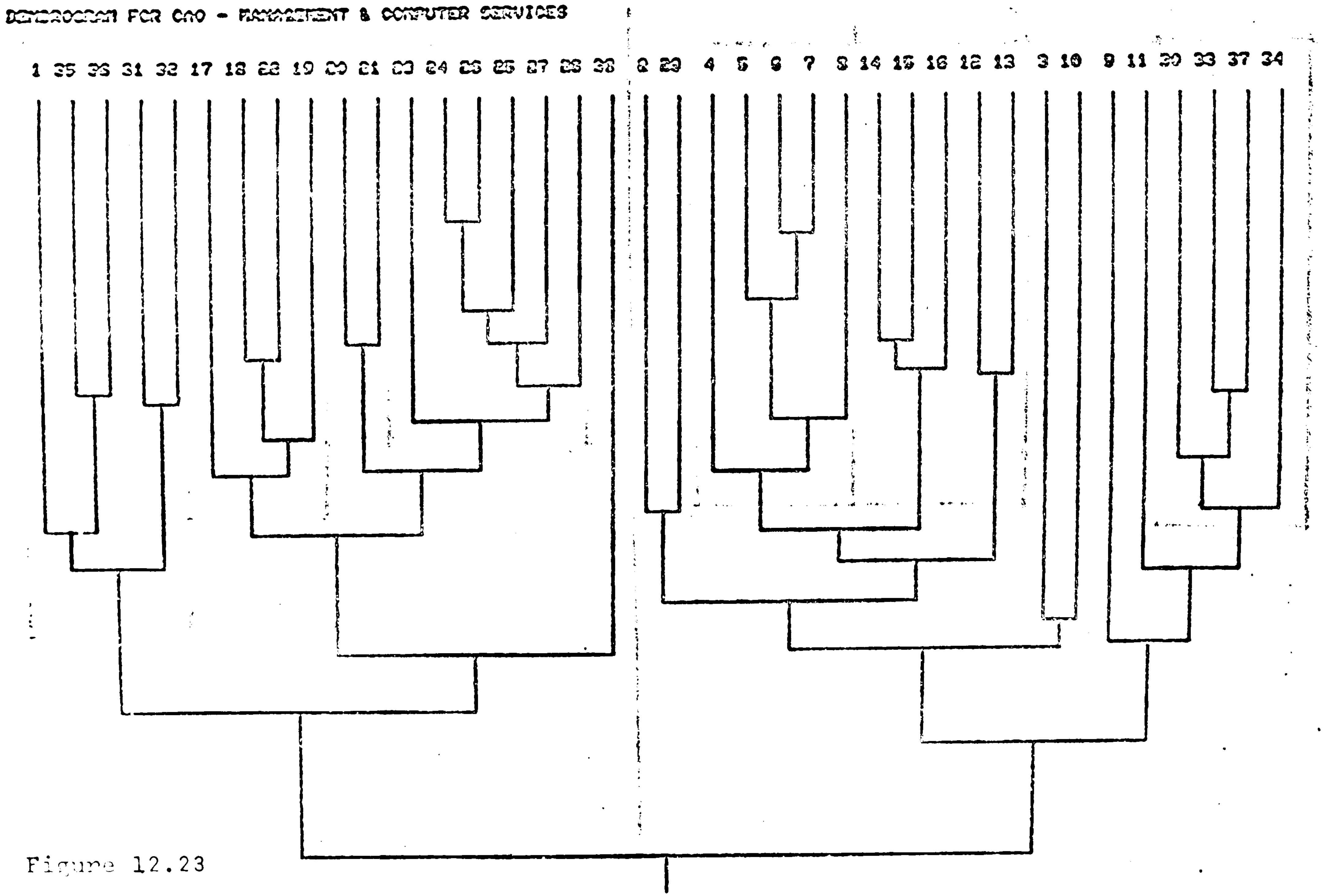
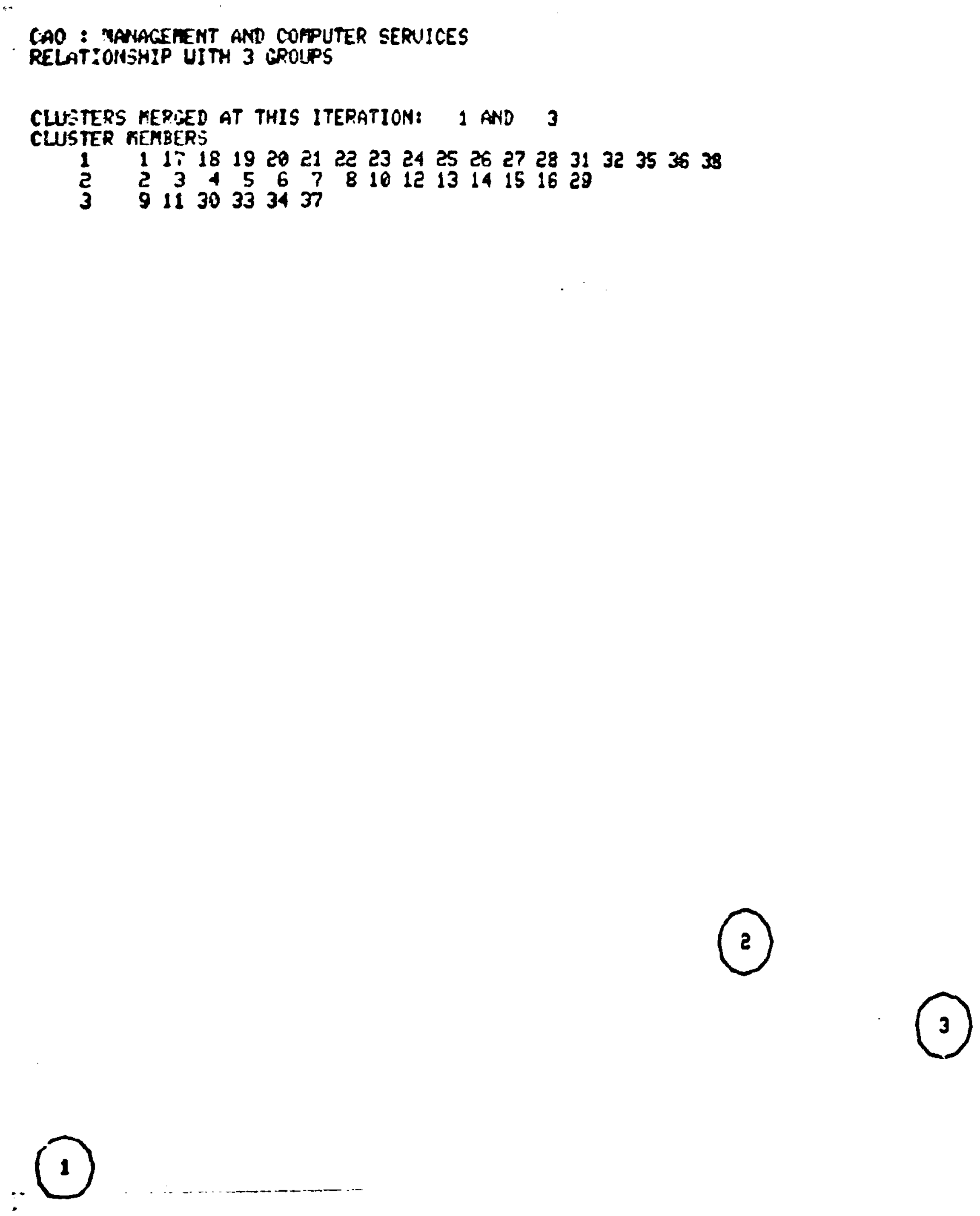


Figure 12.23

Management and Computer Services Department, annotated tree diagram



```
CAO : MANAGEMENT AND COMPUTER SERVICES
RELATIONSHIP WITH 3 GROUPS


CLUSTERS MERGED AT THIS ITERATION:   1 AND   3
CLUSTER MEMBERS
    1    1 17 18 19 20 21 22 23 24 25 26 27 28 31 32 35 36 38
    2    2  3  4  5  6  7  8 10 12 13 14 15 16 29
    3    9 11 30 33 34 37
```

Figure 12.24

Example of "seperate groups" output

Discrete Groupings

Maximum of 13 groups isolated as follows

Group No.	members
1	2, 4, 5, 6, 7, 8
2	10
3	38
4	11
5	3, 12, 13, 14, 15, 16, 29
6	20, 21
7	1, 31, 32, 35, 36
8	30, 33, 34, 37
9	17, 18, 19
10	9
11	24, 25, 26, 27
12	22, 23
13	28

Reducing the number of groups one by one we obtain the following.

First groups 11 and 13 join together. Group 11 becomes 24, 25, 26, 27 and 28. Group 13 ceases to exist and all others remain unchanged.

Next group 6 joins group 12, but items 22 and 23 are also reallocated to groups 9 and 11 respectively. This leaves the following situation.

Group No.	members
1	2, 4, 5, 6, 7, 8
2	10
3	38
4	11
5	3, 12, 13, 14, 15, 16, 29
6	20, 21
7	1, 31, 32, 35, 36
8	30, 33, 34, 37
9	17, 18, 19, 22
10	9
11	23, 24, 25, 26, 27, 28

Next groups 6 and 11 join, and item 20 moves from group 6 to group 9. Groups 2 and 5 then join so with 9 groups we have:

Figure 12.25

Summary of seperate groups output

Group No.	members
1	2, 4, 5, 6, 7, 8
2	3, 10, 12, 13, 14, 15, 16, 29
3	38
4	11
5	21, 23, 24, 25, 26, 27, 28
6	1, 31, 32, 35, 36
7	30, 33, 34, 37
8	17, 18, 19, 20, 22
9	9

Groups 4 and 7 join next (causing groups 8 and 9 to be renumbcred 7 and 8). Groups 3 and 5 then join to leave:

Group No.	members
1	2, 4, 5, 6, 7, 8
2	3, 10, 12, 13, 14, 15, 16, 29
3	21, 23, 24, 25, 26, 27, 28, 38
4	11, 30, 33, 34, 37
5	1, 31, 32, 35, 36
6	17, 18, 19, 20, 22
7	9

Groups 4 and 7 join next, then groups 3 and 6. Item 17 moves from cluster 3 to cluster 5, so we have:

Group No.	members
1	2, 4, 5, 6, 7, 8
2	3, 10, 12, 13, 14, 15, 16, 29
3	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 38
4	9, 11, 30, 33, 34, 37
5	1, 17, 31, 32, 35, 36

Next groups 1 and 2 join and item 10 moves from group 1 to group 4. Groups 2 and 4 then join and force item 10 back into group 1. So the situation is

Group No.	members
1	2, 3, 4, 5, 6, 7, 8, 10, 12, 13, 14, 15, 16, 29
2	1, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 35, 36, 38
3	9, 11, 30, 33, 34, 37

Finally groups 1 and 3 join together.

Figure 12.25

Summary of seperate groups output (continued)

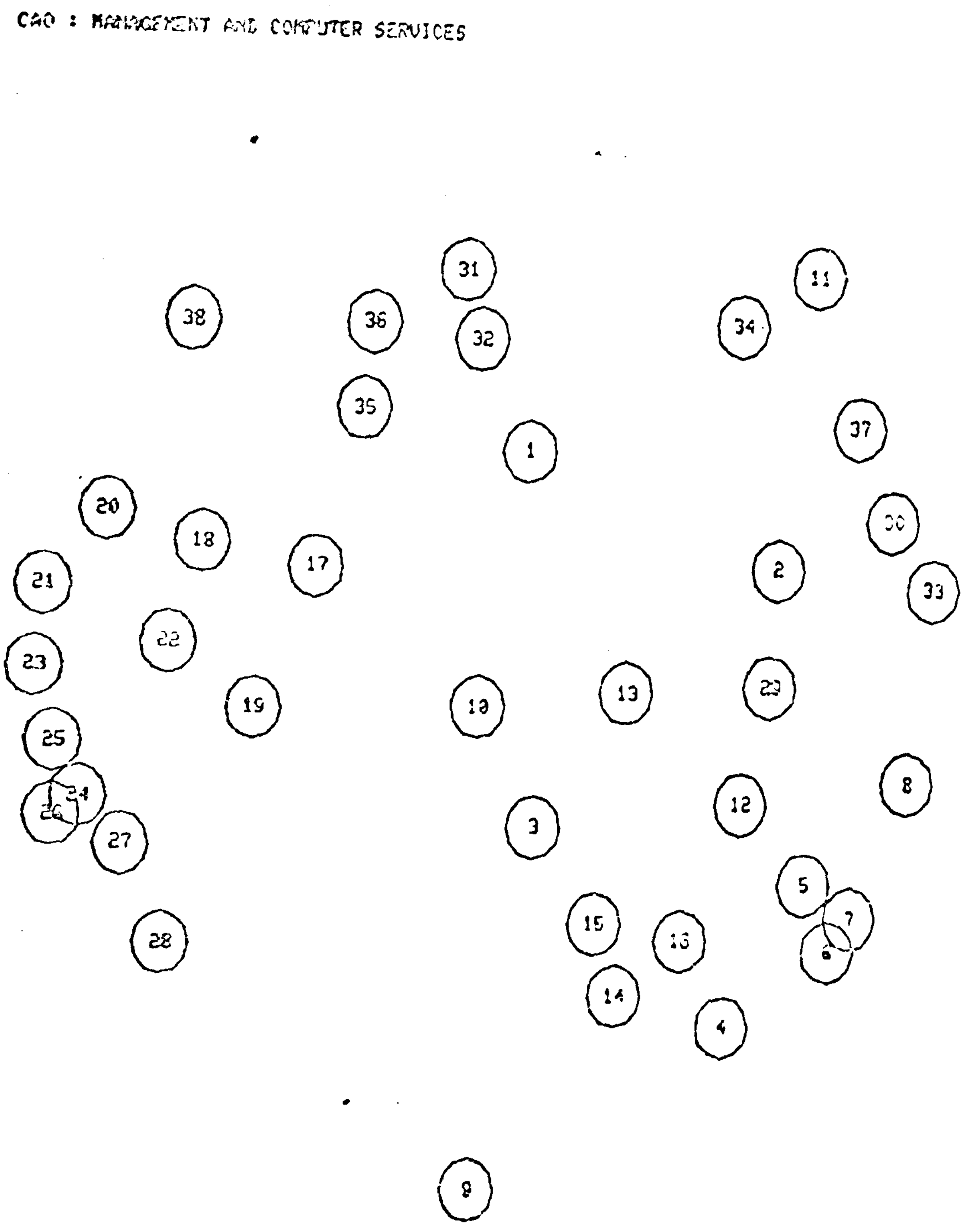


Figure 12.26

Management and Computer Services Department bubble diagram

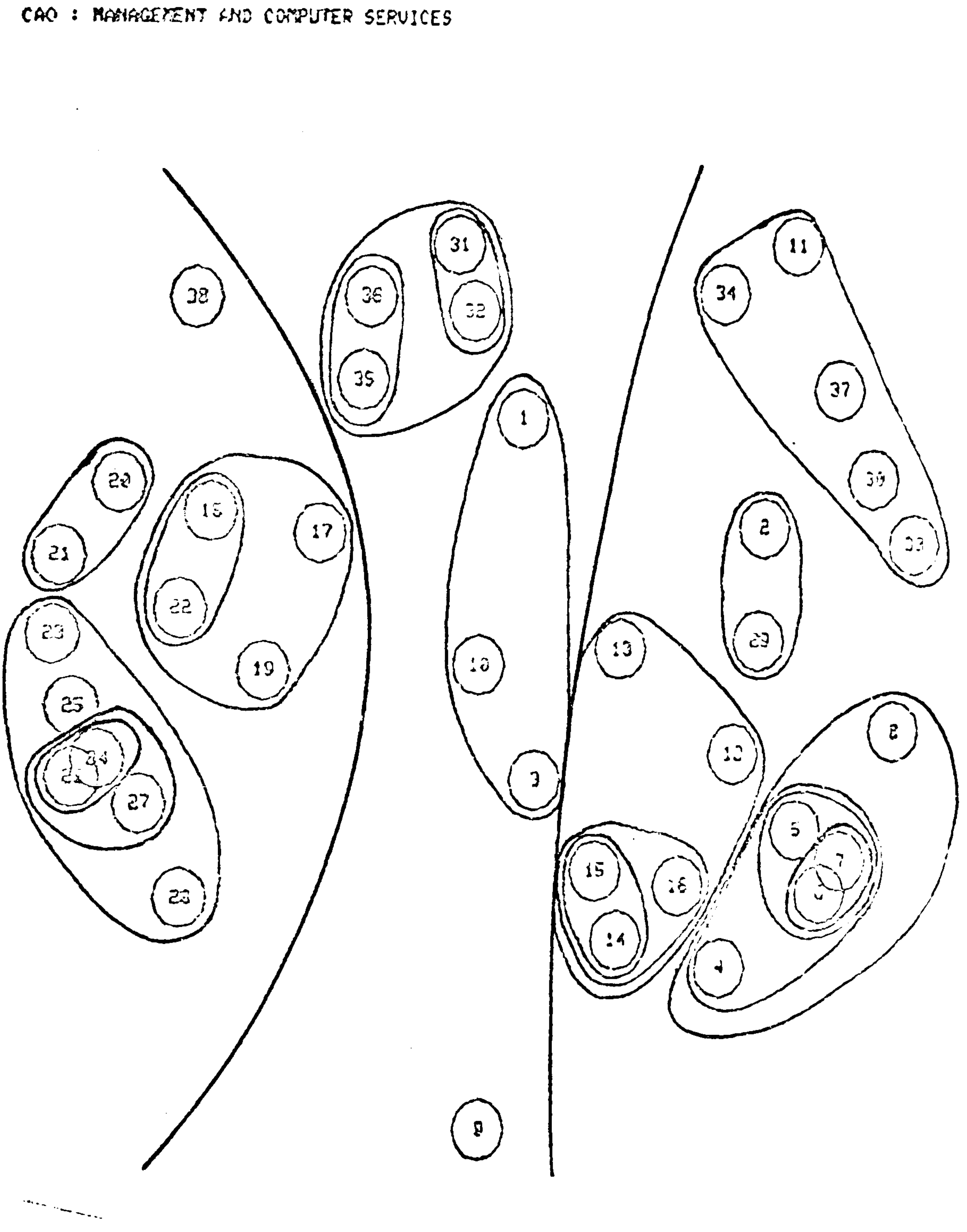


Figure 12.27

Management and Computer Services Department

annotated bubble diagram

The reordered matrix (figure 12.19) simply attempts to shuffle the list order of the elements to bring together the associated activities to push the zero elements (or unrelated activities) out to the corners. By examining the matrix and isolating blocks of 1, 2, and 3's groups of related activities can be identified. One possible grouping is identified on one of the printouts.

The hierarchical tree or grouping diagram (figure 12.22) shows a list of the 38 activities across the top of the page. Pairs of activities are then successively joined together; the nearer the top of the page the joining occurs the more the activities have in common. On one printout (figure 12.23) the more obvious groups have been identified; these groups compare very well with the "seperate groups" analysis showing 9 groups (figure 12.25). One further important point is the very striking division into two seperate hierarchies (shown divided by the dashed line). This division is confirmed by the identical splitting of the activities into two groups by the seperate groups analysis.

The bubble diagram (figure 12.26) is open to a number of interpretations. The one shown (figure 12.27) is based on comparisons with other analyses. Groups contained in larger groups are indicated by the concentric circles. The other lines are there to suggest that activities (1,3,10) and (31,32,35,36) have links to both sets of groups outside the lines. Thus two groups are again identified but with mutual links to common services and chiefs of staff.

Further uses of the data analyses include identifying the weak links for the splitting of activities in different offices or on different floors. The areal implications of the different groups in the seperate groups analysis could br calculated and compared with available floor areas. If the result is too large or small it is only necessary to look at what happens with more or fewer groups as required. In doing this exercise it is useful to keep referring to the hierarchical analysis to ensure that no important (high-up) link becomes accidentally seperated.

12.3 POE OF GENETICS BUILDING EDINBURGH

The MRC facilities are spread over four sites at the Western General Hospital (figure 12.28).

1. Animal House. The MRC have a 2/5 share, with the University and the NHS taking the remainder.

2. Radiotherapy (West Building). A variety of MRC work is carried on in the radiotherapy wing of the main hospital.

3. 1964 Building (Centre Building). The Experimental Studies Section is housed here.

4. MRC/Human Genetics (the study building). Cytogenetics and pattern recognition work is located here.

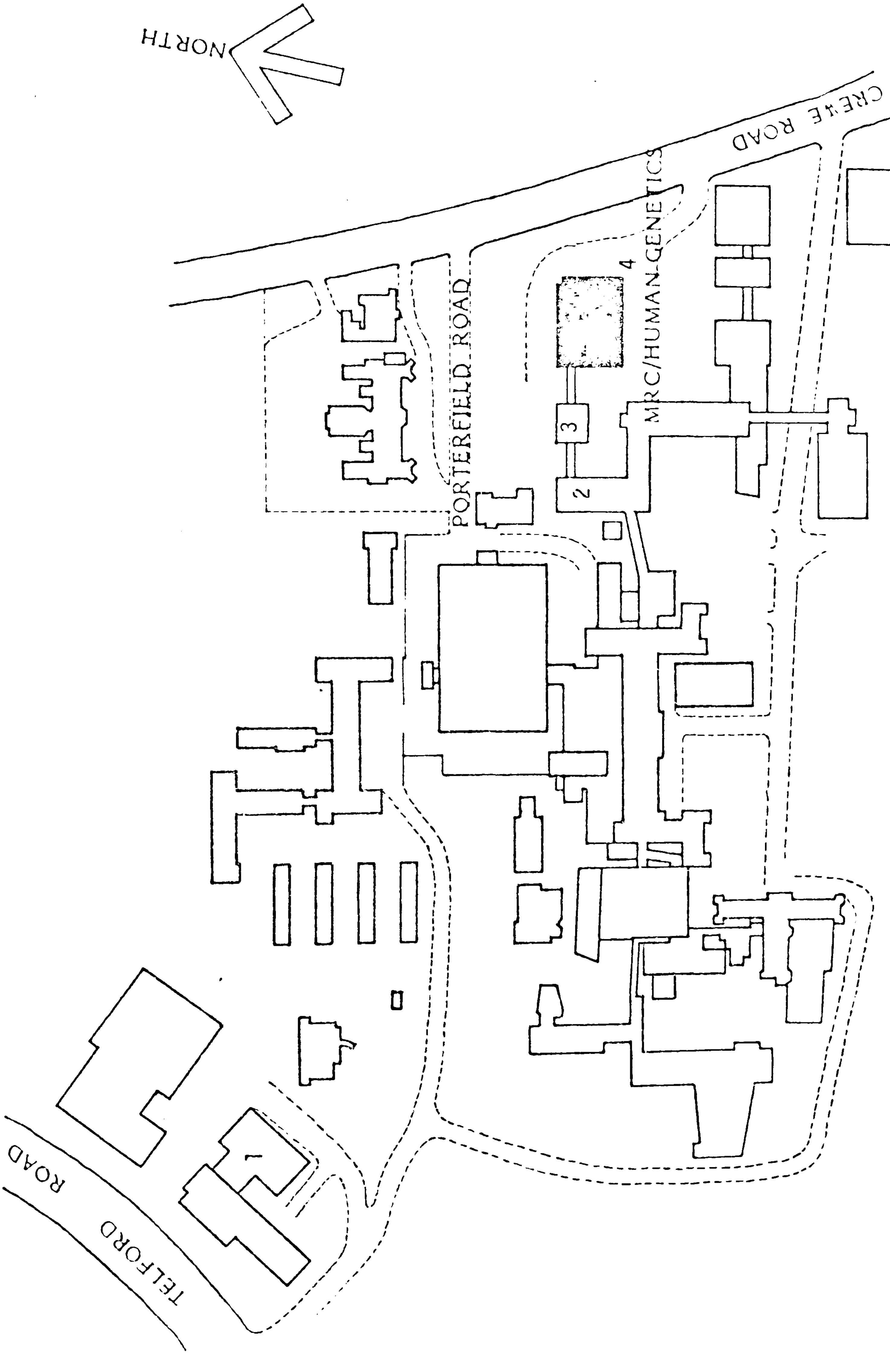


Figure 12.28 - Study building location

12.3.1 The Tasks And Aims Of The Unit

The main work of the unit is the study of chromosome variation and its consequences in man. This work is organised under the supervision of a Director, into five main research sections as follows.

12.3.1.1 Cytogenetics - The general aims of the section are: to establish the incidence and nature of "normal" chromosome variation and of constitutional anomalies in human populations and (in collaboration with the Clinical Studies Section) to define the biological, clinical and social implications of these variations. In practice the work of the section is loosely organised into four groups:

- a group that is largely concerned with population cytogenetics, family studies including work on linkage and polymorphisms and studies on the development and application of new techniques for looking at chromoscmes.

- a group that is concerned with the clinical and psychological study of the chromosomally abnormal children identified in the Unit's earlier newborn baby studies.

- a group that is concerned with cytogenetic aspects of reproductive biology.

- a group that is largely concerned with the cytogenetic and allied effects of environmental mutagens.

12.3.1.2 Director's Section - This is a small section which comprises a number of service groups, including the Unit Cytogenetics Registry, Photography, Electron Microscopy, and one research group - the molecular cytogenetic group. The Cytogenetics Registry acts as the repository for cytogenetic and clinical data for the unit. The Electron Microscopy group is involved in joint projects with the Cytogenetic section.

12.3.1.3 Cinical Studies - This is also a small section and its general aims are: to identify and establish the prevalence of chromosome abnormalities in various populations, in collaboration with the Cytogenetics Section; to undertake clinical correlations of chromosome abnormalities and the pathogenesis of associated diseases; and to provide a specialised service for family "follow-up" genetic studies and for the collection of blood and biopsy specimens for the rest of the Unit.

12.3.1.4 The Pattern Recognition Section - This sections concern is with research and development in the field of instrumentation for automatic location, identification and measurement of cells, chromosomes, and similar microscopic objects. The section also has a substantial service commitment to pattern measurement, statistics, computer programming and operation, and electronic and maintenance work.

12.3.1.5 Experimental Studies - This section is mostly concerned with the malignant transformation of human cells, and in particular lymphoid cells; genetic aspects of the immune response; and human somatic cell genetics. A small group have an active programme in studying cell-cell interaction in vitro and in vivo, as well as investigating organo-genesis in early embryos. The section also provides a general tissue culture service.

12.3.2 Staffing

The formal staff organisation is summarised in figure 12.29.

Formal Staff Organisation

Unit Director
(overall supervision)

Section	Location	S & RO	VS	T
Cytogenetics	4 .	14	*	12(5Pt)
Directors	2,3,4(a)	6	-	12
Clinical Studies	3	8(4Pt)	-	8
Pattern Recog.	4	16(1pt)	-	8(1pt)
Exp. Studies	3	12(2Pt)	-	11

Notes

Director's Unit
Activities carried out in study building -
Photographic Unit - 3T

Administration & Maintenance staff in study building
Admin - 7 AO (2Pt)
Maintenance - 5 MO (1Pt)

Total MRC staff in study building: Director plus 30 S & RO (1Pt); 1 VS; 23T (6Pt); 7 AO (2Pt); 5 MO (1Pt). Total 67 people (10Pt).

Key to abbreviations:

S & RO	Scientific and Research Officer
VS	Visiting Scientist
T	Technician
AO	Administration Officer
MO	Maintenance Officer
Pt	Part-time

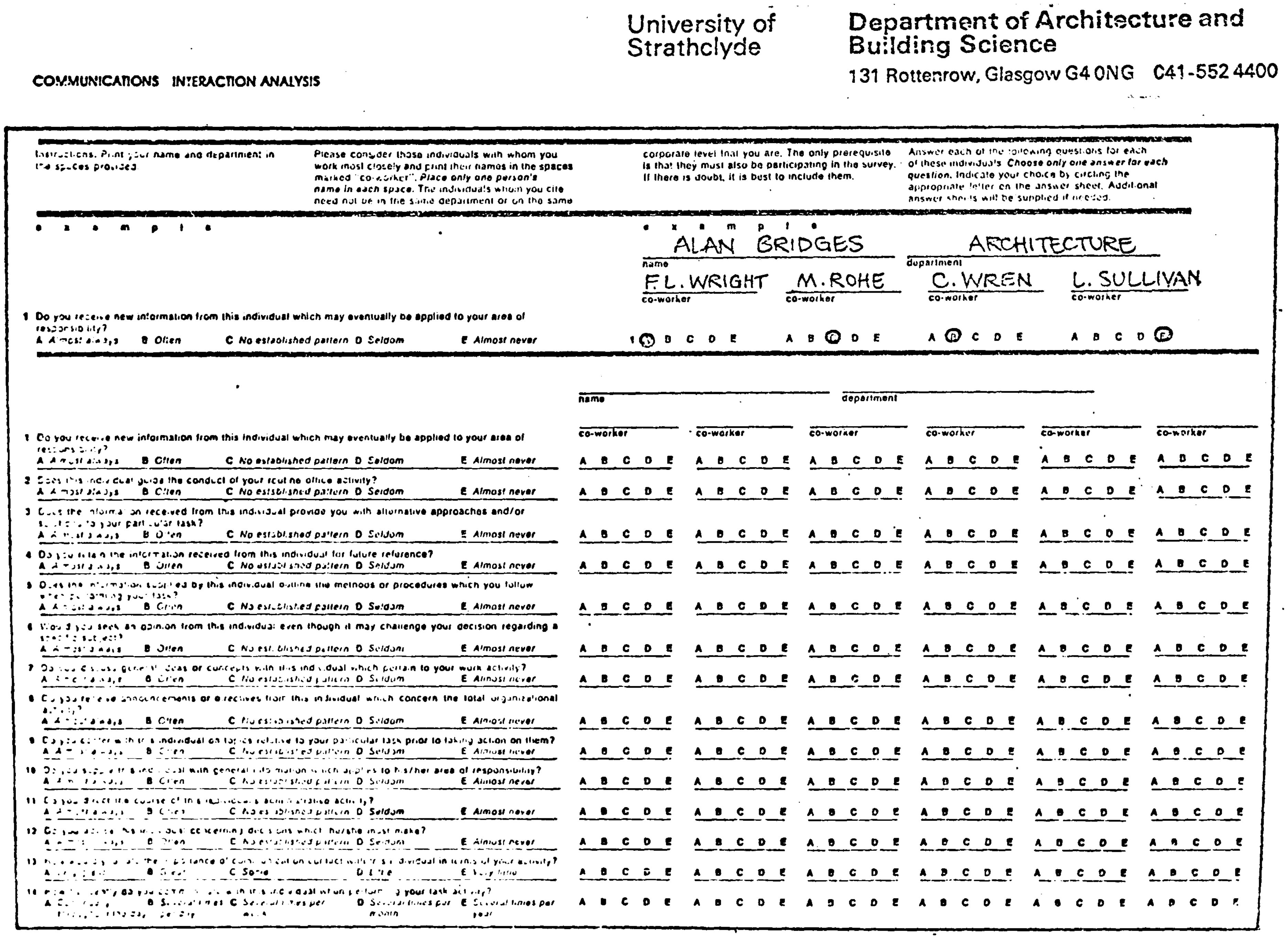
Figure 12.29

Staffing

12.3.3 Communications Interaction Analysis

The previous section has defined the "formal" staff organisation of the unit. All of the following analyses attempt, in different ways, to discover the actual communication patterns and work groups operating in the building. The same analysis is also applied to the Brief information to compare actuality with what the architects may have been able to analyse during their sketch design.

The first analysis is produced from survey returns on the form illustrated in figure 12.30. Here individual workers are asked to define those individuals they need to liaise with during their work. By an analysis of this information the formal organisation can be broken down into each individual's communication patterns, which are then expanded into task related groups, and then shown aggregating into the complete research unit. These relationships are presented diagrammatically in the following pages, and can be read as the formal inter- and intra- departmental links actually existing (as opposed to the formal organisation supposed to exist) within the organisation.



University of Strathclyde

Department of Architecture and Building Science

131 Rottenrow, Glasgow G4 0NG C41-552 4400

COMMUNICATIONS INTERACTION ANALYSIS

Instructions. Print your name and department in the spaces provided.

Please consider those individuals with whom you work most closely and print their names in the spaces marked "co-worker". *Place only one person's name in each space.* The individuals whom you cite need not be in the same department or on the same corporate level that you are. The only prerequisite is that they must also be participating in the survey. If there is doubt, it is best to include them.

Answer each of the following questions for each of these individuals. *Choose only one answer for each question.* Indicate your choice by circling the appropriate letter on the answer sheet. Additional answer sheets will be supplied if needed.

example

name: ALAN BRIDGES department: ARCHITECTURE

	F.L. WRIGHT	M. ROHE	C. WREN	L. SULLIVAN
	co-worker	co-worker	co-worker	co-worker
1 Do you receive new information from this individual which may eventually be applied to your area of responsibility? A Almost always B Often C No established pattern D Seldom E Almost never	1 [✓ A] B C D E	A B [✓ C] D E	A [✓ B] C D E	A B C D [✓ E]

name ______ department ______

	co-worker	co-worker	co-worker	co-worker	co-worker	co-worker
1 Do you receive new information from this individual which may eventually be applied to your area of responsibility? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
2 Does this individual guide the conduct of your routine office activity? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
3 Does the information received from this individual provide you with alternative approaches and/or solutions to your particular task? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
4 Do you retain the information received from this individual for future reference? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
5 Does the information supplied by this individual outline the methods or procedures which you follow when performing your task? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
6 Would you seek an opinion from this individual even though it may challenge your decision regarding a specific subject? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
7 Do you discuss general ideas or concepts with this individual which pertain to your work activity? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
8 Do you receive announcements or directives from this individual which concern the total organizational activity? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
9 Do you confer with this individual on topics relative to your particular task prior to taking action on them? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
10 Do you supply this individual with general information which applies to his/her area of responsibility? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
11 Do you direct the course of this individual's administrative activity? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
12 Do you advise this individual concerning decisions which he/she must make? A Almost always B Often C No established pattern D Seldom E Almost never	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
13 How would you rate the importance of communication contact with this individual in terms of your activity? A Very great B Great C Some D Little E Very little	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E
14 How frequently do you communicate with this individual when performing your task activity? A Continually throughout the day B Several times per day C Several times per week D Several times per month E Several times per year	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E	A B C D E

Figure 12.30 - Communications interaction analysis: survey form

Due to the survey sample not all individuals are represented, but typical individuals from each group are defined by their job titles and room numbers. The digrams are generated by each individual describing those other individuals he or she considers to be part of their functional workgroup, or those individuals to whom they give (or from whom they receive) instructions relating to their work activity. The individual work groups are finally joined together to show the complete departmental structure. In the individual work-unit diagrams (figures 12.31 to 12.53) an arrow indicates the direction of information flow (a double-headed arrow thus signifying information passing both ways). In the departmental diagram arrow heads are omitted as virtually all links at this level are two-way.

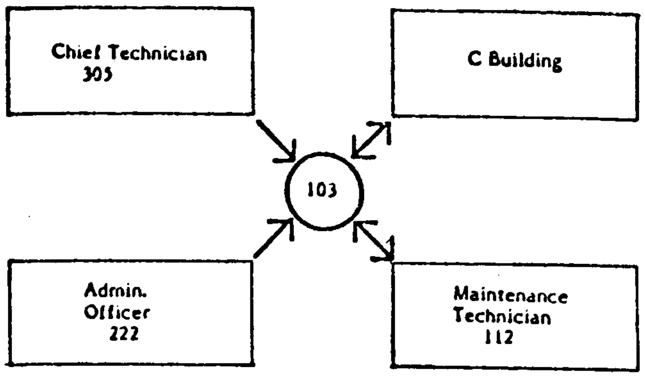


Figure 12.31

Stores technician links

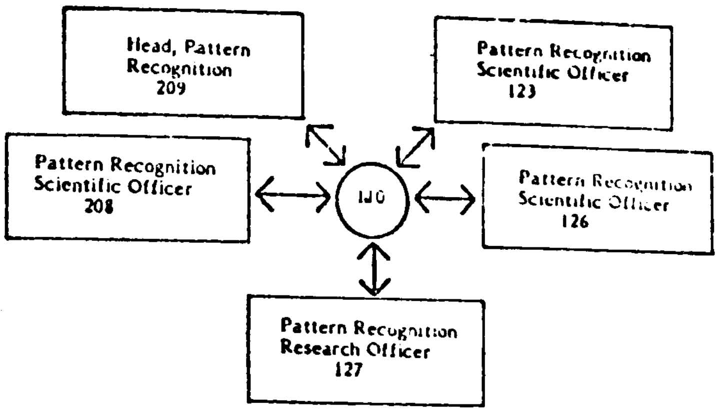


Figure 12.32

Computer section RO links

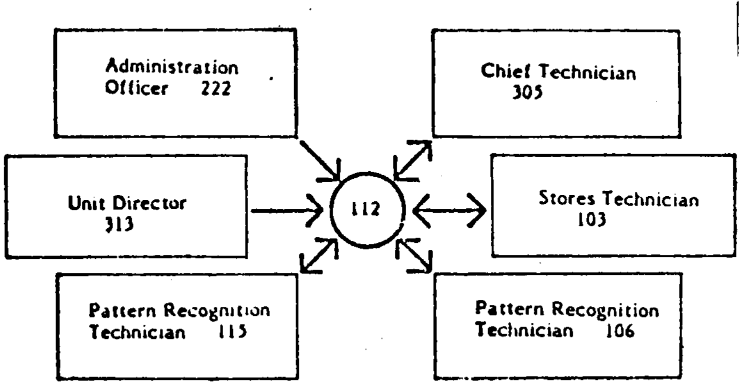


Figure 12.33

Maintenance technician

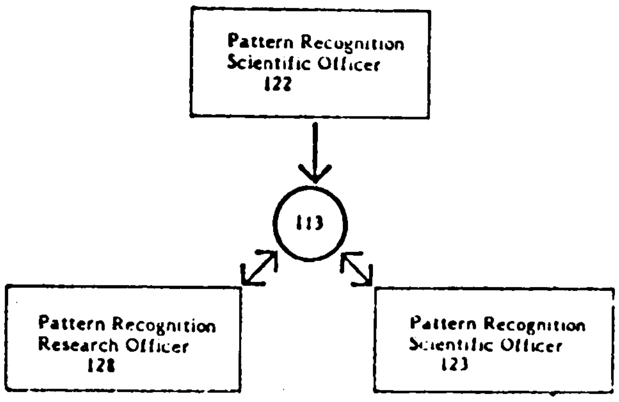


Figure 12.34

PR technician

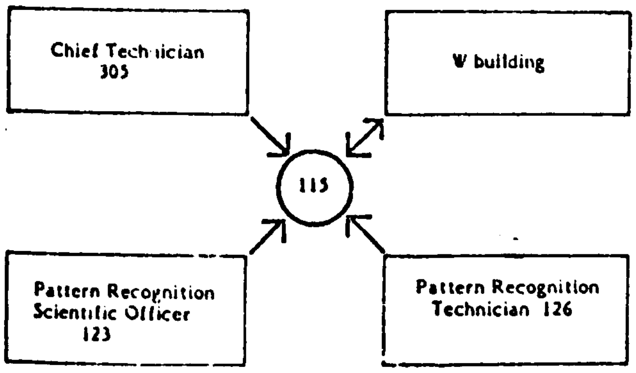


Figure 12.35

PR technician

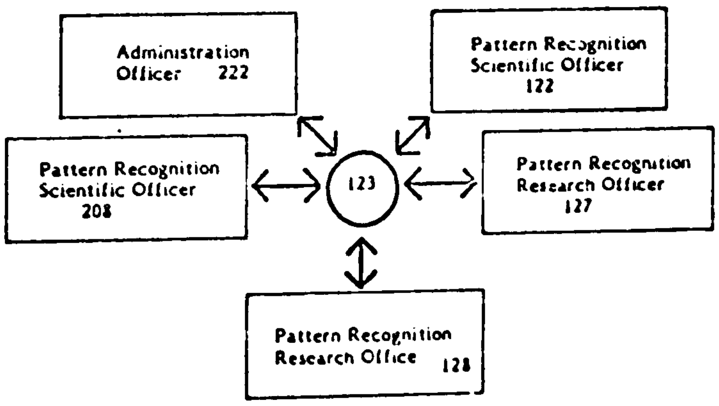


Figure 12.36

PR scientific officer

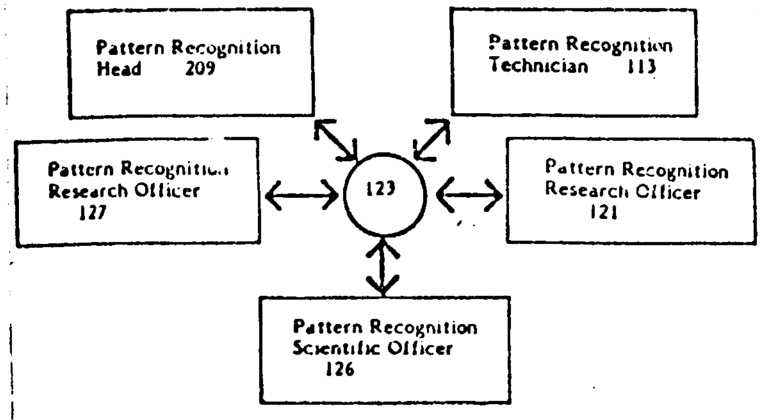

Pattern Recognition Head 209
Pattern Recognition Technician 113
Pattern Recognition Research Officer 127
123
Pattern Recognition Research Officer 121
Pattern Recognition Scientific Officer 126

Figure 12.37

PR scientific officer

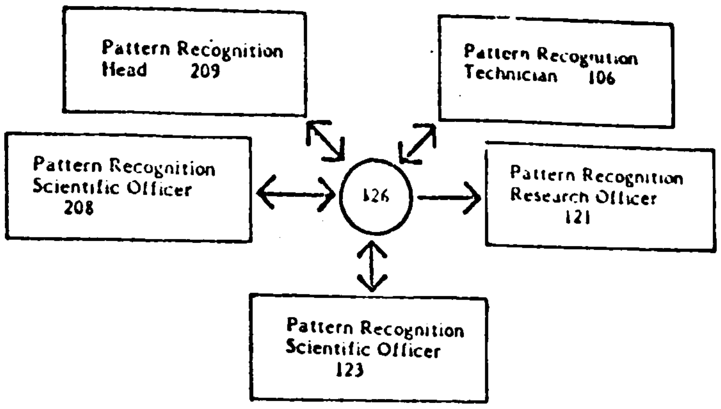

Pattern Recognition Head 209
Pattern Recognition Technician 106
Pattern Recognition Scientific Officer 208
126
Pattern Recognition Research Officer 121
Pattern Recognition Scientific Officer 123

Figure 12.38

PR scientific officer

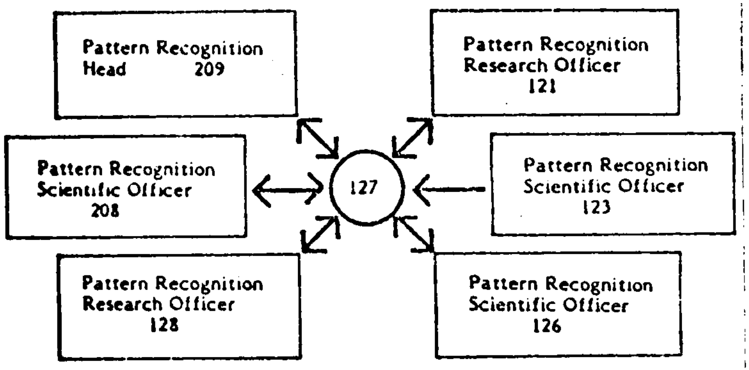

Pattern Recognition Head 209
Pattern Recognition Research Officer 121
Pattern Recognition Scientific Officer 208
127
Pattern Recognition Scientific Officer 123
Pattern Recognition Research Officer 128
Pattern Recognition Scientific Officer 126

Figure 12.39

PR research officer

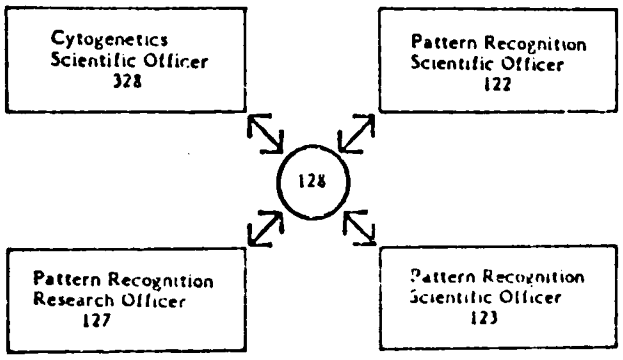

Cytogenetics Scientific Officer 328
Pattern Recognition Scientific Officer 122
128
Pattern Recognition Research Officer 127
Pattern Recognition Scientific Officer 123

Figure 12.40

PR research officer

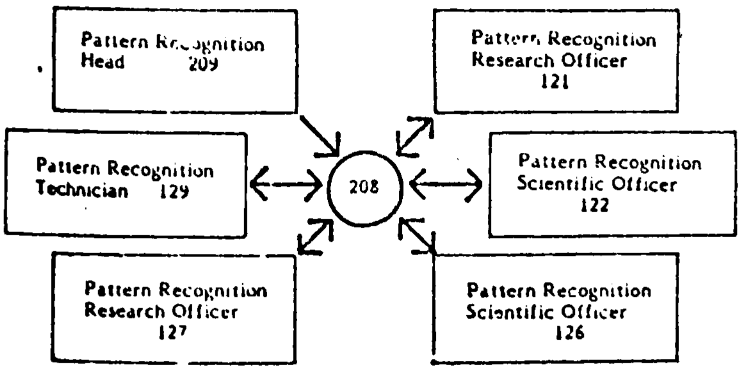

Pattern Recognition Head 209
Pattern Recognition Research Officer 121
Pattern Recognition Technician 129
208
Pattern Recognition Scientific Officer 122
Pattern Recognition Research Officer 127
Pattern Recognition Scientific Officer 126

Figure 12.41

PR scientific officer

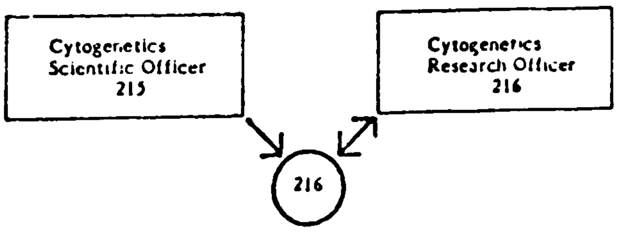

Cytogenetics Scientific Officer 215
Cytogenetics Research Officer 216
216

Figure 12.42

Cytogenetics technicians

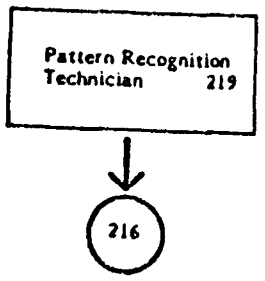


Figure 12.43

Cytogenetics technicians

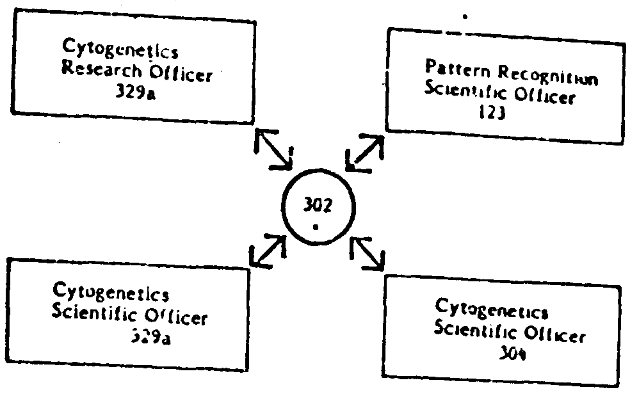

Cytogenetics Research Officer 329a
Pattern Recognition Scientific Officer 123
302
Cytogenetics Scientific Officer 329a
Cytogenetics Scientific Officer 304

Figure 12.44

Cytogenetics SO

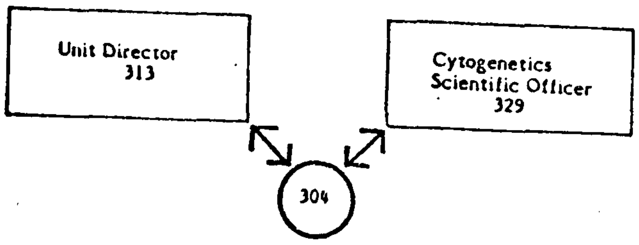


Figure 12.45

Cytogenetics SO

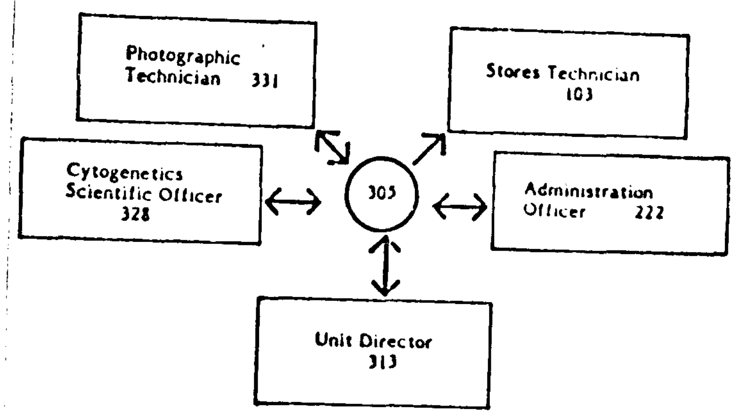

Photographic Technician 331
Stores Technician 103
Cytogenetics Scientific Officer 328
305
Administration Officer 222
Unit Director 313

Figure 12.46

Chief technician

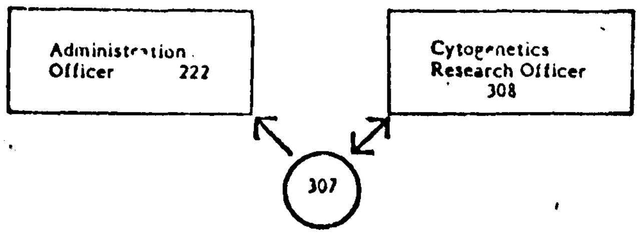


Figure 12.47

Cytogenetics SO

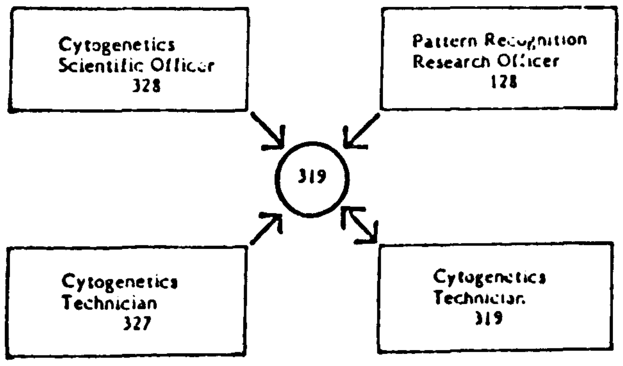

Cytogenetics Scientific Officer 328
Pattern Recognition Research Officer 128
319
Cytogenetics Technician 327
Cytogenetics Technician 319

Figure 12.48

Cytogenetics technician

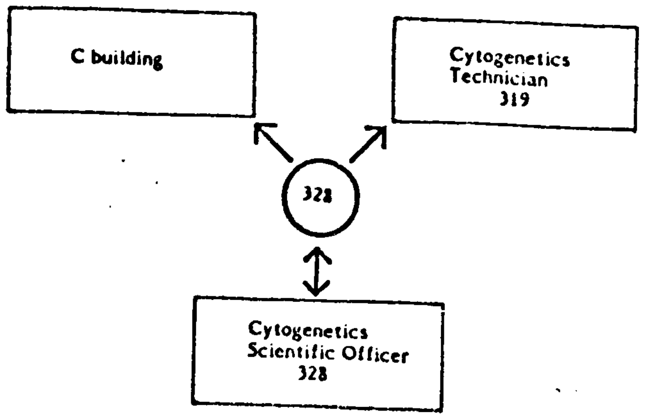

C building
Cytogenetics Technician 319
328
Cytogenetics Scientific Officer 328

Figure 12.49

Cytogenetics SO

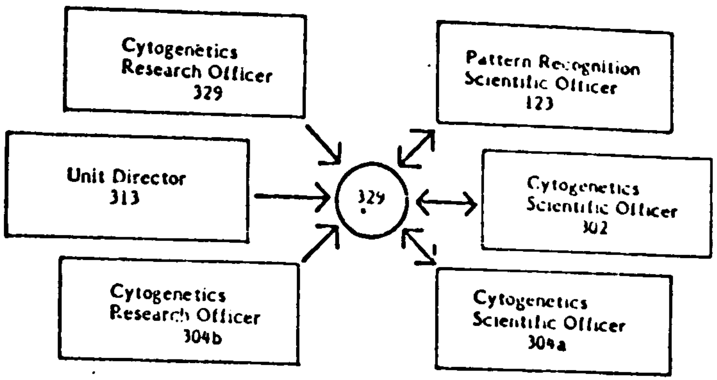

Cytogenetics Research Officer 329
Pattern Recognition Scientific Officer 123
Unit Director 313
329
Cytogenetics Scientific Officer 302
Cytogenetics Research Officer 304b
Cytogenetics Scientific Officer 304a

Figure 12.50

Cytogenetics SO

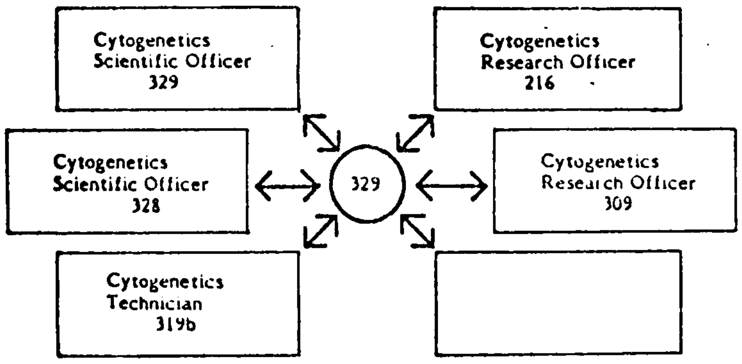

Cytogenetics Scientific Officer 329
Cytogenetics Research Officer 216
Cytogenetics Scientific Officer 328
329
Cytogenetics Research Officer 309
Cytogenetics Technician 319b

Figure 12.51

Cytogenetics RO

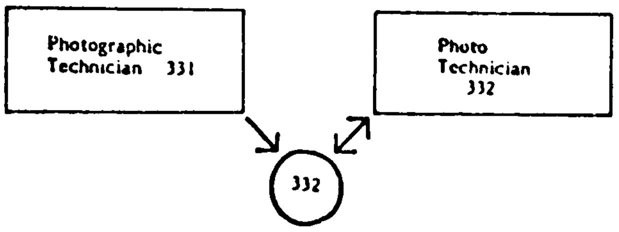


Figure 12.52

Photo technician

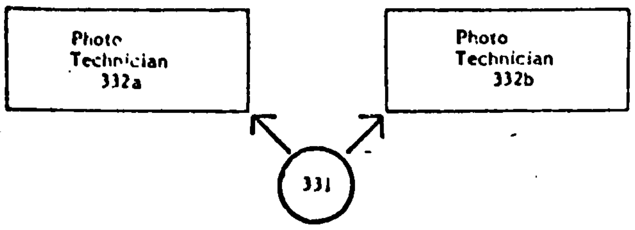


Figure 12.53

Photo technician

The individual diagrams indicate a very fragmented organisation with little interaction between individuals working on different projects. The combined diagram (figure 12.54), showing the departmental organisation, indicates four major groupings centred around rooms 103 and 112 as the stores and maintenance functions; room 329 as the cytogenetic section; and room 123 as the pattern recognition section: this compares very well with the formal organisational divisions. The only members of the Director's Unit in the study building (the photographic unit) also appear quite distinctly located (rooms 331 and 332).

The final figures (12.55, 56, 57) in this section map these individual links onto plans of the study building. Lines indicate communications between linked rooms; not all rooms are included in these linkages due to the nonavailability of staff at the time of the surveys.

Room No.	Occupant(s)
103	Stores technician
106	P.R. technician
112	Maintenance technician
113	P.R. technician
114	P.R. technician
115	P.R. technician
117	F.R. technician
120	P.R. technician
121	P.R. Research Officer
122	P.R. Scientific Officer
123	P.R. Scientific Officer
124	P.R. Scientific Officer
125	P.R. Research Officer
126	P.R. Scientific Officer
127	P.R. Research Officer
128	P.R. Research Officer
129	P.R. technician
130	Computer Room
201	Secretaries/ administration
202	Librarian
203	Library
208	P.R. Scientific Officer
209	P.R. Section Head
214	Cytogenetics Scientific Officer
215	Cytogenetics Scientific Officer
216	Cytogenetics technician/R.O.
219	P.R. technician
222	Administration Officer
224	Cytogenetics Scientific Officer
301	P.R. Scientific Officer
302	Cytogenetics Scientific Officer
303	P.R. Research Officer
304	Cytogenetics S.O./R.O.
305	Chief Technician
307	Cytogenetics Scientific Officer
308	Cytogenetics Research Officer
309	Cytogenetics Research Officer
313	Unit Director
317	Cytogenetics Section Head
318	Cytogenetics Research Officer
319	Cytogenetics technician
320	Cytogenetics technician
327	Cytogenetics technician
328	Cytogenetics Scientific Officer
329	Cytogenetics S.O./R.O.
331	Photographic technician
332	Photographic technician

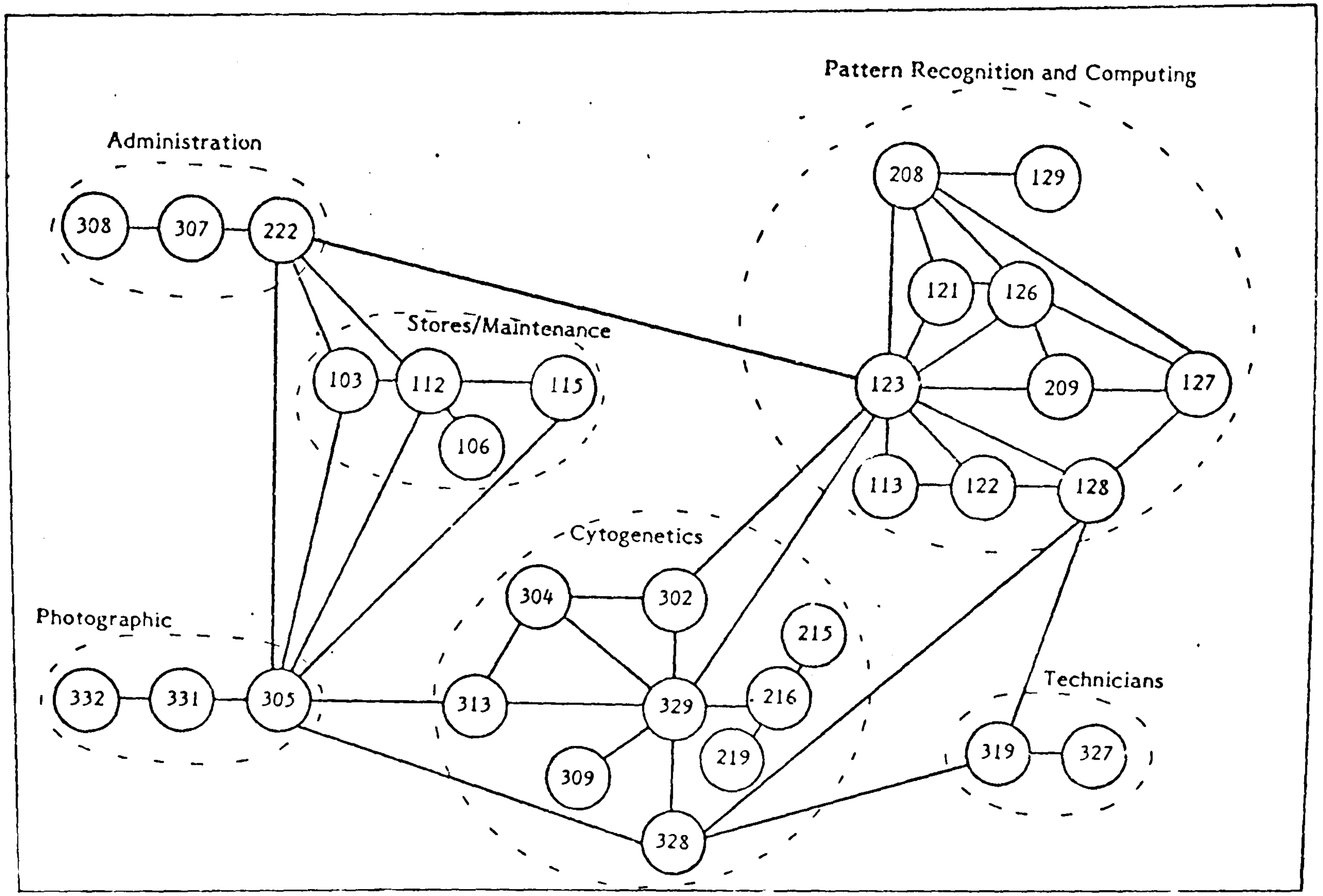


Figure 12.54

Combined interaction diagrams

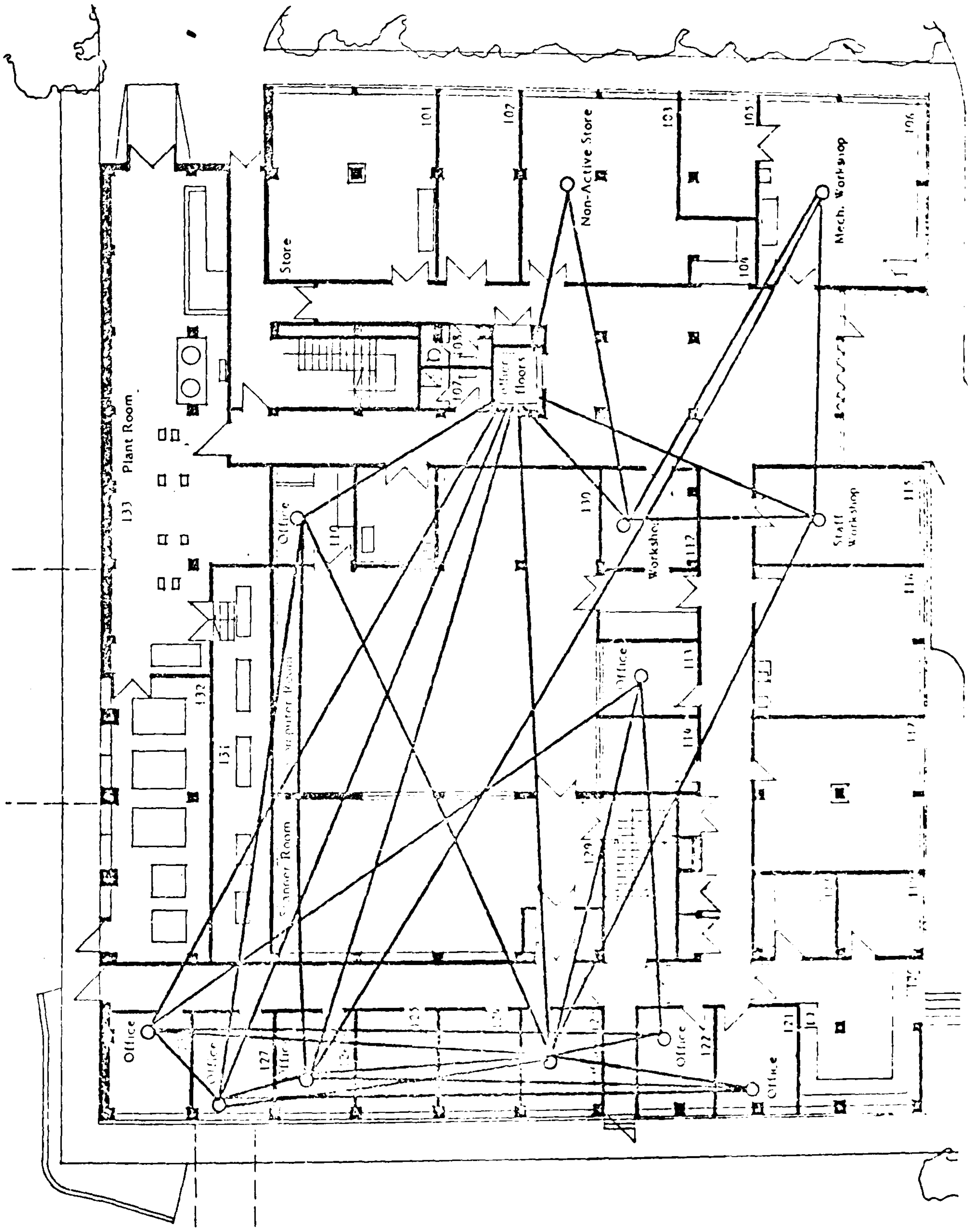


Figure 12.55
Department organisation - individual links, level 1

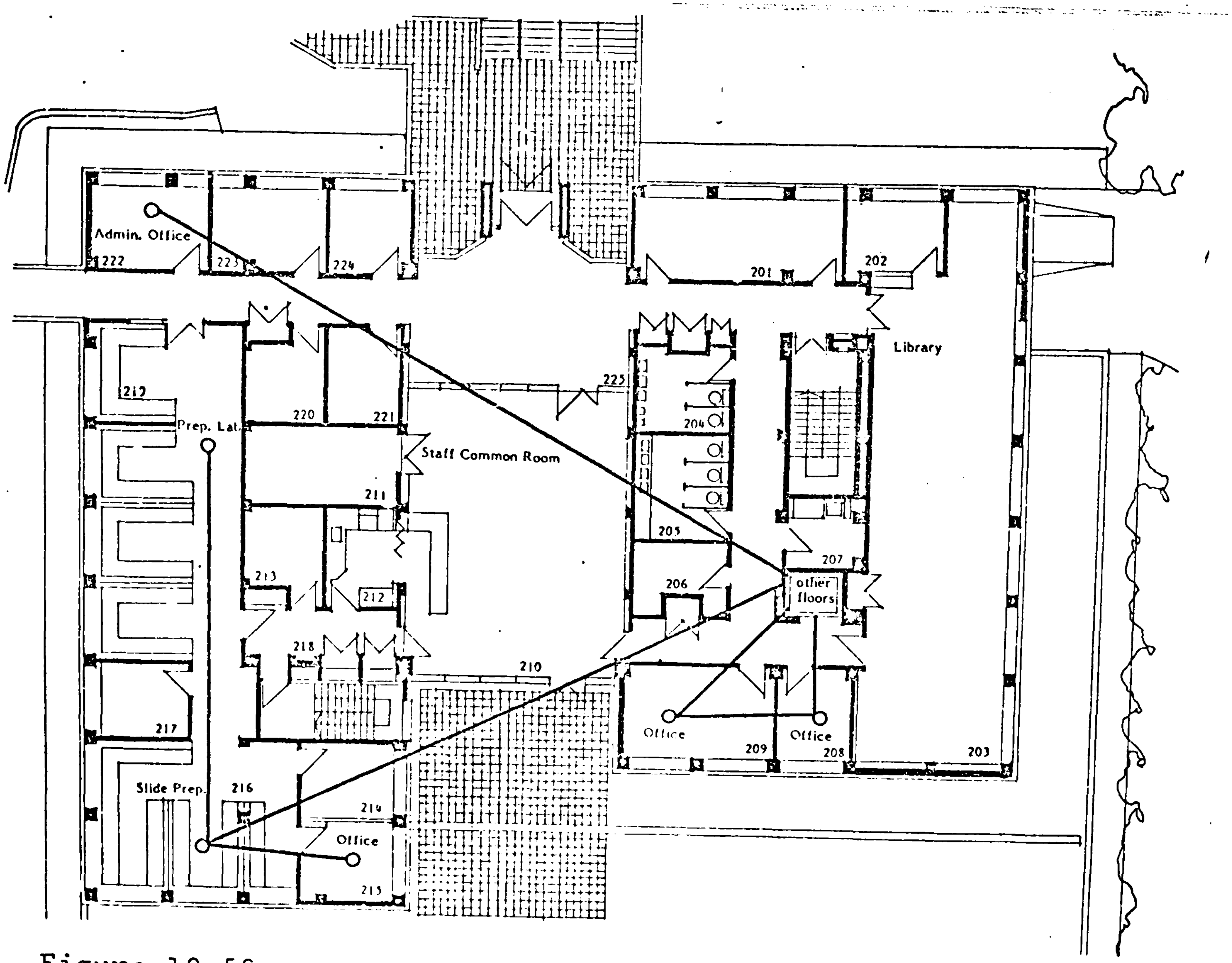


Figure 12.56

Department organisation - individual links, level 2

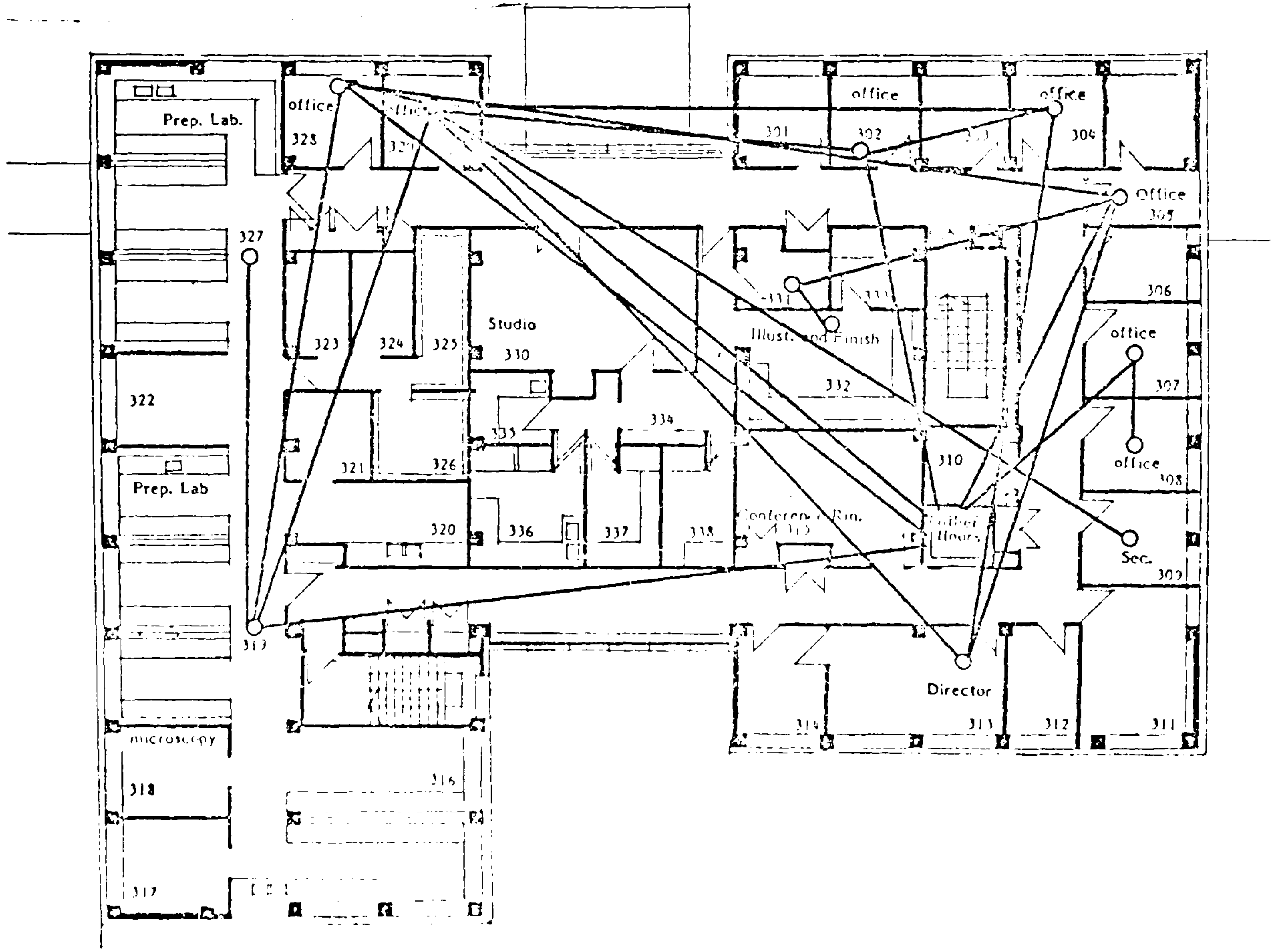
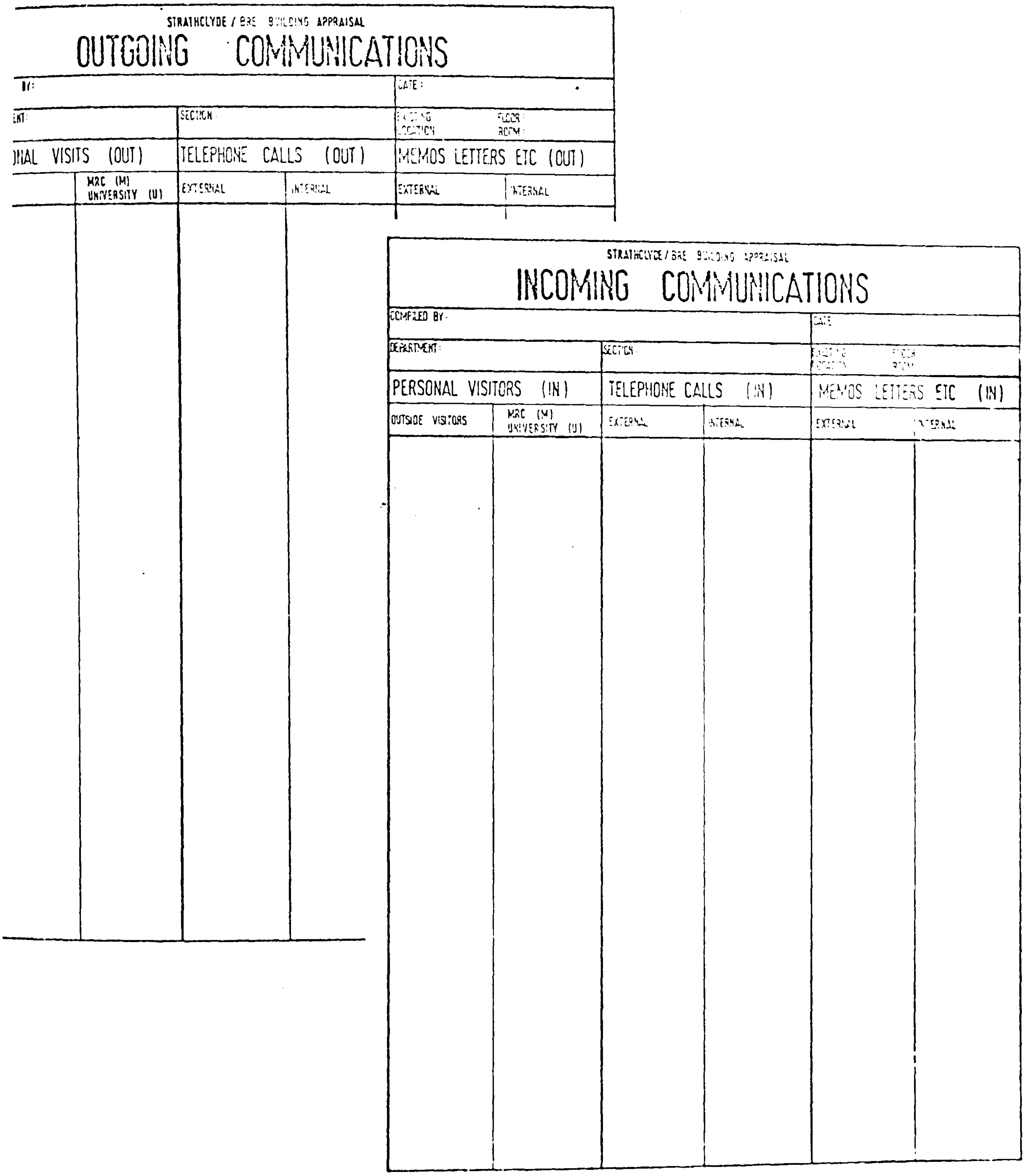


Figure 12.57

Department organisation - individual links, level 3



STRATHCLYDE / BRE BUILDING APPRAISAL

OUTGOING COMMUNICATIONS

Y:

DATE:

ENT:

SECTION:

EXISTING LOCATION

FLOOR: ROOM:

NAL VISITS (OUT)		TELEPHONE CALLS (OUT)		MEMOS LETTERS ETC (OUT)	
	MRC (M) UNIVERSITY (U)	EXTERNAL	INTERNAL	EXTERNAL	INTERNAL

STRATHCLYDE / BRE BUILDING APPRAISAL

INCOMING COMMUNICATIONS

COMPILED BY:

DATE:

DEPARTMENT:

SECTION:

EXISTING LOCATION

FLOOR: ROOM:

PERSONAL VISITORS (IN)		TELEPHONE CALLS (IN)		MEMOS LETTERS ETC (IN)	
OUTSIDE VISITORS	MRC (M) UNIVERSITY (U)	EXTERNAL	INTERNAL	EXTERNAL	INTERNAL

Figure 12.58

Communications analysis: survey forms

12.3.4 Communication Analysis

The perceived formal links just defined are not, of course, the only interactions operating in a building. To determine actual communication links a survey of written and person-to-person communications was undertaken. This survey information was gathered on the forms shown in figure 12.58 to obtain data on information flows and journeys undertaken within the department. The data were subjected to computer analysis to obtain quantitative measures of functional workgroups in the department, defined by their needs for communication.

For comparison, the originally perceived needs of the organisation as evidenced in the room data sheets (forming part of the architects brief) were examined and notes made of the specific required adjacencies. Unfortunately the activity groups defined in the brief are not identical to the currently existing groups but the analyses are comparable in general terms. It is thus possible to check the building plan against the brief and also, using the survey data, to see how the current organisation fits in the building and compares with its own perceived organisation as defined in sections 12.3.1 and 12.3.2.

12.3.5 Analysis Of The Brief Information

Some 79 individual space titles were identified and then inter - relationships specified in the briefing documentation (figure 12.59). In the course of the design a number of these activities were amalgamated, and, in terms of post-hoc analysis, a lot of the fine segregation becomes irrelevant. For the purpose of appraising the design as built it is possible to aggregate the briefing information into the following twelve functional units:

1 Stores

2 Workshop Technicians

3 Maintenance Engineers

4 Pattern Recognition Researchers

5 Reception

6 Second Floor Laboratory Researchers

7 Canteen

8 Administration

9 Senior Technician

10 Third Floor Laboratory Researchers

11 Laboratory Technicians

12 Photographic Technicians

The analysis of this data is shown here in two ways. The first of these, figure 12.60, shows a nonlinear mapping analysis representation of the adjacency matrix in the form of a bubble diagram. Distance between bubbles is proportional to the need for association defined in the brief. Activities shown as close together in the diagram should be close together in the building plan derived from the brief. The second type of analysis, shown in figure 12.61, is a tree-diagram or dendrogram produced from a hierarchical cluster analysis performed on the data. This diagram is read from the top downwards - activities 2 and 4 are the most closely related, then activities 3 and 5, and then 6 and 10. Activity 7 then joins 3 and 5, and so on until all the activities join together to form the complete organisation.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Cytology Reception Area		E																																					
Preparative Laboratory	E		E	E																																			
Slide Preparation		E		E			E																																
Senior Technician			E																																				
Offices, Cytogeneticists		E					E	E																															
Slide Storage							D																																
Scanner Room				D		D		E													E																		
Computer Room							E		D	D						E			D																				
Computer Stationery								D				D																											
Mag. Tape Archives								E																															
DP and Prog Library												D		D	D			D																					
Puch Card & Tape Room								D			D																												
Technical Link								D																															
Office, Scientific								D																															
Office, Comp. Manager								D					D				D		D																				
Office, Engineer								D													D																		
Secretaries														D	D																								
Clerical Assistants											E			D																									
Off Duty Room																																							
Technicians Offices							E														E																		
Electronics Workshop							E	D												E		E																	
E. W. Stores																					E																		
Mechanical Workshop																																							
M. W. Stores																							E																
Drawing Office														D						D																			
Dark Room (Photometric)																					D																		
Prep. Laboratory																													D	E	E	D	D						
Sex Chromatin Labs.																												D											
Office Areas																												E						E					
Tissue Culture																												E											
Media Handling																												D											
Wash-Up																																							
Microscope Areas																												E											
Cold Store																																							
Cytology Section Store																																							
Warm Room																																							
Slide Records Store																																							
Secretarys' Room																												E									D		

E - Essential Link

D - Desirable Link

Figure 12.59

Specified inter-relationships from MRC brief

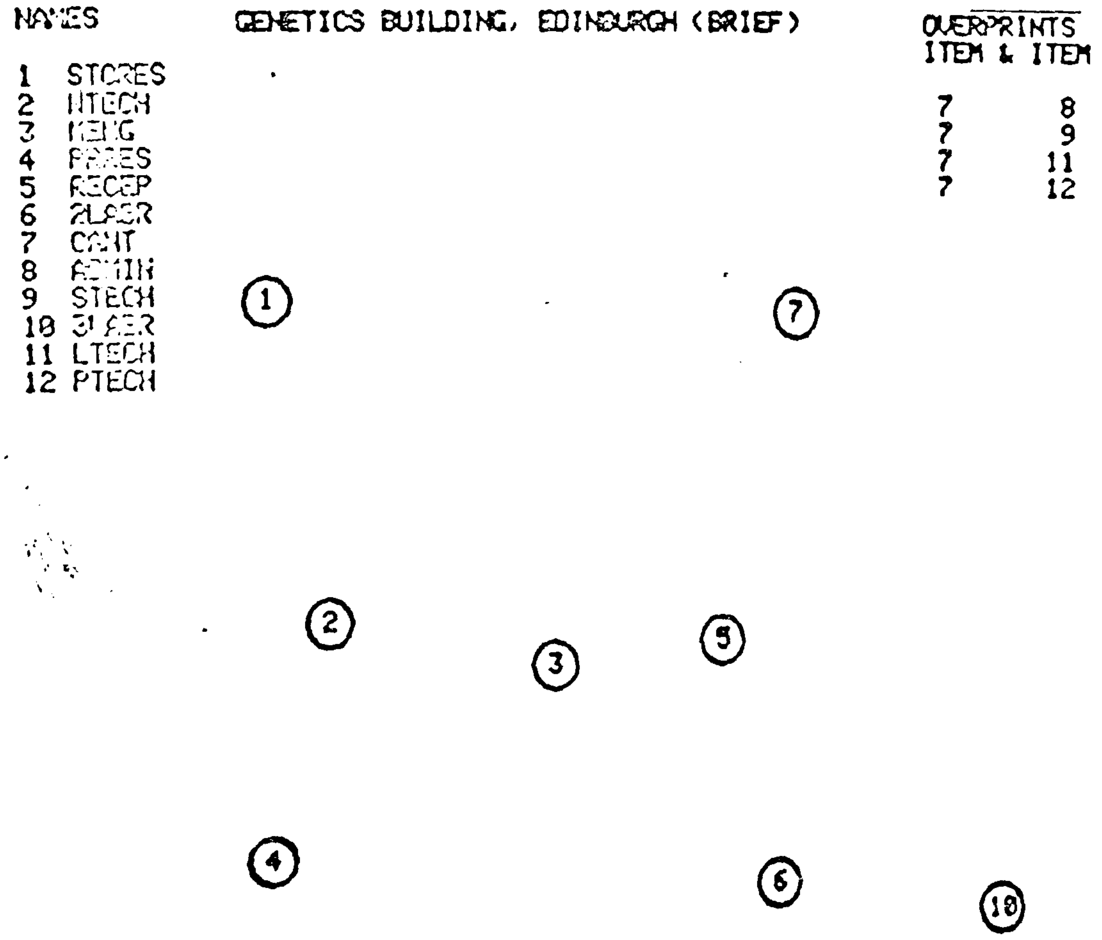
	40	41	42	43	44	45	46	47	48	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	70	71	72	73	75	76	78	79
40 Photographic Studio																																
41 Photomicrography																																
42 Negative Dark Room																																
43 Printing			E			E																										
44 Colour Dark Room																																
45 Finishing & Illustration				E																												
46 Store & Solution Room																																
47 Stores, negatives									D																							
48 Senior Technicians Offices																																
51 Reception Area																																
52 Directors Office												D	D	E																		
53 Admin. Assistants										D				E																		
54 Directors Res. Ass.											E																					
55 Secretarys' Office											E	E																				
56 Registery Admin.												E		E																		
57 Consumables Store																				E												
58 Large Equipment Store																				E												
59 Furniture Store																				E												
60 Gas Cylinder Store																				D												
61 Office																E																
62 Unpacking & Empties																D	D			E												
63 Chemicals - non-inf.																																
64 Chemicals - inf.																					D											
65 Waste Disposal																																
70 Library Reading Space																										E	E	D				
71 Office Workspace																									E		E	D				
72 Reprint Library																									E	E						
73 Stockroom																									D	D						
75 Conference Room																																
76 Staff Common Room																													E			
78 Male Toilets																														D		
79 Female Toilets																																

E - Essential Link

D - Desirable Link

Figure 12.59

Specified inter-relationships from MRC brief (continued)



NAMES

GENETICS BUILDING, EDINBURGH (BRIEF)

1 STORES
2 HTECH
3 HENG
4 PRAES
5 RECEP
6 RLAER
7 CANT
8 ADMIN
9 STECH
10 SLAER
11 LTECH
12 PTECH

OVERPRINTS

ITEM &	ITEM
7	8
7	9
7	11
7	12

Figure 12.60

Bubble diagram of brief data

```
MEDIAN CLUSTERING STRATEGY


PAIRING SEQUENCE

ITEM     JOINS     ITEM   AT  DISTANCE

  2                  4          1.776
  3                  5          1.951
  6                 10          2.496
  3                  7          2.682
  3                  8          2.834
  3                  9          2.878
  3                 11          2.879
  3                 12          2.882
  3                  6          3.141
  2                  3          3.596
  1                  2          3.959

FIT IS  47.% ACCURATE
```

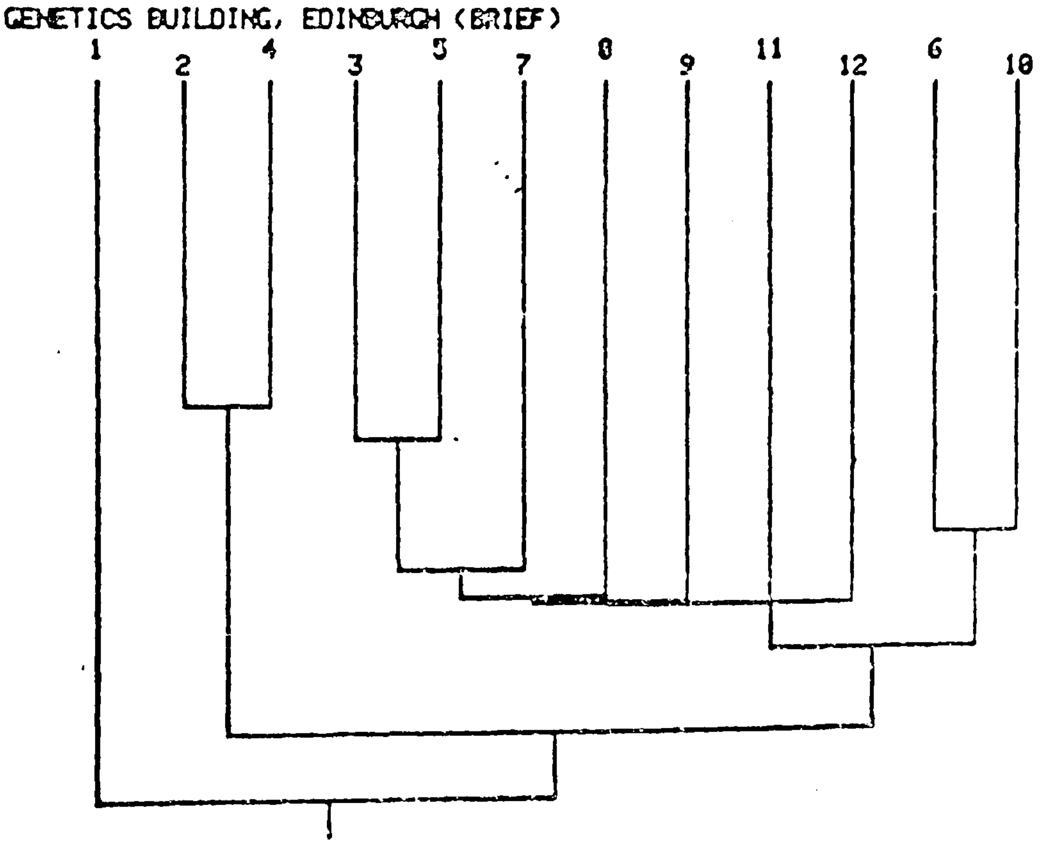


Figure 12.61

Tree diagram of brief data

These diagrams indicate a requirement in the brief for a close connection between activities 2 and 4 (workshop technician and pattern recognition research), activities 6 and 10 (the laboratories) and activities 8, 9, 11 and 12 (administration, senior technician, laboratory technicians, photographic technicians). Activity 7 which the bubble diagram shows as identical to 8, 9, 11, and 12 is the canteen. In terms of the job functions for the canteen staff this placing seems reasonable. All of these groupings are mirrored in the current organisation and to a large extent successfully incorporated into the building design.

12.3.6 Analysis Of Survey Data

Three types of analysis were carried out on the survey data; nonlinear mapping and cluster analysis as before, with the addition of Euclidean cluster analysis to determine the discrete groupings in the organisation. A summary of the survey data is shown in figure 12.62. This shows down the left-hand side the survey sample of 22 individuals who logged all incoming and outgoing communications during the survey period and, along the top, the locations (in the form of room numbers) with which they communicated. In each square under these room numbers the top figure represents the number of communications received from the occupant of that room, whilst the lower figure represents the number of communications issued to the occupant of that room.

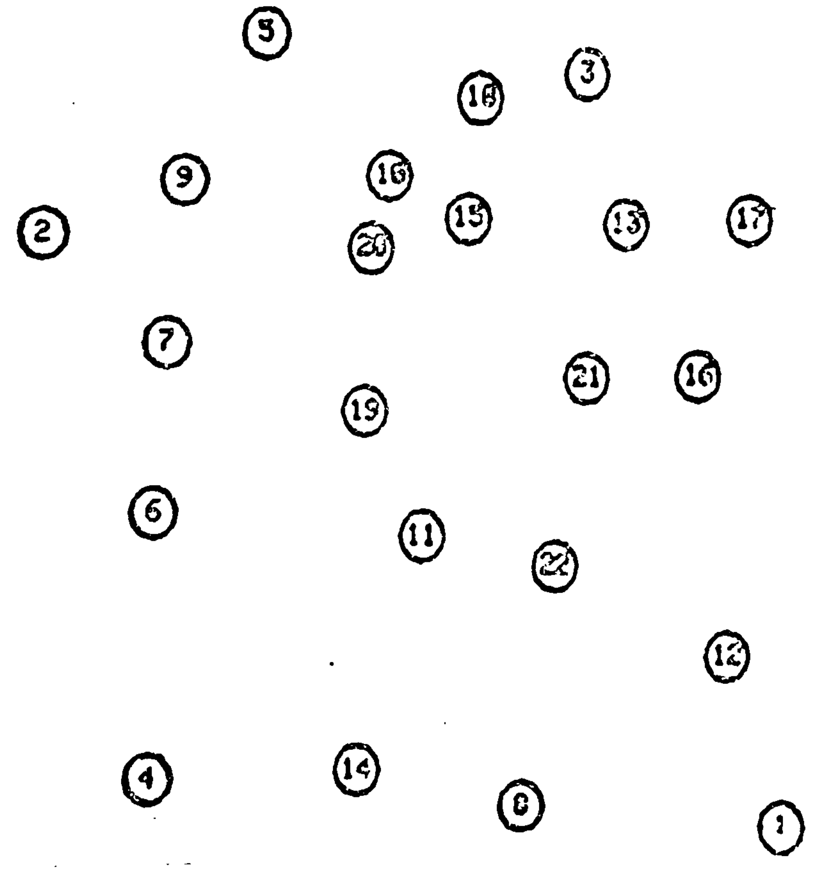
The following output from the computer analysis represents the organisation in a number of different ways. The bubble diagram (figure 12.63) attempts to give an overall picture. Bubble size is not significant, but the distance between bubbles represents strength of communications between activities: closely related individuals (i.e. activities with large numbers of communications) are shown as closely related bubbles. To represent such a complex web of communications in a two-dimensional diagram obviously requires some compromise. This diagram is the best fit to all of the communications shown in the survey data summary, but to obtain detailed information on the strength of links it is necessary to refer to the next diagram (figure 12.64).

NUMBER OF COMMUNICATION $^{a}_{b}$ a FROM ROOM NUMBER SHOWN
b TO ROOM NUMBER SHOWN

LOCATION	103	106	112	115	124	126	127	200	208	210	216	222	304	305	307	309	317	317	319	320	328	331
ACTIVITY	Stores	Workshop Tech.	Maint. Eng.	Tech.	S.O.	S.O.	R.O.	Reception	S.O.	Canteen	R.O.	Admin.	R.O.	S.T. Admin.	S.O.	R.O.	S.O.	R.O.	Lab. Tech.	Lab. Tech.	S.O.	Photo Tech.
101				8																		
103	X	$^{1}_{1}$		$^{4}_{15}$		$^{2}_{3}$		$^{4}_{6}$			$^{1}_{1}$	$^{3}_{2}$		$^{1}_{16}$		$^{1}_{3}$			$^{1}_{4}$			6
104							2															
106	$^{3}_{2}$	X	6	$^{13}_{19}$	2	$^{1}_{2}$	2		2			1		4		1						
112	2	2	X	17				1				2										
113						1																
115	10	$^{4}_{1}$		X		8						2										
116				10										8							3	
117	5	1				17			1					1								
118											1											
120	$^{1}_{1}$			1			6	1	19													1
121	$^{1}_{1}$	$^{2}_{1}$		$^{7}_{1}$		5	$^{2}_{1}$	1				1		1								
122					$^{3}_{5}$	1		3	2			1										
123		1		4				1	2			6		1								
124	1				X		1	1														
125				1	1																	
126	$^{1}_{3}$	4		13		X																
127	1	$^{1}_{1}$		3	3		X		$^{1}_{1}$			1										
128	1				$^{2}_{4}$	$^{1}_{1}$	1		1													
130				5	37	5	4		$^{3}_{6}$					1					4			$^{4}_{4}$
133			1	2		2																
160	1			6																		
200	$^{10}_{7}$							X				6										
201					10			$^{1}_{1}$				3		5								
203	2					3		$^{2}_{3}$			5	1	8	6		8	1		1		8	1
208	2	1		4	$^{6}_{8}$	$^{1}_{3}$	1		X													
209							1															
210	$^{2}_{3}$			$^{1}_{19}$			12	$^{3}_{3}$		X	6	$^{7}_{3}$		12		11				15	15	8
214	1																1	$^{1}_{1}$				
215	3							1				2		1			1					
216	$^{10}_{3}$			4				3			X	2		$^{2}_{5}$				1				
219	$^{7}_{9}$		8					$^{1}_{1}$						2					1			
222	3			19	6			3			1	X		18							1	
224				1								$^{2}_{1}$		2			$^{1}_{1}$					
304	3			1								1	X									
305	7	$^{1}_{1}$		3				4			1	14		X			$^{1}_{1}$		1			
307								1				3			X						1	
308												1										
309	$^{4}_{1}$			2				1								X	$^{7}_{3}$					
310																14						2
313				6							2	$^{2}_{3}$		$^{1}_{5}$			$^{1}_{2}$					
314	1							3				$^{3}_{1}$		2	2						1	$^{1}_{2}$
318			1	2									2	1		1	$^{4}_{2}$	X				$^{8}_{2}$
319			1					1					1	3		$^{1}_{33}$	$^{23}_{26}$	1	X		$^{2}_{31}$	
320		1		2								1				1				X		1
327	$^{10}_{12}$		8	3				4			2		33									
328	5		1					2				1	$^{6}_{1}$			6	2				X	
329	4			$^{1}_{1}$		1		1					$^{9}_{3}$	2		3			1		1	
331	4	1		8		1		$^{3}_{1}$	1					4			1					X
332												1										
C	$^{61}_{60}$	$^{2}_{2}$	9	$^{30}_{33}$		$^{3}_{1}$		$^{17}_{3}$			1	22	4	$^{12}_{23}$		1	1	[illegible]			[illegible]	$^{4}_{6}$
W				1				2				1		$^{4}_{2}$			1					[illegible]
U	$^{2}_{5}$	1		2				2	1			$^{1}_{1}$		$^{2}_{1}$				2	3			

Figure 12.62 - Summary of communications survey data

GENETICS BUILDING, EDINBURGH



GENETICS BUILDING, EDINBURGH

NAMES

1	STORE
2	W.TEC
3	M.ENG
4	TECH
5	SO
6	SO
7	RO
8	RECEP
9	SO
10	CANT
11	RO
12	ADMIN
13	RO
14	ST
15	SO
16	RO
17	SO
18	RO
19	LABT
20	LABT
21	SO
22	PTECH

Figure 12.63

Bubble diagram of survey data

```
GENETICS BUILDING, EDINBURGH

 NAMES                    FURTHEST NEIGHBOUR CLUSTERING STRATEGY

 1  STORE
 2  M.TEC
 3  M.ENG                 PAIRING SEQUENCE
 4  TECH
 5  SO                    ITEM    JOINS     ITEM   AT   DISTANCE
 6  SO
 7  RO                     16                20          151.777
 8  RECEP                  16                15          231.631
 9  SO                     16                18          340.601
10  CANT                   13                21          352.619
11  RO                     13                16          385.877
12  ADMIN                   5                 9          393.618
13  RO                      5                 7          421.197
14  ST                     16                19          437.639
15  SO                      3                13          481.644
16  RO                      2                 6          457.149
17  SO                     16                11          488.939
18  RO                      8                14          493.318
19  LABT                    3                17          513.113
20  LABT                    3                16          517.612
21  SO                      2                 5          555.873
22  PTECH                   3                22          570.638
                            8                12          604.181
                            1                 4          664.521
                            2                 3          669.717
                            1                 8          673.302
                            1                 2          724.534
```

GENETICS BUILDING, EDINBURGH

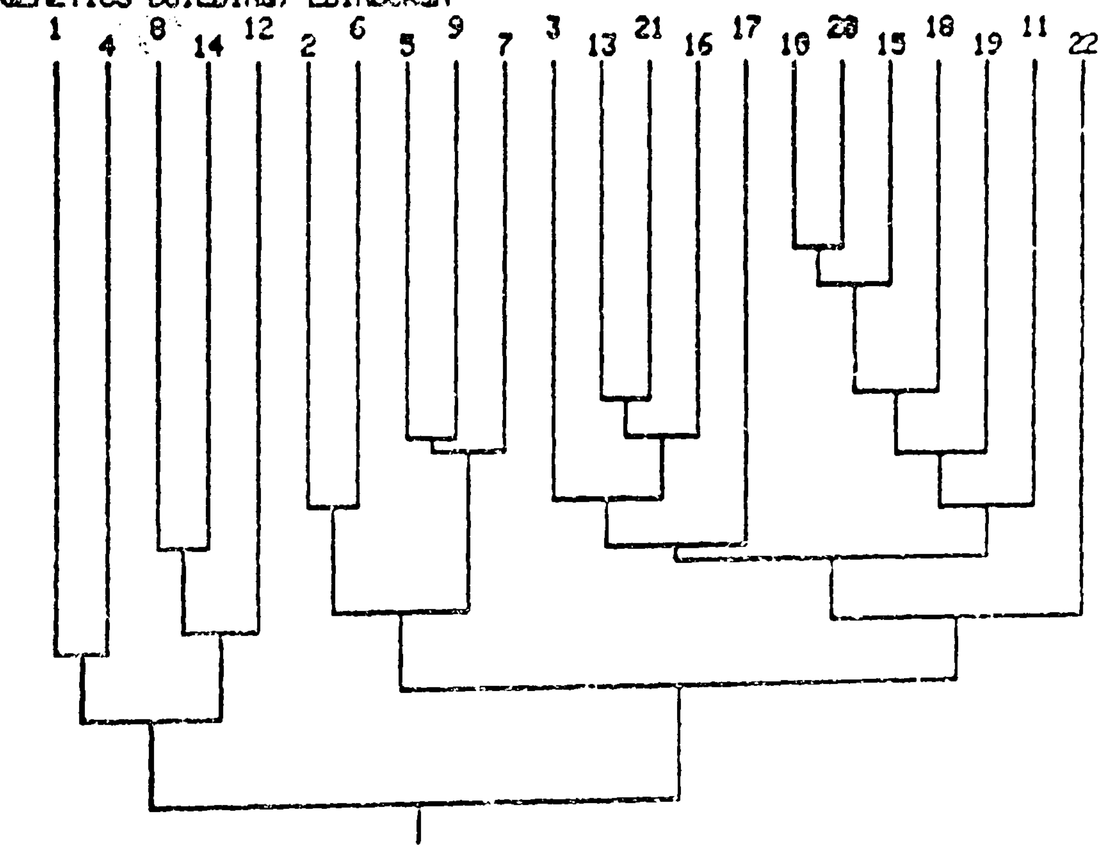


Figure 12.64

Tree diagram of survey data

For comparison with the perceived organisation the individuals in this analysis compare as follows:

	Activity	Room No.
1.	Stores technician	103
2.	Workshop technician	106
3.	Mechanical engineer	112
4.	Technician	115
6.	Scientific Officer (PR)	124
7.	Research Officer (PR)	126
8.	Receptionist	200
9.	Scientific Officer (PR)	208
10.	Canteen	210
11.	Research Officer (C)	216
12.	Admin. Officer	222
13.	Research Officer (C)	304
14.	Senior Technician	305
15.	Scientific Officer (C)	307
16.	Research Officer (C)	309
17.	Scientific Officer (C)	317
18.	Research Officer (C)	318
19.	Laboratory Technician	319
20.	Laboratory Technician	320
21.	Scientific Officer (C)	328
22.	Photographic Technician	331

Figure 12.65

Individuals activity and room number key

The hierarchical cluster analysis attempts to show strength of relations between individuals. In the tree diagram in figure 12.64 each individual is represented along the top of the tree by their key number (see figure 12.65 for key numbers of individuals). The diagram is then read downwards, the most closely related individuals being linked in the diagram nearest the top (the individual level). Finally all individuals become coalesced into groups which eventually merge at the bottom of the diagram (the complete organisation level). Looking more closely we see activities 10 and 20 (the canteen and laboratory technician) are the most closely associated. The next closest link is between activity 15 (cytogenetics scientific officer) and the two activities just merged together (10 and 20). The table in the figure gives the pairing sequence in which the individual activities group together, and the whole picture is represented diagramatically in the tree diagram which is drawn to scale.

The third type of analysis, Euclidean cluster analysis, looks at the grouping of activities in a different way (figure 12.66). The hierarchical cluster analysis showed the way individuals join together to form groups, which then join together to form the organisation as a whole. The Euclidean cluster analysis attempts to identify distinctly seperate groups within the organisation. The computer can find only six discrete groups in the 22 activities included in the survey. In other words if one were seeking to divide the organisation into subgroups then these six groups are the largest number of subgroups it is possible to divide the organisation into without destroying important communication links. In figure 12.66 the bubbles are very self contained with little interrelationship. Successive illustrations (figures 12.67, 68, 69, 70) then show the organisation divided into 5, 4, 3, and finally, just 2 subgroups. In this analysis the bubble size represents group cohesiveness. A large bubble thus represents a distinct group which is only loosely interrelated internally, and a small bubble signifies tight intragroup relationships. The distance between bubbles represents relationships in the same way as in the previous diagrams. The group membership is shown in tabular form on each printout. Throughout all of the Euclidean cluster analysis output groups are tightly cohesive and independent of each other.

Looking at each of the printouts in turn, the interpretation of the first output (with 6 groups) is:

Group 1 - activities 10, 11, 15, 18, 19, 20, 22. Roughly based around a cytogenetics section workgroup.

Group 2 - activities 3, 13, 16, 17, 21. Another cytogenetics workgroup.

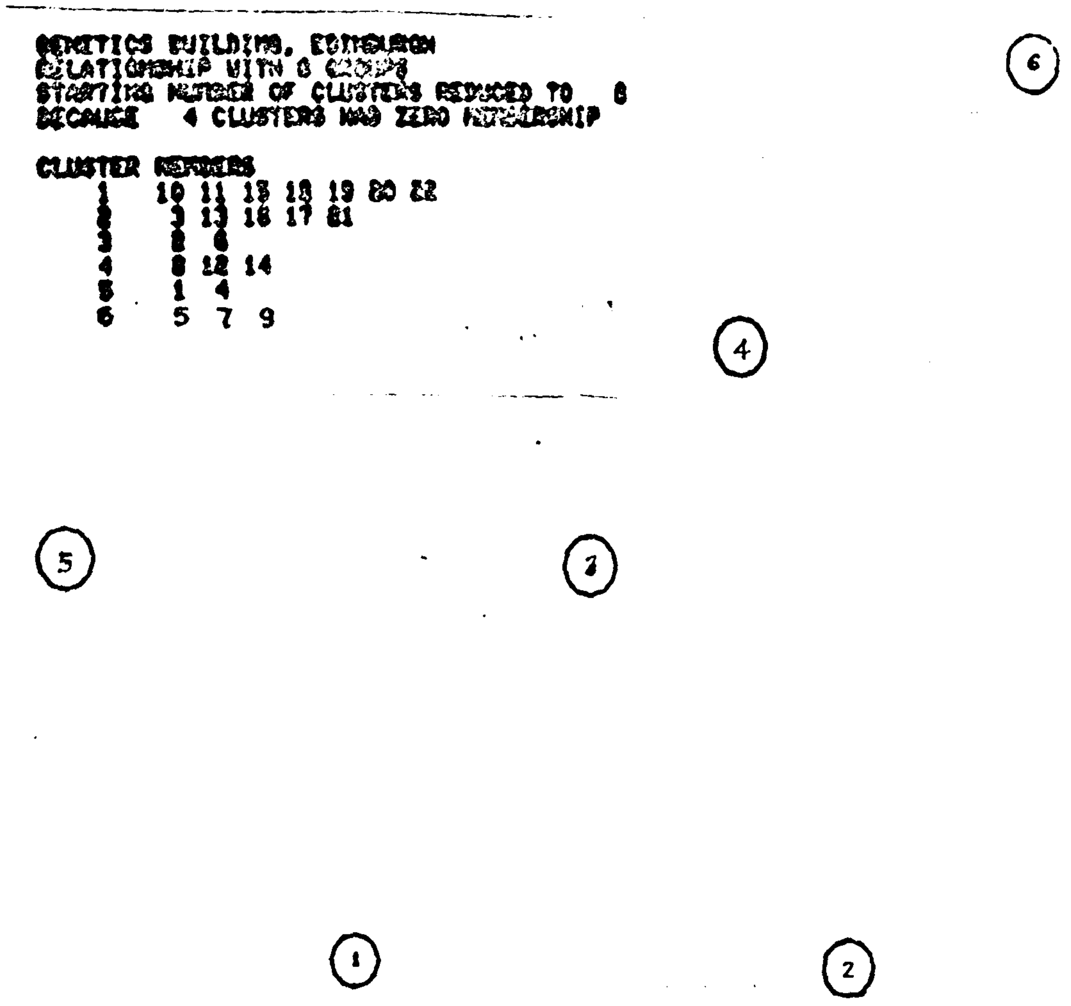
Group 3 - activities 2 and 6. Pattern recognition workshop.

Group 4 - activities 8, 12, 14. Administration.

Group 5 - activities 1 and 4. Stores and maintenance.

Group 6 - activities 5, 7, 9. Pattern recognition workgroup.

At the next iteration groups 5 and 6 merge, showing the strong connection between the stores (particularly) and the pattern recognition groups research. Next groups 1 and 2 merge, showing the whole cytogenetics unit seperate from the rest of the organisation. Activity 5 (PRSO) moves from one pattern recognition workgroup to the other. Then, with three groups a further redistribution occurs, showing the three main functional groups very distinctly as group 1 cytogenetics, group 2 pattern recognition, and group 3 administration. In the final printout with just two groups the scientific staff are shown as being quite distinct from the administrative staff.



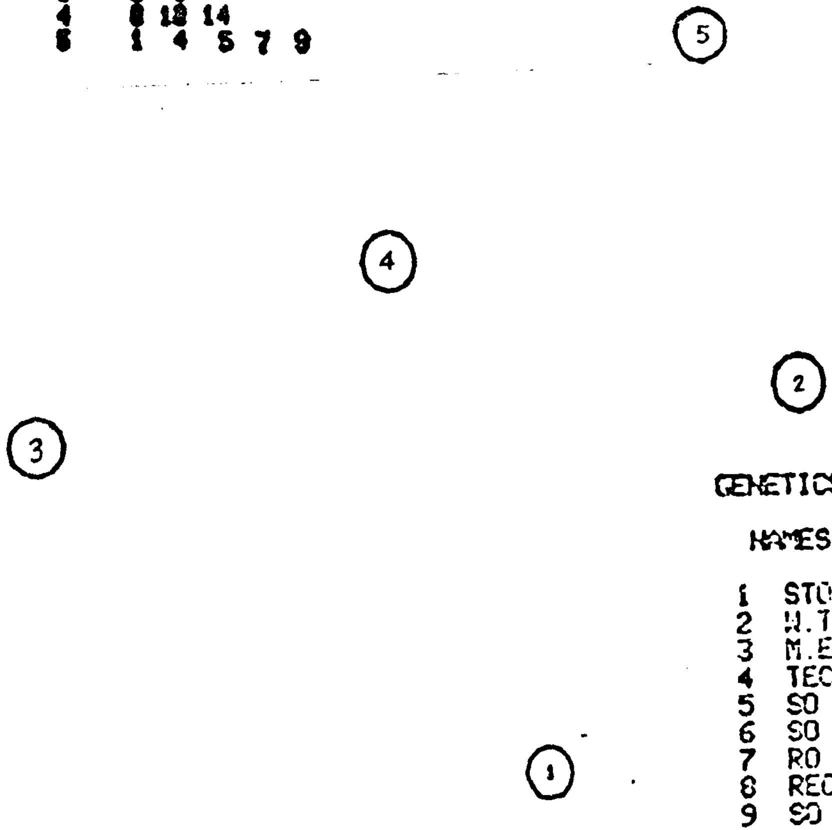
```
GENETICS BUILDING, EDINBURGH
RELATIONSHIP WITH 6 GROUPS
STARTING NUMBER OF CLUSTERS REDUCED TO   6
BECAUSE   4 CLUSTERS HAD ZERO MEMBERSHIP

CLUSTER MEMBERS
     1    10 11 15 18 19 20 22
     2     3 13 16 17 21
     3     2  6
     4     8 12 14
     5     1  4
     6     5  7  9
```

```
GENETICS BUILDING, EDINBURGH

  NAMES

 1  STORE
 2  W.TEC
 3  M.ENG
 4  TECH
 5  SO
 6  SO
 7  RO
 8  RECEP
 9  SO
10  CANT
11  RO
12  ADMIN
13  RO
14  ST
15  SO
16  RO
17  SO
18  RO
19  LABT
20  LABT
21  SO
22  PTECH
```

Figure 12.66

Seperate groups analysis of survey data - 6 groups

```
GENETICS BUILDING, EDINBURGH
RELATIONSHIP WITH 5 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   5 AND   6
CLUSTER MEMBERS
    1    10 11 15 18 19 20 22
    2     3 13 16 17 21
    3     2  6
    4     8 12 14
    5     1  4  5  7  9
```

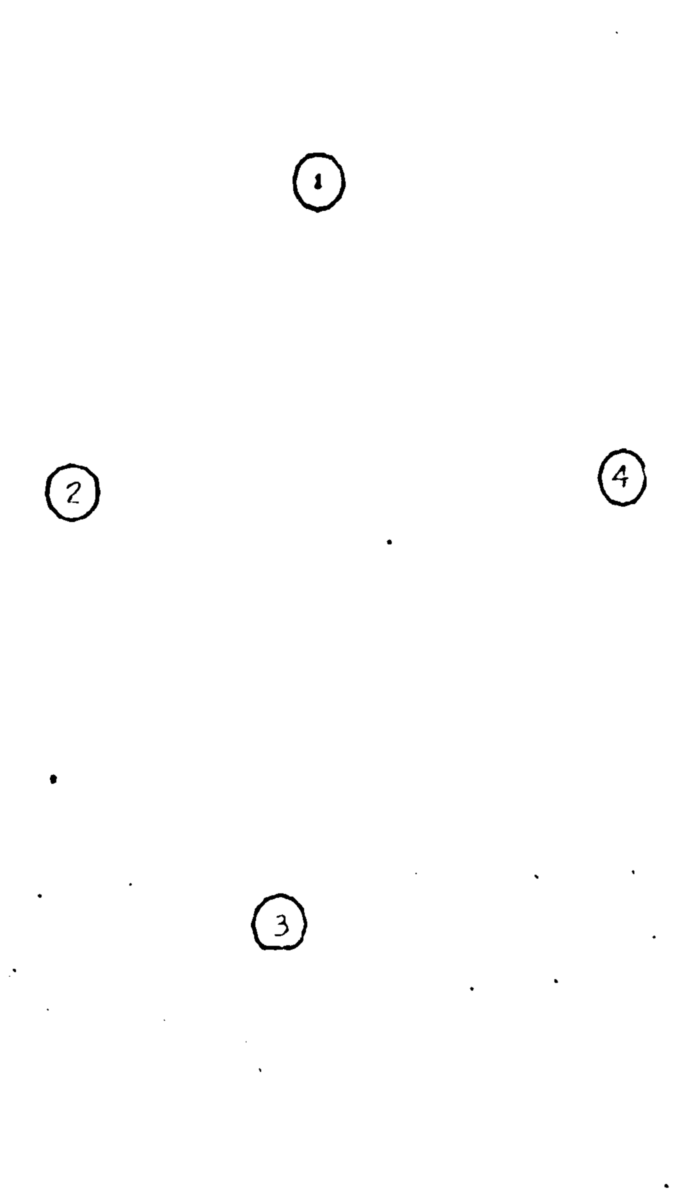


```
GENETICS BUILDING, EDINBURGH

 NAMES

 1  STORE
 2  N.TEC
 3  M.ENG
 4  TECH
 5  SO
 6  SO
 7  RO
 8  RECEP
 9  SO
10  CANT
11  RO
12  ADMIN
13  RO
14  ST
15  SO
16  RO
17  SO
18  RO
19  LABT
20  LABT
21  SO
22  PTECH
```

Figure 12.67

Seperate groups analysis of survey data - 5 groups

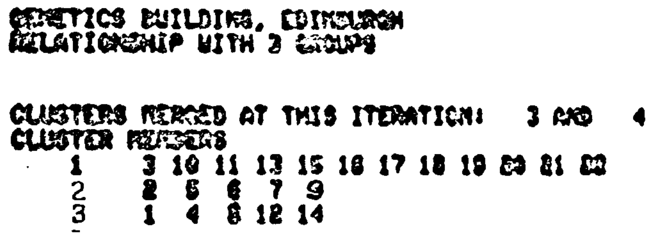
```
GENETICS BUILDING, EDINBURGH
RELATIONSHIP WITH 4 GROUPS

CLUSTERS FORMED AT THIS ITERATION:   1 AND   8
CLUSTER NUMBERS
     1     3 10 11 13 15 16 17 18 19 20 21 22
     2     2  5  6
     3     8 12 14
     4     1  4  7  9
```



```
GENETICS BUILDING, EDINBURGH

  NAMES

  1  STORE
  2  W.TEC
  3  M.ENG
  4  TECH
  5  SO
  6  SO
  7  RO
  8  RECEP
  9  SO
 10  CANT
 11  RO
 12  ADMIN
 13  RO
 14  ST
 15  SO
 16  RO
 17  SO
 18  RO
 19  LABT
 20  LABT
 21  SO
 22  PTECH
```

Figure 12.68

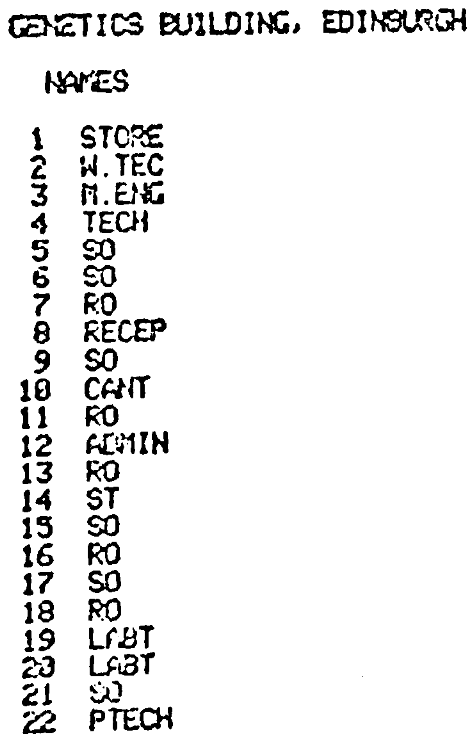
Seperate groups analysis of survey data - 4 groups



```
GENETICS BUILDING, EDINBURGH
RELATIONSHIP WITH 3 GROUPS

CLUSTERS MERGED AT THIS ITERATION:   3 AND   4
CLUSTER MEMBERS
    1     3 10 11 13 15 16 17 18 19 20 21 22
    2     2  5  6  7  9
    3     1  4  8 12 14
```

1

3



```
GENETICS BUILDING, EDINBURGH

  NAMES

 1  STORE
 2  W.TEC
 3  M.ENG
 4  TECH
 5  SO
 6  SO
 7  RO
 8  RECEP
 9  SO
10  CANT
11  RO
12  ADMIN
13  RO
14  ST
15  SO
16  RO
17  SO
18  RO
19  LABT
20  LABT
21  SO
22  PTECH
```

2

Figure 12.69

Seperate groups analysis of survey data - 3 groups

```
GENETICS BUILDING, EDINBURGH
RELATIONSHIP WITH 2 GROUPS

CLUSTERS FORMED AT THIS ITERATION:   1 AND   2
CLUSTER MEMBERS
    1     2  3  5  6  7  9 10 11 13 15 16 17 18 19 20 21 22
    2     1  4  8 12 14
```

```
GENETICS BUILDING, EDINBURGH

  NAMES

 1  STORE
 2  W.TEC
 3  M.ENG
 4  TECH
 5  SO
 6  SO
 7  RO
 8  RECEP
 9  SO
10  CANT
11  RO
12  ADMIN
13  RO
14  ST
15  SO
16  RO
17  SO
18  RO
19  LABT
20  LABT
21  SO
22  PTECH
```

Figure 12.70

Seperate groups analysis of survey data - 2 groups

Having worked through all these analyses it is constructive to reconsider them in conjunction. The two forms of cluster analysis particularly help in the interpretation of the initial bubble diagram. Looking again at the tree diagram (figure 12.71) the three main groups are easily identifiable. The breakdown of the administration group into the storeman and technician and the more exclusively deskbound jobs of administration officer, reception and chief technician is clearly shown. The two groups in the pattern recognition group and the cytogenetics group are again easily seen, as is the final division into scientific and non-scientific staff.

The bubble diagram emphasises the very segmented nature of the organisation with no easily discernable pattern. However, in the light of the other analysis the groupings become more apparent and are shown in figure 12.72.

```
GENETICS BUILDING, EDINBURGH
 NAMES                         FURTHEST NEIGHBOUR CLUSTERING STRATEGY

 1  STORE
 2  N.TEC                      PAIRING SEQUENCE
 3  M.ENG
 4  TECH                       ITEM    JOINS    ITEM   AT   DISTANCE
 5  SO
 6  SO                          10               20         191.777
 7  RO                          10               15         331.031
 8  RECEP                       10               18         343.061
 9  SO                          13               21         349.633
10  CANT                        13               16         355.877
11  RO                           5                9         393.815
12  ADMIN                        5                7         401.137
13  RO                          10               19         407.829
14  ST                           3               13         431.644
15  SO                           2                6         457.143
16  RO                          10               11         465.000
17  SO                           8               14         483.919
18  RO                           3               17         503.113
19  LABT                         3               10         517.078
20  LABT                         2                5         555.233
21  SO                           3               22         579.032
22  PTECH                        8               12         584.101
                                 1                4         604.501
                                 2                3         644.337
                                 1                8         673.233
                                 1                2         704.534
```

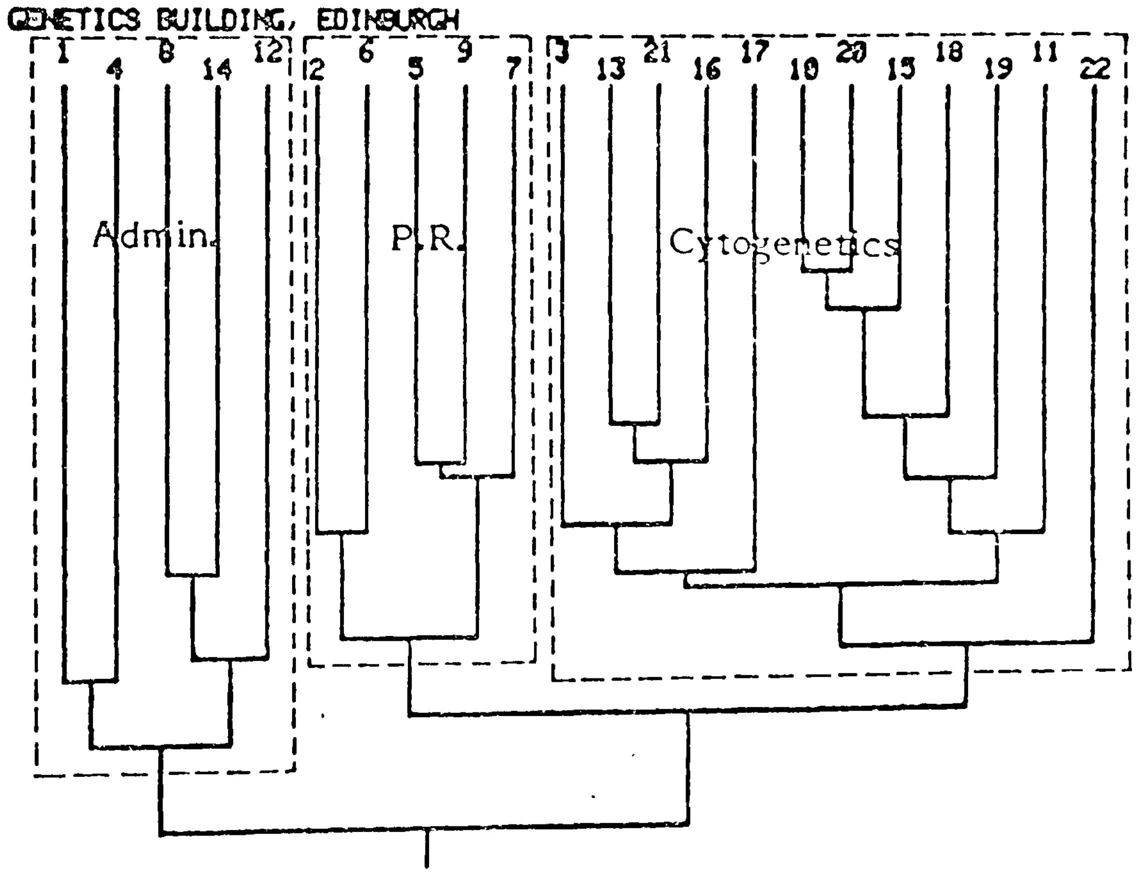
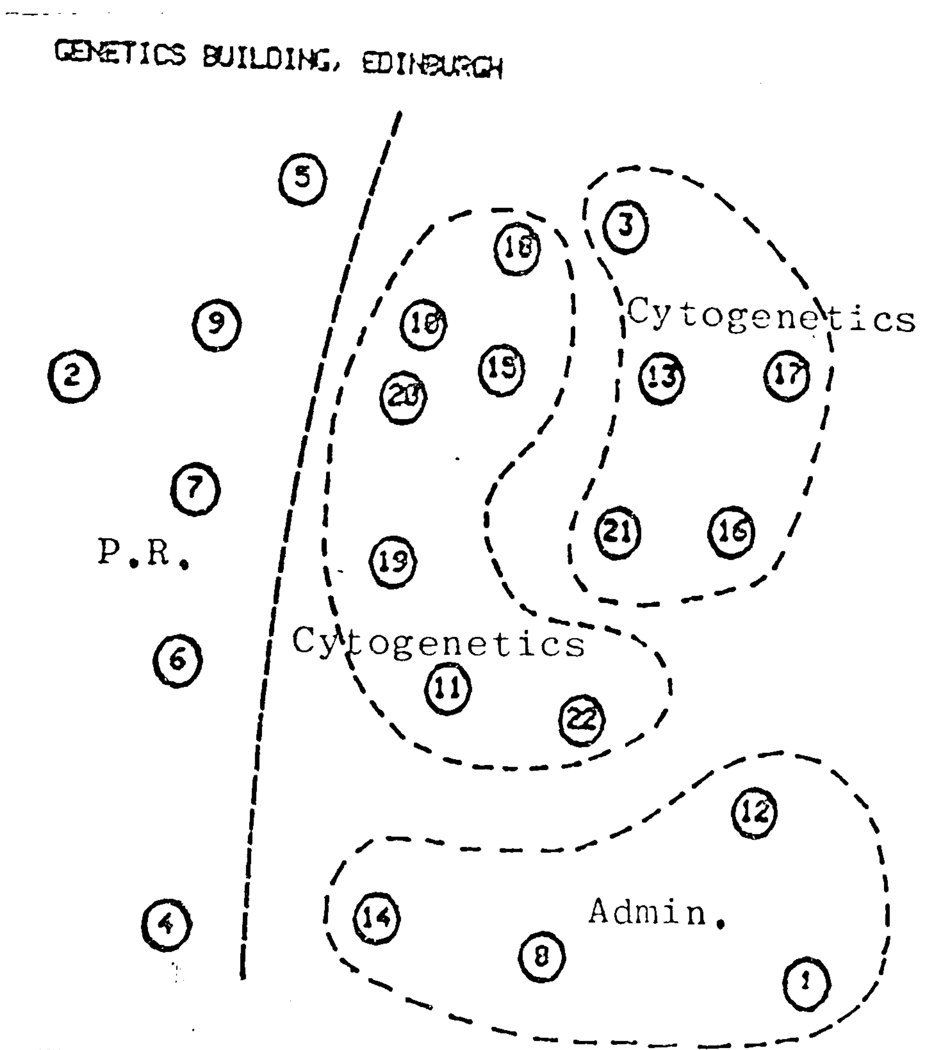


Figure 12.71

Tree diagram of survey data, with main groupings indicated



GENETICS BUILDING, EDINBURGH

NAMES

1	STORE
2	W.TEC
3	M.ENG
4	TECH
5	SO
6	SO
7	RO
8	RECEP
9	SO
10	CANT
11	RO
12	ADMIN
13	RO
14	ST
15	SO
16	RO
17	SO
18	RO
19	LAST
20	LABT
21	SO
22	PTECH

Figure 12.72

Bubble diagram of survey data, with main groupings indicated

As different data were used to obtain the diagrams of perceived interrelations, brief analysis interrelations, and actual interrelations as surveyed, cross comparisons are difficult to make. However the overwhelming opinion must be that the main functional units defined in the brief have remained constant. Provided facilities are available for their own specialist work the individuals concerned show little need for interrelations with other groups or any further support from the building.

CHAPTER 13

CONCLUSION

13.1 SUMMARY

The theoretical basis of a statistical approach to problem structuring has been established and its application shown in an architectural example. The embodiment of the techniques into a computer program with a sophisticated user interface has demonstrated that these advanced techniques may be made available for use by "mathematically naive" users. The provision of a range of complimentary techniques enables a number of insights to be gained into the data structure, defining its overall interrelationships (nonlinear mapping, principal co-ordinates analysis), its discrete components (Euclidean cluster analysis), and its hierarchical structure (hierarchical cluster analysis).

13.2 POSSIBLE EXTENSIONS OF THE WORK

The number of techniques which may be brought to bear on this problem obviously include more methods than have been discussed here. The discussion has been deliberately constrained to keep the size of the thesis within reasonable bounds, but the modular nature of MAGIC will enable further techniques to be added as and when they are required.

One of the more interesting developments may be to extend the range of classification techniques to include non-exclusive (overlapping) techniques. MAGIC indicates "shades of classifications" by presenting different analyses which may be compared, thus enabling alternative possible structures to be identified. The mapping techniques, whilst presenting an overall picture, enable groupings to be identified by eye and possible overlapping clusters or alternative classifications to be evaluated in conjunction with the clustering output. However, in some circumstances, it may be considered inappropriate to insist that an object should belong to only one group and the incorporation of a non-exclusive classification technique would then be desirable.

There is a growing body of theory on the subject of overlapping classifications. One of the major contributions is by Jardine and Sibson (1968). They describe a sequence of clustering methods which they call "beta dendrograms"; however, the algorithm they describe for obtaining the β_k clusters makes heavy demands on computing time, and, additionally, the results are very difficult to assimilate for larger data sets. At the moment it seems that the most feasible interactive use of this technique would be the post-hoc analysis of a small data set extracted from the complete data set by, say, the Euclidean cluster analysis already in MAGIC.

Another technique which attempts to incorporate objects whose group membership may be variable is fuzzy clustering. The concept behind this approach is relatively simple: a probability density function P(x) is assumed known, $P(S_m|x_i)$ then denotes the "degree of belongingness" of the vector x_i to the class S_m. However, as the approach seeks to optimise an intuitively derived criterion a number of statistical problems arise in comparing alternative solutions. More recently attempts have been made to incorporate a fuzzy clustering concept into the sum of squares clustering criterion, but until some fundamental mathematical problems have been solved any development for incorporation within MAGIC must be viewed with caution.

One further interesting extension of the classificatory procedures would be constrained classification. If one has predetermined requirements (or external information which makes the imposition of constraints appropriate) then it may be of use to be able to constrain the set of allowable classifications.

13.3 FURTHER APPLICATIONS

The examples of the application of MAGIC described here are in the context of layout planning and post-occupancy evaluation. The sophisticated multivariate clustering techniques incorporated in MAGIC are, however, capable of dealing with a much broader range of general analysis. For example, in the subject area of building costing, MAGIC has already been used to analyse historic cost data to find the major cost determinants, which are then used to predict capital costs of similar new buildings. In a social science application MAGIC has been used to analyse survey data collected over a wide range of incompatible measures (age, occupation, address, etc) and sensibly analysed what appeared, previously, to be quite intractable data. It is hoped, therefore, that MAGIC will be of use in almost any situation where someone who knows a lot about the data, but little about statistics, wishes to subject multivariate data to exploratory data analysis. The emphasis on the use of visual displays to reveal the structure of the data helps the user insofar as he may concentrate on the interpretation of the picture in terms of his application, rather than on the interpretation of abstract statistics. To the extent that there is an isomorphism between the elements and interrelations of the data and the representation spaces via the intermediary of an appropriate set of transformations in data analysis then the task of interpretation is made that much easier.

REFERENCES

ADAMS, E.N. (1972). Consensus techniques and the comparison of taxonomic trees. Systematic Zoology, vol. 21, pp 390 - 397.

ALEXANDER, C. (1964). Notes on the Synthesis of Form. Harvard University Press. Cambridge, MA.

ALEXANDER, C. (1965). A city is not a tree. Architectural Forum, vol. 122, no. 1, pp 58 - 62, and no. 2, pp 58 - 61.

ALEXANDER, C. (1971). The state of the art in design methods. DMG Newsletter vol. 5, no 3. Later reprinted as Max Jacobson interviews Christopher Alexander in Architectural Design, Dec. 1971, pp 768 - 770.

ANDERSON, A.J.B. (1971). Ordination methods in ecology. Journal of Ecology, vol. 59, no. 3, Nov. 1971, pp 713 - 726.

ANDREWS, H.C. (1972). Introduction to Mathematical Techniques in Pattern Recognition. Wiley Interscience, NY.

ARCHER, L.B. (1979). Design as a discipline: the three R's. Design Studies, vol. 1, no. 1 (July), pp

17 - 20.

ARMOUR, G.C., and E.S. BUFFA (1963). A heuristic algorithm and simulation approach to relative location of facilities. Management Science, vol. 9, pp 294 - 309.

AUSTIN, M.P. and L. ORLOCI (1966). Geometric models in Ecology II - an evaluation of some ordination techniques. Journal of Ecology, vol. 54, no. 1,. pp 217 - 227.

BAKER, F.B. and L.J. HUBERT (1975). Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association. vol. 70, no. 349, March 1975, pp 31 - 38.

BALCHIN, W.G.V. (1972). Graphicacy. Geography, vol. 57, pp 185 - 195.

BARTLETT, M.S. (1951). The effect of standardisation of an approximation in factor analysis. Biometrika, vol. 38, pp 337 - 344.

BEALE, E.M.L. (1969). Euclidean cluster analysis. Bulletin of the International Statistics Institute, vol. 43, no. 2, pp 92 - 94.

BEAUMONT, M.J.S. (1967). Computer-aided techniques for the synthesis of layout and form with respect to circulation. PhD thesis, University of Bristol.

BELLMORE, M., and J.C. MALONE (1971). Pathology of traveling salesman subtour elimination algorithms. Operations Research, vol. 19, p 1766.

BELLMORE, M., and G.L. NEMHAUSER (1968). The travelling salesman problem: a survey. Operations Research, vol. 16, p 538.

BLACKITH, R.E. and R.A. REYMENT (1971). Multivariate Morphometrics. Academic Press, London.

BLINN, J.F. (1977). A homogenous formulation for lines in 3-space. Computer Graphics, vol. 11, no. 2, Summer 1977, pp 237 - 241.

BOYCE, A.J. (1969). Mapping diversity: a comparative study of some numerical methods. In Cole (1969).

BRIDGES, A.H. (1978). Reanimation of the Eastern Electricity Board Central Accounting Office: office layout using the MAGIC program. ABACUS report R45, University of Strathclyde.

BROADBENT, G. (1979). The development of design methods. Design Methods and Theories, vol. 13, no. 1,

pp 41 - 45.

BROADBENT, G. and J. O'KEEFE (1972). An appraisal of the role of space planning techniques within the design process. Preprints of the RIBA Conference, Computers in Architecture, York 1972, pp 110 - 117.

BURR, E.J. (1968). Cluster sorting with mixed character types I. Standardisation of character values. Australian Computer Journal, vol. 1, pp 97 - 99.

BURR, E.J. (1970). Cluster sorting with mixed character types II. Fusion strategies. Australian Computer Journal, vol. 2, pp 98 - 103.

CALINSKI, T. and J. HARABASZ (1974). A dendrite method for cluster analysis. Communications in Statistics, vol. 3, pp 1 - 27.

CALVERT, T.W. and T.Y. YOUNG (1969). Randomly generated nonlinear transformations for pattern recognition. IEEE Transactions on Systems Science and Cybernetics, vol. ssc-5 no. 4, Oct. 1969, pp 266 - 273.

CARRIE, A.S. (1973). Numerical taxonomy applied to group technology and plant layout. International

Journal of Production Research, vol. 11, no. 4, pp 399 - 416.

CARTER, D.J. and B. WHITEHEAD (1975a). Data for generative layout planning programs. Building Science, vol. 10, pp 95 - 102.

CARTER, D.J. and B. WHITEHEAD (1975b). The use of cluster analysis in multi-storey layout planning. Building Science, vol. 10, pp 287 - 296.

CARTER, D.J. and B. WHITEHEAD (1976). A study of pedestrian movement in a multi-storey office block. Building and Environment, vol. 11, no. 4, pp 239 - 247.

CHANG, C.L. and R.C.T. LEE (1973). A heuristic relaxation method for nonlinear mapping in cluster analysis. IEEE Transactions on Systems, Man, and Cybernetics, vol. smc-3, no. 2, March 1973, pp 197 - 200.

CHRISTOFIDES, N. (1975). Graph Theory: an Algorithmic Approach. Academic Press, London.

COLE, A.J. (ed) (1969). Nunerical Taxonomy. Academic Press, London.

CORMACK, R.M. (1971). A review of classification.

Journal of the Royal Statistical Society, vol. A134, pp 321 - 367.

CORSTEN, L.C.A. and T. POSTELNICU (eds) (1975). Proceedings of the 8th International Biometric Conference.

COUSIN, J. (1970). Topological organisation of architectural space. Architectural Design, October 1970, pp 491 - 493.

CROSS, N. (ed) (1972). Design Participation. Academy Editions, London.

CROSS, N., J. NAUGHTON and D. WALKER (1981). Design method and scientific method. Design Studies, vol. 2, no. 4, October 1981, pp 195 - 201.

CUNNINGHAM, K.M. and J.C. OGILVIE (1972). Evaluation of hierarchical grouping techniques: a preliminary study. Computer Journal, vol. 15, no. 3, pp 209 - 213.

DARKE, J. (1979). The primary generator and the design process. Design Studies, vol. 1, no. 1, pp 36 - 44.

DORN, W.S. and D.D. MCCRACKEN (1972). Numerical Methods With FORTRAN IV Case Studies. Wiley, NY.

DUBES, R. and A.K. JAIN (1976). Clustering techniques: the users dilemma. Pattern Recognition, vol. 8, pp 247 - 260.

DUDNIK, E.E. and R. KRAWCZYK (1973). An evaluation of space planning methodologies. Proceedings of EDRA 4 (ed W.F.E. Preiser), pp 414 - 427.

EADES, D.C. (1965). The inappropriatness of the correlation coefficient as a measure of taxonomic resemblence. Systematic Zoology, vol. 14, no. 2, pp 98 - 100.

EASTMAN, C. (1972a). Logical methods of building design: a synthesis and review. Design Research and Methods, vol. 6, no. 3, pp 79 - 90.

EASTMAN, C. (1972b). Preliminary report on a system of general space planning. Communications of the ACM, vol. 15, no. 2, pp 76 - 87.

EILON, S., C.D.T. WATSON-GANDY and N. CHRISTOFIDES (1971). Distribution Management: Mathematical Modelling and Practical Analysis. Griffin, London.

ENSLEIN, K., A. RALSTON and H.S. WILF (eds) (1977). Statistical Methods for Digital Computers. Wiley, NY.

FARRIS, J.S. (1969). On the cophenetic correlation coefficient. Systematic Zoology, vol. 18, no. 3, pp 279 - 285.

FASHAM, M.J.R. (1977). A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology, vol. 58, pp 551 - 561.

FERGUSON, E.S. (1977). The mind's eye: non-verbal thought in technology. Science, 26 August 1977, pp 827 - 836.

FEYERABEND, P. (1975). Against Method. New Left Books, London.

FISHER, L. and J.W. VAN NESS (1971). Admissable clustering procedures. Biometrika, vol. 58, no. 1, pp 91 - 104.

FLOREK, K., J. LUKASZEWICZ, J. PERKAL, H. STEINHAUS and S. ZUBRZYCKI (1951). Sur la liason et la division des points d'un ensemble fini. Colloquium Mathematicum, vol. 2, pp 282 - 285.

FOLEY, D.H. (1972). Considerations of sample and feature size. IEEE Transactions on Information Theory, vol. it-18, no. 5, pp 618 - 626.

FORGEY, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classification. Biometrics, vol. 21, pp 235 - 250.

FORTIN, G. (1978). BUBBLE: Relationship diagrams using iterative vector approximation. Proceedings of 5th Design Automation Conference, Las Vegas, June 1978.

FOULDS, L.R. and D.F. ROBINSON (1976). A strategy for solving the plant layout problem. Operational Research Quarterly, vol. 27, no. 4(i), pp 845 - 855.

FREW, R.S. (1976). Clustering methods as partial solutions to the space allocation problem. Proceedings of the 13th Design Automation Conference, S.F., pp 16 -21.

FREW, R.S., R.K.RAGADE and P.H.ROE (1972). The animals of architecture. Proceedings of EDRA 3, pp 23.2.1 - 23.2.7.

FRIEDMAN, H.P. and J. RUBIN (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, vol. 62, pp 1159 - 1178.

GAWAD, M.T.A. and B. WHITEHEAD (1976). Addition of communication paths to diagrammatic layouts. Building and Environment, vol. 11, pp 249 - 258.

GEORGE, J.E. (1975). Algorithms to reveal the representation of characters, integers and floating point numbers. ACM Transactions on Mathematical Software, vol. 1, no. 3, pp 210 - 216.

GNANADESIKAN, R. and M.B. WILK (1969). Data analytic methods in multivariate statistical analysis. In Krishnaiah (ed) (1969), pp 593 - 638.

GOODALL, D.W. (1966). A new similarity measure based on probability. Biometrics, vol. 22, pp 882 - 907.

GOODMAN, L.A. and W.H. KRUSKAL (1954). Measures of association for cross classifications. Journal of the American Statistical Association, vol. 49, pp 732 - 764.

GOODMAN, L.A. and W.H. KRUSKAL (1959). Measures of association for cross classifications II. Further discussions and references. Journal of the American Statistical Association, vol. 54, pp 123 - 163.

GOWER, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, vol. 53, pp 325 - 338.

GOWER, J.C. (1967a). A comparison of some methods of cluster analysis. Biometrics, vol. 23, no. 4, pp 623 - 637.

GOWER, J.C. (1967b). Multivariate analysis and multidimensional geometry. The Statistician, vol. 17, no. 1, pp 14 - 28.

GOWER, J.C. (1970). A note on Burnaby's character weighted similarity coefficient. Journal of the International Association for Mathematical Geology, vol. 2, pp 39 - 45.

GOWER, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, vol. 27, pp 857 - 874.

GOWER, J.C. (1972). Discussion of paper by R. Sibson. Journal of the Royal Statistical Society, vol. B34, pp 340 - 343.

GOWER, J.C. (1975). Generalised Procrustes analysis. Psychometrika, vol. 40, no. 1, pp 33 - 51.

GOWER, J.C. and C.F. BANFIELD (1975). Goodness of fit criteria for hierarchical classification and their empirical distributions. In Corsten and Postelnicu (eds) (1975), pp 347 - 361.

GRASON, J. (1969). Fundamental description of a floor plan design program. Proceedings of EDRA 1, pp 175 - 180.

GRASON, J. (1970). A dual linear graph representation for space-filling location problems of the floor plan type. In Moore, G.T. (ed) (1970), pp 170 - 178.

GREEN, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. Psychometrika, vol. 17, pp 429 - 440.

GREEN, P.E. and F.J. CARMONE (1969). Multidimensional scaling - an introduction and empirical classification of unfolding techniques. Journal of Marketing Research, vol. 6, no. 4.

GREEN, P.E. and V.R. RAO (1972). Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. Holt, Rinehart and Winston, NY.

GREGORY, R.T. and D.L. KARNEY (1969). A Collection of Matrices for Testing Computational Algorithms. Wiley, NY.

GREGORY, S.A. (1967). The Design Method. Butterworth, London.

GRUVAEUS, G.T. (1970). A general approach to procrustes pattern rotation. Psychometrika, vol. 35,

pp 493 - 505.

GUTTMAN, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, vol. 33, pp 469 - 506.

HALL, A.V. (1969). Avoiding informational distortions in automatic grouping programs. Systematic Zoology, vol. 18, no. 3, pp 318 - 329.

HARARY, F. and J. ROCKEY (1976). A city is not a semi-lattice either. Environment and Planning A, vol. 8, pp 375 - 384.

HARTIGAN, J.A. (1967). Representation of similarity matrices by trees. Journal of the American Statistical Association, vol. 62, pp 1140 - 1158.

HARTIGAN, J.A. (1977). Distribution problems in clustering. In Van Ryzin (ed) (1977), pp 45 - 71.

HARTIGAN, J.A. (1978). Asymptotic distributions for clustering criteria. Annals of Statistics, vol. 6, pp 117 - 131.

HENRION, M. (1978). Automatic space-planning: a postmortem? Institute of Physical Planning, Research

Report 72, Carnegie Mellon University.

HILLIER, F.S. (1963). Quantitative tools for plant layout analysis. Journal of Industrial Engineering, vol. 14, pp 33 - 40.

HILLIER, F.S. and M.M. CONNORS (1966). Quadratic assignment problem algorithms and the location of indivisable facilities. Management Science, vol. 13, no. 1, pp 42 - 57.

HILLIER, W., J. MUSGROVE and P. O'SULLIVAN (1972). Knowledge and design. In Mitchell, W.J. (ed) (1972), pp 29.3.1 - 29.3.14.

HOLGERSSON, M. (1978). The limited value of cophenetic correlation as a clustering criterion. Pattern Recognition, vol. 10, no. 4, pp 287 - 295.

JACKSON, B. (1977). Evolution of a spatial allocation system. ACM SIGDA Newsletter, vol. 7, no. 2, pp 12 - 14.

JARDINE, C.J., N. JARDINE and R. SIBSON (1967). The structure and construction of taxonomic hierarchies. Mathematical Biosciences, vol. 1, pp 173 - 179.

JARDINE, N. and R. SIBSON (1968). The construction of hierarchic and nonhierarchic classifications. Computer Journal, vol. 11, pp 177 - 184.

JARDINE, N. and R. SIBSON (1971). Numerical Taxonomy. Wiley, London.

JOHNSON, L.A.S. (1970). Rainbow's end: the quest for an optimal taxonomy. Systematic Zoology, vol. 19, pp 203 - 239.

JOHNSON, S.C. (1967). Hierarchical clustering schemes. Psychometrika, vol. 32, no. 3, pp 241 - 254.

JOHNSON, T.E. (1970). IMAGE: An interactive graphics based computer system for multi-constrained spatial synthesis. MIT Department of Architecture Report.

JOLICOEUR, P. (1963). The multivariate generalisation of the allometry equation. Biometrics, vol. 19, pp 497 - 499.

JONES, J.C. (1970). Design Methods - Seeds of Human Futures. Wiley Interscience, London.

JONES, J.C. (1977). How my thoughts about design methods have changed over the years. Design Methods and Theories, vol. 11, no. 1, pp 48 - 62.

JUEL, H. and R.F. LOVE (1976). An efficient computational procedure for solving the multi-facility rectilinear facilities location problem. Operational Research Quarterly, vol. 27, no. 3(ii), pp 697 - 703.

KING, B.F. (1966). Market and industry factors in stock price behaviour. Journal of Business, vol. 39, pp 139 - 190.

KING, B.F. (1967). Step-wise clustering procedures. Journal of the American Statistical Association, vol. 62, pp 86 - 101.

KREJCIRIK, M. (1969). Computer aided plant layout. Computer Aided Design, vol. 2, p 7.

KRISHNAIAH, P.R. (ed) (1969). Multivariate Analysis II. Academic Press, NY.

KRUSKAL, J.B. (1964a). Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. Psychometrika, vol. 29, pp 1 - 27.

KRUSKAL, J.B. (1964b). Nonmetric multidimensional scaling: a numerical method. Psychometrika, vol. 29, pp 115 - 129.

KRUSKAL, J.B. (1971). Comments on a nonlinear mapping for data structure analysis. IEEE Transactions on Computers, vol. c-20, p 1614.

KRUSKAL, J.B. and J.D. CARROLL (1969). Geometrical models and badness of fit functions. In Krishnaiah (ed) (1969), pp 639 - 671.

KUIPER, F.K. and L. FISHER (1975). A Monte Carlo comparison of six clustering procedures. Biometrics, vol. 31, pp 777 - 783.

LANCE, G.N. and W.T. WILLIAMS (1967a). A general theory of classificatory sorting strategies I. Hierarchic systems. Computer Journal, vol. 9, pp 373 - 380.

LANCE, G.N. and W.T. WILLIAMS (1967b). A general theory of classificatory sorting strategies II. Clustering systems. Computer Journal, vol. 10, pp 271 - 276.

LEE, R.C. and J.M. MOORE (1967). CORELAP - COmputerised RElationship LAyout Planning. Journal of Industrial Engineering, vol. 18, p 195.

LEE, R.C.T., J.R. SLAGLE and H. BLUM (1977). A triangulation method for the sequential mapping of points from N-space to two-space. IEEE Transactions on Computers, vol. c-26, no. 3, pp 288 - 292.

LEVIN, P.H. (1964). Use of graphs to decide the optimum layout of buildings. Architects Journal, 7 October 1964, pp 809 - 815.

LEW, I.P. and P.H. BROWN (1970). Evaluation and modification of CRAFT for an architectural methodology. In Moore, G.T. (1970), pp 155 - 161.

LIGGETT, R.S. (1972). Floor plan layout by implicit enumeration. Proceedings of EDRA 3, pp 23.4.1 - 23.4.12.

LIGGETT, R.S. (1980). A partitioning approach to large floor plan layout problems. Proceedings of CAD 80, pp 705 - 714.

LIGGETT, R.S. and W.J. MITCHELL (1981). Optimal space planning in practice. Computer Aided Design, vol. 13, no. 5, pp 277 - 288.

LINGOES, J.C. (1973). The Guttman-Lingoes Nonmetric Program Series. Mathesis Press, Ann Arbor, Michigan.

LINGOES, J.C. and E.E. ROSKAM (1973). A mathematical and empirical analysis of two multidimensional scaling algorithms. Psychometrika Monograph Supplement no. 19.

LINGOES, J.C. and P.H. SCHONEMANN (1974). Alternative measures of fit for the Schonemann - Carroll matrix fitting algorithm. Psychometrika, vol. 39, pp 423 - 427.

LONGLEY, J.M. (1967). An appraisal of least-squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, vol. 62, pp 819 - 829.

LOOMIS, H.H. (1977). Site selection and placement techniques. ACM SIGDA Newsletter, vol. 7, no. 2, pp 28 - 36.

MALCOLM, M.A. (1972). Algorithms to reveal properties of floating point arithmetic. Communications of the ACM, vol. 15, no. 11, pp 949 - 951.

MARCH, L.J. (1976). The logic of design and the question of value. In March, L.J. (ed) (1976), pp 1 - 40.

MARCH, L.J. (ed) (1976). The Architecture of Form. Cambridge University Press, Cambridge.

MARCH, L.J. and R. MATELA (1974). The animals of architecture: some census results on N-omino populations for N = 6, 7, 8. Environment and Planning B, vol. 1, pp 193 - 216.

MARKUS, T.A. and G.M. AYLWARD (1980). An Appraisal of the Genetics Building, Edinburgh. Research Report, Department of Architecture and Building Science, University of Strathclyde.

MAROY, J.P. and J.P. PENEAU (1973). Multivariate statistical analysis in architectural design. DRS Conference Preprints, pp 2.21.1 - 2.21.6.

MARRIOT, F.H.C. (1971). Practical problems in a method of cluster analysis. Biometrics, vol. 27, pp 501 - 514.

MASSER, I. and P.J.B. BROWN (1975). Hierarchical aggregation procedures for interaction data. Environment and Planning A, vol. 7, pp 509 - 523.

MATELA, R. and E. O'HARE (1976). Graph theoretic aspects of polyominoes and related spatial structures. Environment and Planning B, vol. 3, no. 1, pp 79 - 110.

MAXWELL, E.A. (1946). Methods of Plane Projective Geometry Based on the Use of General Homogenous Co-ordinates. Cambridge University Press, Cambridge.

MAXWELL, E.A. (1951). General Homogenous Co-ordinates in Space of Three Dimensions. Cambridge University Press, Cambridge.

MILLER, W.R., V. KHACHOONI and J. OLSTEN (1969). Matrix method for grouping an inter-related set of elements. Proceedings of EDRA 1, pp 304 - 317.

MILNE, M.A. (1971). CLUSTR: A program for structuring design problems. Proceedings of Design Automation Workshop, pp 242 - 249.

MITCHELL, W.J. (1970a). Notes on approaches to computer aided space planning. Proceedings of Kentucky Workshop on Computer Applications to Environmental Design, pp 82 - 88.

MITCHELL, W.J. (1970b). A computer aided approach to complex building layout problems. Proceedings of EDRA 2, pp 391 - 397.

MITCHELL, W.J. (ed) (1972). Environmental Design: Research and Practice. University of California at Los Angeles, LA.

MITCHELL, W.J. (1975a). Techniques of automated design in architecture: a survey and evaluation. Computers and Urban Society, vol. 1, pp 49 - 76.

MITCHELL, W.J. (1975b). The theoretical foundation of computer aided architectural design. Environment and Planning B, vol. 2, pp 127 - 150.

MITCHELL, W.J. and R.L. DILLON (1972). A polyomino assembly procedure for architectural floor planning. Proceedings of EDRA 3, pp 23.5.1 - 23.5.12.

MITCHELL, W.J., J.P. STEADMAN and R.S. LIGGETT (1976). Synthesis and optimisation of small rectangular floor plans. Environment and Planning B, vol. 3, no. 1, pp 37 - 70.

MOJENA, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. Computer Journal, vol. 20, pp 359 - 363.

MOORE, G.T. (ed) (1970). Emerging Methods in Environmental Design and Planning. MIT Press, Cambridge, Mass.

MOORE, J.M. (1974). Computer aided facilities design: an international survey. International Journal of Production Research, vol. 12, no. 1, pp 21 - 44.

MOSELY, D.L. (1963). A rational design theory for planning buildings based on the analysis and solution of circulation problems. Architects Journal, 11 September 1963, pp 525 - 537.

MOSIER, C.I. (1939). Determining a simple structure when loadings for certain tests are known. Psychometrika, vol. 4, pp 149 - 162.

MOSS, W.W. (1968). Experiments with various techniques of numerical taxonomy. Systematic Zoology, vol. 17, pp 31 - 47.

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

MCRAE, D.J. (1971). MICKA: a FORTRAN IV iterative k-means cluster analysis program. Behavioral Science, vol. 16, pp 423 - 424.

NUGENT, C.E., T.E. VOLLMANN and J. RUML (1968). An experimental comparison of techniques for the assignment of facilities to locations. Operations Research, vol. 16, no. 1, pp 150 - 173.

ORLOCI, L. (1966). Geometric methods in Ecology I. the theory and application of some ordination methods. Journal of Ecology, vol. 54, no. 1, pp 193 - 215.

ORLOCI, L. (1967). Data centering: a review and evaluation with reference to component analysis. Systematic Zoology, vol. 16, pp 208 - 212.

PEARSON, K. (1926). On the coefficient of racial likeness. Biometrika, vol. 18, pp 105 - 117.

PEARSON, W.H. (1966). Estimation of a correlation measure from an uncertainty measure. Psychometrika, vol. 31, no. 3, pp 421 - 433.

PENNINGTON, R.H. (1970). Introductory Computer Methods and Numerical Analysis. Macmillan, NY.

PEREIRA, L.M. N. PORTAS, L.F. MONTEIRO and F. PEREIRA (1973). Interactive dimensional layout schemes from adjacency graphs. Preprints of DRS Conference, pp 2.19.1 - 2.19.5.

PHILLIPS, R.J. (1969). Computerised approaches to circulation. Building, 18 April 1969, pp 117 - 122.

POPPER, K.R. (1963). Conjectures and Refutations. Routledge and Kegan Paul, London.

POPPER, K.R. (1968). The Logic of Scientific Discovery. Hutchison, London.

PORTLOCK, P.C. and B. WHITEHEAD (1971). A program for practical layout planning. Building Science, vol. 6, pp 213 - 220.

PORTLOCK, P.C. and B. WHITEHEAD (1974). Three dimensional layout planning. Building Science, vol. 9,

pp 45 - 53.

POYNER, B. (1966). Activity Data Method. Research and Development Bulletin. HMSO, London.

PRENTICE, I.C. (1977). Nonmetric ordination methods in Ecology. Journal of Ecology, vol. 65, pp 85 - 94.

PRITCHARD, N.M. and A.J.B. ANDERSON (1971). Observations on the use of cluster analysis in botany with an ecological example. Journal of Ecology, vol. 59, pp 727 - 747.

RAO, C.R. (1965). The use and interpretation of principal component analysis in applied research. Sankhya, pp 329 - 358.

RITTEL, H. (1973). The state of the art in design methods. Design Research and Methods, vol. 7, no. 2, pp 143 - 147.

RITTEL, H. and M. WEBBER (1973). Dilemmas in a general theory of planning. Policy Sciences, vol. 4, pp 155 - 163.

ROHLF, F.J. (1970). Adaptive hierarchic clustering schemes. Systematic Zoology, vol. 19, pp 58 - 82.

ROHLF, F.J. (1972). An empirical comparison of three ordination techniques in numerical taxonomy. Systematic Zoology, vol. 21, pp 271 - 280.

ROHLF, F.J. (1973). Hierarchical clustering using the minimum spanning tree. Computer Journal, vol. 16, pp 93 - 95.

ROSSI, B. (1970). A Survey of Classification Techniques. Working Paper, Department of Architecture, University of Bristol.

ROY, G.G. (1979). Multidimensional analysis for spatial organisation. Building and Environment, vol. 14, pp 241 - 246.

SAMMON, J.W. (1969). A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, vol. c-18, pp 401 - 409.

SCHONEMANN, P.H. (1966). A generalised solution of the orthogonal procrustes problem. Psychometrika, vol. 31, pp 1 - 10.

SCHONEMANN, P.H. (1968). On two-sided orthogonal procrustes problems. Psychometrika, vol. 33, pp 19 - 33.

SCHONEMANN, P.H. (1970). On metric multidimensional unfolding. Psychometrika, vol. 35, pp 349 - 366.

SCHONEMANN, P.H. and R.M. CARROLL (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, vol. 35, pp 245 - 255.

SCOTT, A.J. (1971). An Introduction to Spatial Allocation Analysis. Association of American Geographers, Resource Paper 9. Washington, DC.

SCOTT, A.J. and M.J. SYMONS (1971). On the Edwards and Cavalli-Sforza method of cluster analysis. Biometrics, vol. 27, pp 217 - 219.

SEPPANEN, J. and J.M. MOORE (1970). Facilities planning with graph theory. Management Science, vol. 17, no. 4, pp B242 - B253.

SHAPIRA, H. and R.S. FREW (1974). A procedure for generating floor plans. Proceedings of Design Automation Conference, Denver, pp 229 - 236.

SHAVIV, E. and D. GALI (1974). A model for space allocation in complex buildings: a computer graphics approach. Build International, vol. 7, pp 493 - 518.

SHAVIV, E., R. HASHIMSHONY and A. WACHMAN (1977). Decomposition of a multicell complex - a problem in physical design. DMG/DRS Journal, vol. 11, no. 2, pp 1 - 8.

SHAVIV, E., R. HASHIMSHONY and A. WACHMAN (1978). A decomposition - recompostion model for multi-cell systems. Building and Environment, vol. 13, pp 109 - 123.

SHEPARD, R.N. (1962a,b). The analysis of proximities: multidimensional scaling with an unknown distance factor. Psychometrika, vol. 27, part I pp 125 - 139, part II pp 219 - 246.

SHEPARD, R.N. (1972). Introduction to volume 1. Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. (eds R.N. Shepard, A.K. Romney and S.B. Nerlove), Seminar Press, NY. pp 1 - 20

SIBSON, R. (1970). A model for taxonomy II. Mathematical Biosciences, vol. 6, pp 405 - 430.

SIBSON, R. (1971). Some observations on a paper by Lance and Williams. Computer Journal, vol. 14, no. 2, pp 156 - 157.

SIBSON, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. Computer Journal, vol. 16, no. 1, pp 30 -34.

SIBSON, R. (1978). Studies in the robustness of multidimensional scaling: procrustes statistics. Journal of the Royal Statistical Society, vol. B40, pp 234 - 238.

SIMON, H.A. (1969). The Sciences of the Artificial. MIT Press, Cambridge, MA.

SIMON, H.A. (1973). The structure of ill-structured problems. Artificial Intelligence, vol. 4, pp 181 - 201.

SNEATH, P.H.A. (1957). The application of computers to taxonomy. Journal of General Microbiology, vol. 17, pp 201 - 226.

SNEATH, P.H.A. (1966). A comparison of different clustering methods as applied to randomly spaced points. Classification Society Bulletin, vol. 1, no. 2, pp 2 - 18.

SNEATH, P.H.A. and R.R. SOKAL (1973). Numerical Taxonomy. Freeman, NY.

SOKAL, R.R. and C.D. MICHENER (1958). A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, vol. 38, pp 1409 - 1438.

SOKAL, R.R. and C.D. MICHENER (1967). The effects of different numerical techniques on the phenetic classification of bees of the Hoplitis complex (Megachilidae). Proceedings of the Linnean Society of London, vol. 178, pp 59 - 74.

SOKAL, R.R. and F.J. ROHLF (1962). The comparison of dendrograms by objective methods. Taxon, vol. 11, pp 33 - 40.

SOKAL, R.R. and F.J. ROHLF (1970). The intelligent ignoramus, an experiment in numerical taxonomy. Taxon, vol. 19, pp 305 - 319.

SOKAL, R.R. and P.H.A. SNEATH (1963). Principles of Numerical Taxonomy. Freeman, NY.

SORENSEN, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. Biologiske Skrifter, vol. 5, pp 1 - 34.

STEADMAN, J.P. (1970). The automated generation of minimum standard house plans. Proceedings of EDRA 2.

STEADMAN, J.P. (1973). Graph theoretic representation of architectural arrangement. Architectural Research and Teaching, no. 2/3, pp 161 - 172.

STEWART, W. and K. LEE (1972). COMPROSPACE: interactive computer graphics in the real world. Proceedings of EDRA 3.

STRAUSS, J.S., J.J. BARTKO and W.T. CARPENTER JR. (1973). The use of clustering techniques for the classification of psychiatric patients. British Journal of Psychiatry.

TABOR, P. (1976). Analysing communication patterns. Chapter 9 in March (ed) (1976), pp 284 - 351.

TEAGUE, L.C., JR. (1970). Network models of configurations of rectangular parallelepipeds. In Moore G.T. (ed) (1970), pp 162 - 169.

TEN BERGE, J.M.F. (1977). Orthogonal procrustes rotation for two or more matrices. Psychometrika, vol. 42, pp 267 - 276.

TEREKHINA, A.Y. (1973). Methods of multidimensional data scaling and visualisation - a survey. Automation and Remote Control, vol. 34, no. 7, pt. 1, pp 1109 - 1121.

THOMPSON, H.K. and M.A. WOODBURY (1970). Clinical data representation in multidimensional space.

Computers and Biomedical Research, vol. 3, pp 58 - 73.

TUKEY, J.M. (1977). Exploratory Data Analysis. Addison Wesley, Reading, Mass.

ULLRICH, J.R. and H.M. BRAUNSTEIN (1977). The use of multidimensional scaling as a tool in architectural design. Design Methods and Theories, vol. 11, no. 2, pp 121 - 127.

VAN NESS, J.W. (1973). Admissable clustering procedures. Biometrika, vol. 60, pp 422 - 424.

VAN RYZIN, J. (ed) (1977). Classification and Clustering. Academic Press, NY.

VOLLMANN, T.E. and E.S. BUFFA (1966). The facilities layout problem in perspective. Management Science, vol. 12, no. 10, pp B450 - B468.

WAMPLER, R.H. (1970). A report on the accuracy of some widely used least squares computer programs. Journal of the American Statistical Association, vol. 65, pp 549 - 565.

WARD, J.H. (1963). Hierarchical grouping to optimise an objective function. Journal of the American Statistical Association, vol. 58, pp 236 - 244.

WHITE, I. (1972). Comments on a nonlinear mapping for data structure analysis. IEEE Transactions on Computers, vol. c-21, no. 2.

WHITEHEAD, B, and M.Z. ELDARS (1964). An approach to the optimum layout of single storey buildings. Architects Journal, 17 June 1964, pp 1373 - 1380.

WHITEHEAD, B. and M.Z. ELDARS (1965). The planning of single storey layouts. Building Science, pp 127 - 139.

WILKS, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypothesis. Annals of Mathematical Statistics, vol. 9, pp 60 - 62.

WILLIAMS, W.T., H.T. CLIFFORD and G.N. LANCE (1971). Group size dependence: a rationale for choice between numerical classifications. Computer Journal, vcl. 14, no. 2, pp 157 - 162.

WILLIAMS, W.T. and G.N. LANCE (1977). Hierarchical classificatory methods. Chapter 11 in Enslein, Ralston and Wilf (1977), pp 269 - 295.

WILLIAMS, W.T., G.N. LANCE, M.B. DALE and H.T.

CLIFFORD (1971). Controversy concerning the criteria for taxonomic strategies. Computer Journal, vol. 14, no. 2, pp 162 - 165.

WILLOUGHBY, T.M. (1970). Computer aided design of a university campus. Architects Journal, 25 March 1970, pp 753 - 758.

WOLFE, J.H. (1970). Pattern clustering by multivariate mixture analysis,. Multivariate Behavioral Research, vol. 5, pp 329 - 350.

WRIGHT, W.E. (1973). A formalisation of cluster analysis. Pattern Recognition, vol. 5, pp 273 - 282.

YOUNG, T.Y. and T.W. CALVERT (1974). Classification Estimation and Pattern Recognition. Elsevier, NY.

YOUNGS, E.A. and E.M. CRAMER (1971). Some results relevant to the choice of sum and sum of product algorithms. Technometrics, vol. 13, pp 657 - 665.