A Survey of the Literature of Cluster Analysis

Julie Scoltock

Sub-Department of Industrial Management, University of Newcastle upon Tyne,
Stephenson Building, Newcastle upon Tyne NE1 7RU, UK

The purpose of this paper is to provide a survey of the published literature devoted to the subject of cluster analysis. It may be used as a guide for finding further articles, in any field of study, relatively quickly.

INTRODUCTION

Classification and clustering has become an increasingly popular method of multivariate analysis over the past two decades, and with it has come a vast amount of published material. Since there is no journal devoted exclusively to cluster analysis as a general topic and since it has been used in many fields of study, the novice user is faced with the daunting prospect of searching through a multitude of journals for appropriate references. In order to organize this diverse and voluminous material the following points will be considered: the terminology associated with cluster analysis; the fields of study in which the techniques have been practised; the published books; the journals containing the most significant papers; the broad nature of the published papers and a list of the computer packages available for performing a cluster analysis.

An exhaustive list of books is given but a comprehensive list of papers is not included, mainly because these are provided by the books cited and by Cormack's 1971 review. Finally, a few thoughts on the state of the literature of cluster analysis are given.

TERMINOLOGY

In his review Cormack states that 'terminology is used somewhat haphazardly in the literature', and Anderberg[1] attributes the diversity to 'a mixture of professional jealousy, relative isolation among the fields and genuine differences of viewpoint'. The major force at work is the variety of study disciplines in which cluster analysis has been used.

Terminology differs from one field to another. In biology, a significant field of study for the use of cluster analysis, the term 'numerical taxonomy' is frequently used as a substitute for cluster analysis. In pattern recognition the terms are generally 'clustering' or 'classification', but in cybernetics the term 'unsupervised learning' is often found. Geographers use the term 'regionalization' and anthropologists sometimes use cluster analysis to solve a problem they call 'seriation'. Other terms which are frequently used in most fields are 'classification', 'grouping' and 'clumping', 'typology' and 'Q-analysis'.

Further differences in terminology occur throughout the formulation of problems and the description of algorithms, not so much as a result of the use of different words but usually as a result of the use of one word to mean different things. For example, classification is used by some authors to describe techniques for assigning individuals to groups having *a priori* labels and by others to describe the allocation of individuals to initially undefined groups. Fortunately, most authors seem to be acutely aware of this problem and state explicitly their intended meaning of such terms.

FIELDS OF STUDY

It has already been suggested that cluster analysis techniques have been applied to data from numerous and diverse fields of study. For the purpose of this survey they are presented in five major groups:

Biology and zoology

Clustering techniques were first developed in this field. They are invariably used to group animals and plants, etc., and hence develop taxonomies (naming schemes). So much significant work has been done in this area that the terms biological/numerical taxonomy have come to be associated with this application of cluster analysis.

Medicine and psychiatry

The principal application in these fields has been the classification of disease, both mental and physical. Although a characteristic difficulty of classification of mental illness is the subtle and variable character of the symptoms, numerical techniques have in fact gained more acceptance in this area (psychiatry) than in medical diagnosis. In both areas cluster analysis has been slow to develop because medical and psychiatric data is difficult to assimilate into the standard data matrices used in cluster analysis.

Sociology and criminology, anthropology and archaeology

Such entities as training methods, organizations, criminals and crimes, cultures, racial mixture in human populations and objects found in excavation sites have all been subject to clustering techniques.

Geology, geography and remote sensing

Rock samples and sediments, cities and land-use patterns have been studied using cluster analysis.

CCC–0010–4620/82/0025–0130 $02.50

Information retrieval, pattern recognition, market research and economics

Objects of analysis have been pictures and scenes, documents, industries, consumers, products, markets and productivity ratios. Market research has been the source of perhaps the most innovative applications of cluster analysis over the past few years—Paul Green has done much work in this field. Information retrieval and classification have also been the subject of clustering attempts in recent years.

In addition to these major groups the techniques of cluster analysis have been used in a few other areas. In literature samples of prose have been analysed for rhythmic differences, in electrical engineering circuit designs have been analysed and in production engineering machines have been grouped for efficient production.

PUBLISHED BOOKS

The number of books devoted to this subject is limited—the list given at the end of this paper is, to the best of the author's knowledge, exhaustive. The earlier books such as Sokal and Sneath[10] and Jardine and Sibson,[9] tend to be heavily slanted toward the field of biology. However, the more recently published books do cater for the general user. Everitt's book[6] is compendious, giving any novice a good introduction to the main ideas of cluster analysis. He briefly introduces the subject, reviews the various techniques, discusses the problems associated with them, investigates some of the methods using artificial data (generated to have a particular structure) and finally provides a guide to the practical use of clustering techniques. Duran and Odell[5] also provide a brief exposition of cluster analysis. Theirs, however, is more mathematically detailed than Everitt's with little reference to practical problems or detailed examples. Tryon and Bailey[11] devote their book on cluster analysis to a comprehensive description of their factor analysis/cluster analysis package called BC TRY. Three applications of cluster analysis are used repeatedly in the book and all of the data are from Tryon's field of interest—psychology. Hartigan[8] provides a treatment of modern clustering theory from the statistical point of view. His book contains detailed discussions of a variety of algorithms and their application to real data sets ranging from medical and biological data to political data.

Anderberg's book,[1] entitled *Cluster Analysis for Applications* deserves a special mention. It is an excellent introductory text providing the novice user with an adequate appreciation of the topic. Anderberg says that his own education in the subject was through the 'laborious path of extensive reading in a wide variety of books and journals', and consequently he had a good understanding of the problems facing the newcomer to cluster analysis when he wrote this book. He gives a broad view of the subject and discusses in detail the many problems associated with the concepts of cluster analysis. In addition to notes on various methods he provides ideas on the interpretation of results and strategies for using cluster analysis—areas which have been neglected in the past. He also gives an excellent treatment of distance measures. His contribution to the area of applied cluster analysis is certainly very valuable and although he does not refer to practical examples

Table 1. Journals containing significant publications on cluster analysis

Journal	*Coverage*
Applied Statistics: Journal of the Royal Statistical Society, Series C	Emphasis on statistical methodology with illustrative applications
Biometrics	Often theoretical papers with a tendency to relate to the life sciences
Biometrika	Almost exclusively theoretical papers
Computer Journal	Various papers with program listings. A lot of Lance and Williams' work is published in this journal
Educational and Psychological Measurement	This journal contains most of McQuitty's work on cluster analysis
IEEE Transactions on Computers	Cluster analysis is treated as a subfield of pattern recognition
Information Storage and Retrieval	Papers of particular interest to librarians and information scientists
International Journal of Production Research	Cluster analysis is used in the fields of group technology and production flow analysis
Journal of the American Statistical Association	Papers concentrate on various theoretical aspects of cluster analysis
Journal of Business Research	The application of cluster analysis to a variety of business problems
Journal of Ecology	Devoted to plants, animals and other life-forms
Journal of the Marketing Research Society	Applications in marketing: selection of test markets, consumer behaviour, market segmentation and product preference. An area of expansion
Management Science	Marketing and business applications. Paul Green's work is often published in this journal
Multivariate Behavioural Research	Field of psychology
Nature	Papers are usually brief (since this journal is published weekly) and are often comments on methods. There is a tendency for them to be pertinent to biology
Omega	Work of a similar vein to that found in the *International Journal of Production Research*
Operational Research Quarterly	Marketing problems
Pattern Recognition	Clustering techniques strongly related to the ideas of pattern recognition. Often theoretical papers concerned with mode-seeking techniques
Psychometrika	Theoretical papers in the field of psychology
Systematic Zoology	A considerable amount of material relating to cluster analysis may be found in this journal, but it is mainly only applicable in the field of zoology

anyone wishing to use cluster analysis techniques on real data should not fail to read this book.

The most recently published book[12] is a collection of papers given at the Advanced Seminar on Classification and Clustering held in Wisconsin in 1976. It covers a lot of ground, ranging from a general discussion on the background and current directions in the field to aspects of the theory and application of the techniques.

Finally, two early books which have not been mentioned are by Fisher[7] and Cole,[4] and a third book by Anderson *et al.*[2] is an extensive bibliography of multivariate statistical analysis. Fisher develops a theory of clustering and aggregation and applies it to basic types of economic problems. Cole presents a collection of papers given at a colloquium in Numerical Taxonomy which describe developments and applications of such methods. Discussion resulting from the papers is also included.

JOURNALS

The major drawback of the textbooks in this field is that, due to the delay between conception and publication, their contents quickly become out of date. Consequently, the most recent research reports are to be found in journals. Since there is no single journal devoted to cluster analysis it becomes necessary to search a number of publications for relevant papers. Table 1 lists those journals in which significant publications can be found.

At least one reference per journal has been cited at the end of this paper. Many others may be found in the bibliographies of the books on cluster analysis and Cormack's 1971 review.

A BRIEF REVIEW OF PUBLISHED PAPERS

The total population of papers published on the broad topic of cluster analysis falls into one of two categories: the theoretical and the applied. As far as possible indications were given in Table 1 as to the nature of the papers to be found in each journal.

The theoretical papers either discuss the mathematical/statistical aspects of existing techniques or propose new clustering algorithms. For example, Bolshev's paper entitled 'Cluster analysis' is a survey of the papers related to the probabilistic approach to cluster analysis and contains a detailed section on probabilistic interpretation; Sibson's paper discusses the mathematical details of the single linkage method and presents an optimally efficient algorithm (and a corresponding program) for its execution; and Baker and Hubert discuss a statistical concept of power for hierarchical clustering schemes. Occasionally these papers are accompanied by examples, usually on single artificial data-sets. For example, Edwards and Cavalli-Sforza present a (hierarchical) method of cluster analysis and give examples of its application to bacteriological data and Rao's race mixture data.

The applied papers obviously describe the application of one or more techniques to a particular problem. For example, Wishart and Leach apply three different methods of cluster analysis to a data-set consisting of percentage occurrences of 5-syllable sequences throughout 33 passages of Platonic text, in an attempt to analyse Platonic prose rhythm; and Parks applies a hierarchical clustering technique to Purdy's data on the constituent particle composition of recent Bahamian bottom sediment samples. However, sufficient details of the analysis are not always given and motivation for the use of the techniques in the first instance is often omitted.

These and other references are cited at the end.

COMPUTER PROGRAMS

Numerous computer programs and sets of programs have been developed as a result of research in the field. By far the most readily available and complete package is CLUSTAN, developed by Wishart in 1969 and now into its third edition. It is a suite of programs written in FORTRAN and contains a comprehensive choice of techniques. Further details may be obtained from, Computer Centre, University College, London, 19 Gordon Street, London, WC1H 0AH and a copy of the user manual is available from, Program Library Unit, Edinburgh University, 18 Buccleuch Place, Edinburgh EH8 9LN.

Another readily available package is MIDAS, but unlike CLUSTAN, it is a general statistical package which contains only a single clustering subroutine. The output is less detailed but it offers the most popular techniques and the option of grouping the cases or the variables. Further details may be obtained from, Statistical Research Laboratory, University of Michigan. Everitt[6] (p. 100) gives a list of eight other programs (all written in one or other version of FORTRAN) and details of their availability. An abridged version of the user's manual for the BC TRY system of programs, developed by Tryon and Bailey, appears in their book[11] courtesy of Tryon-Bailey Associates, Inc.

The remaining computer programs for cluster analysis are found as listings in the appendices of books and articles. Anderberg's book[1] contains program listings for scale conversions, association measures, hierarchical and non-hierarchical clustering techniques and aids to interpreting the clustering results. Hartigan[8] lists all the programs he describes and a number of the references cited in this paper include listings. However, the latter tend to be specific to the subject matter of the particular paper.

CONCLUSIONS

The literature of cluster analysis is scattered throughout many journals in many fields of study. The aim here was not to produce a comprehensive list of references, but to offer a guide to further reading in the subject and a selection of papers covering as many areas as possible. Numerous other references may be found in the papers cited at the end.

At present there is no journal devoted exclusively to the subject of cluster analysis. It is debatable whether or not there is a real need for one. One might argue that any individual paper will fit naturally into an existing journal;

however, such a journal would clearly be a major benefit to research workers by dramatically reducing the amount of time spent searching for material. Another advantage which might be gained from a 'cluster analysis' journal is a possible increase in the amount of detail given in articles concerning the application of clustering techniques to real data sets. At present authors tend to devote more effort to the description of the techniques used than to the data collection and the motivation, the detailed results and their interpretation. Details of work carried out on 'live' data are lacking at present but might be more forthcoming in a specialist journal with, presumably, a readership particularly interested in them.

Much can be learned by studying the articles published in the field of cluster analysis. The papers concerning the application of the techniques in particular, should give anyone contemplating using cluster analysis a good idea as to whether or not it will be a suitable and useful method of data analysis; and should help the theorists understand the practical problems involved in using these techniques.

REFERENCES

Books

1. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, London (1973).
2. T. W. Anderson, S. Das Gupta and G. P. H. Styan, *A Bibliography of Multivariate Statistical Analysis*, Oliver and Boyd, Edinburgh (1972).
3. H. T. Clifford and W. Stephenson, *An Introduction to Numerical Classification*, Academic Press, London (1975).
4. A. J. Cole, *Numerical Taxonomy*, Academic Press, London (1969).
5. B. S. Duran and P. L. Odell, *Cluster Analysis. A Survey*, Springer-Verlag, Berlin (1974).
6. B. Everitt, *Cluster Analysis*, Heinemann Education Books, London (1974).
7. W. D. Fisher, *Clustering and Aggregation in Economics*, John Hopkins, Baltimore (1968).
8. J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York (1975).
9. N. Jardine and R. Sibson, *Mathematical Taxonomy*, John Wiley & Sons, London (1971).
10. R. R. Sokal and P. H. A. Sneath, *Numerical Taxonomy*, Freeman, San Francisco (1973).
11. R. C. Tryon and D. E. Bailey, *Cluster Analysis*, McGraw-Hill, New York (1970). (This book is now out of print.)
12. J. Van Ryzin, *Classification and Clustering*, Academic Press, New York (1977).

BIBLIOGRAPHY

Papers

The following references, listed alphabetically, include at least one from each of the journals enumerated in this paper and several devoted to both the theory and application of cluster analysis.

F. B. Baker, Information retrieval based on latent class analysis. *Journal of the Association for Computing Machinery* **9**, 512–521 (1962).

F. B. Baker and L. J. Hubert, Measuring the power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association* **70** (No. 349), 31–38 (1975).

G. H. Ball and D. J. Hall, A clustering technique for summarising multivariate data. *Behaviour Science* **12**, 153–155 (1967).

D. N. Baron and P. M. Fraser, Medical applications of taxonomic methods. *British Medical Bulletin* **24**, 236–240 (1968).

L. N. Bolshev, Cluster Analysis. *Bulletin of I.S.I.* **43** (Book I), 411–425 (1969).

D. M. Boulton and C. S. Wallace, A program for numerical classification. *The Computer Journal* **13**, 63–69 (1970).

J. Bryant, On the clustering of multidimensional pictorial data. *Pattern Recognition* **11**, 115–125 (1979).

J. W. Carmichael, J. A. George and R. S. Julius, Finding natural clusters. *Systematic Zoology* **17**, 144–150 (1968).

A. S. Carrie, Numerical taxonomy applied to group technology and plant layout. *International Journal of Production Research* **11** (No. 4), 399–416 (1973).

H. J. Chen, D. M. Dunn and J. M. Landwehr, Grouping companies based on their Operating Environment. *Proceedings of the American Statistical Association*, 278–283 (August 1975).

J. D. Claxton, The use of RMCA to distinguish artifacts from natural groupings. *Market Research Society Journal* **17** (No. 3), 198–200 (1975).

P. Constantinescu, The classification of a set of elements with respect to a set of properties. *The Computer Journal* **8**, 352–357 (1966).

R. M. Cormack, A review of Classification. *Royal Statistical Society Journal, Series A* **134** (No. 3), 321–367 (1971).

P. Doyle and Z. B. Gidengil, Defining International Market Opportunities via Wishart's mode analysis. *Operational Research Quarterly* **29** (No. 2), 147–157 (1978).

A. W. F. Edwards and L. L. Cavalli-Sforza, A Method for cluster analysis. *Biometrics* **21**, 362–375 (1965).

B. Everitt, Unresolved problems in Cluster Analysis. *Biometrics* **35**, 169–181 (1979).

J. L. Fleiss and J. Zubin, On the methods and theory of clustering. *Multivariate Behavioral Research* **4**, 235–250 (1969).

H. P. Friedman and J. Rubin (1967), On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62**, 1159–1178 (1967).

D. W. Goodall, Hypothesis testing in classification. *Nature* **211**, 329–330 (1966).

J. C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338 (1966).

A comparison of some methods of cluster analysis. *Biometrics* **23**, 623–628 (1967).

P. E. Green and F. J. Carmone, Segment congruence analysis: a method for analysing association among alternative bases for market segmentation. *Journal of Consumer Research* **3** (No. 4), 217–222 (1977).

P. E. Green, R. E. Frank and P. J. Robinson, Cluster analysis in test market selection. *Management Science* **13**, 387–400 (1967).

P. Harris, The effect of clustering on costs and sampling errors of random samples. *Journal of the Market Research Society* **19** (No. 3), 112–122 (1977).

P. J. Harrison, A method of cluster analysis and some applications. *Applied Statistics* **17**, 226–236 (1968).

F. R. Hodson, P. H. A. Sneath and J. E. Doran, Some experiments in the numerical analysis of archaeological data. *Biometrika* **53**, 311–324 (1966).

K. Hope, Complete analysis: a method of interpreting multivariate data. *Journal of the Market Research Society* **11**, 267–284 (1969).

S. C. Johnson, Hierarchical Clustering Schemes. *Psychometrika* **32** (No. 3), 241–254 (1967).

K. S. Jones, Some thoughts on classification for retrieval. *Journal of Documentation* **26**, 89–101 (1970).

K. S. Jones and D. M. Jackson, The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval* **5**, 175–201 (1970).

D. G. Kendall, Some problems and methods in statistical archaeology. *World Archaeology* **1** (No. 1), 68–76 (1969).

J. N. Kennedy, A review of some cluster analysis methods. *AIIE Transactions* **6** (No. 3), 216–227 (1974).

B. F. King, Step-wise clustering procedures. *Journal of the American Statistical Association* **62**, 86–101 (1967).

J. R. King, Scheduling and the Problem of computational complexity. *Omega* **7** (No. 3), 233–240 (1979).

Machine-component group formation in group technology. *Omega* **8** (No. 2), 193–199 (1980).

Machine-component grouping in production flow analysis: an approach using a rank order clustering algorithm. *International Journal of Production Research* **18** (No. 2), 213–232 (1980).

J. R. King and A. S. Spachis, Heuristics for flow-shop scheduling. *International Journal of Production Research* **18** (No. 3), 345–357 (1980).

G. N. Lance and W. T. Williams, A general theory of classifactory sorting strategies. I. Hierarchical systems and II. Clustering systems. *Computer Journal* **9** (No. 4), 373–380 (1966), **10** (No. 3), 271–276 (1967) respectively.

J. McAuley, Machine grouping for efficient production. *The Production Engineer* 53–57 (February 1972).

W. T. McCormick, P. J. Schweitzer and T. W. White, Problem decomposition and data reorganisation by a clustering technique. *Operations Research* **20**, 993–1009 (1972).

L. L. McQuitty, Capabilities and improvements of linkage analysis as a clustering method. *Educational and Psychological Measurement* **24**, 441–456 (1964).

F. H. C. Marriott, Practical problems in a method of cluster analysis. *Biometrics* **27**, 501–514 (1971).

E. Mayr, Theory of biological classifications. *Nature* **220**, 545–548 (1968).

D. G. Morrison, Measurement problems in cluster analysis. *Management Science* **13**, 775–780 (1967).

J. I. Naus, An indexed bibliography of clusters, clumps and coincidences. *International Statistical Review* **47**, 47–78 (1979). (A bibliography bringing together an extensive and widely scattered literature on cluster distributions.)

L. Orloci, An agglomerative method for classification of plant communities. *Journal of Ecology* **55**, 193–206 (1967).

J. M. Parks, Cluster analysis applied to multivariate geologic problems. *Journal of Ecology* **74**, 703–715 (1966).

R. A. Peterson, Market structuring by sequential cluster analysis. *Journal of Business Research* **2** (No. 3), 249–264 (1974).

F. J. Rohlf, Adaptive hierarchical clustering schemes. *Systematic Zoology* **19**, 58–82 (1970).

G. Sebestyen and J. Edie, An algorithm for non-parametric pattern recognition. *Institute of Electrical and Electronic Engineers Transactions Computers*, EC-15, 908–915 (1966).

E. Shaffer, Single linkage characteristics of a mode seeking clustering algorithm. *Pattern Recognition* **11**, 65–70 (1979).

R. Sibson, SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal* **16** (No. 1), 30–34 (1973).

J. H. Ward, Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association* **58**, 236–244 (1963).

W. G. Wee, Generalised inverse approach to adaptive multiclass pattern classification. *Institute of Electrical and Electronic Engineers Transactions Computers* **C17**, 1157–1164 (1968).

W. T. Williams and J. M. Lambert, Multivariate methods in plant ecology I and II. *Journal of Ecology* **47**, 83–101 (1959); **48**, 689–710 (1960) respectively.

D. Wishart, An algorithm for hierarchical classifications. *Biometrics* **22**, 165–170 (1969).

D. Wishart and S. V. Leach, A multivariate analysis of Platonic prose rhythm. *Computer studies in Humanities and verbal behavior* **3**, 90–99 (1971).

J. H. Wolfe, Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5**, 329–350 (1970).

Received March 1981