Behavior Research Methods & Instrumentation
1982, Vol. 14 (2), 170-175

Interactive clustering

C. P. WHALEY
Bell-Northern Research, Ottawa K1Y 4H7, Canada

An interactive strategy for applying cluster-analytic techniques in behavioral research is presented. The two-part approach stresses the use of on-line computers for both data collection and analysis. In data collection, an extension of multidimensional unfolding to clustering reduces the number of judgments required of subjects by as much as 50%. During data analysis, the interactive procedures described permit the testing of multiple clustering models from an extensive family. With each selection, the goodness of fit of the model to the data can be tested. In addition to improving efficiency, the interactive strategy promoted here combines the advantages of the original nonmetric clustering procedures (e.g., Johnson, 1967) with those of the latest linear additive models (e.g., Sattath & Tversky, 1977; Shepard & Arabie, 1979).

This paper promotes the use of on-line computers for the collection and analysis of data that are best suited for cluster analysis. As such, it is in keeping with the theme of two papers I have presented to this conference on previous occasions, in which on-line computing procedures were recommended for paired comparison research (Whaley, 1979) and for computer-augmented decision making (Whaley, 1981). The techniques described here are oriented toward the reduction of time and effort in carrying out the research exercise.

DATA COLLECTION

There is an unmatched elegance in putting respondents in front of a terminal for the purposes of data collection. Questionnaires, tachistoscopes, and other such paraphernalia pale by comparison. With the collection of similarity or relatedness data, paper-and-pencil exercises are commonly adequate; but they are usually quite time-consuming with even a moderate set of stimulus materials. One normally expects each respondent to rate all $N(N-1)/2$ unique pairs of stimuli within an experimental session. The end product is a matrix like the one shown in Table 1.

In this case, six graduate students volunteered to generate "relatedness" data for pairs of the 12 nouns shown.[1] The table shows the pooled data. High values indicate high relatedness between the pairs of nouns represented by the corresponding row and column labels. Naturally, as N increases, the number of judgments increases with the square of N.

In the case in which the data are subsequently to be subjected to multidimensional scaling, there are ways of reducing the number of judgments per respondent (e.g., Cliff, Girard, Green, Kehoe, & Doherty, 1977). Unfortunately, thus far there are not comparable techniques for data which are more suited for cluster analysis.

The technique proposed here shows a great deal of promise. The mathematics are by no means original,

Table 1
Relatedness Data for 12 Nouns

1–wife											
2–trout	4										
3–mother	52	6									
4–turtle	4	49	9								
5–tiger	2	23	6	20							
6–husband	56	2	35	1	7						
7–knight	33	1	27	4	14	45					
8–crocodile	0	38	1	42	42	1	3				
9–cook	44	16	52	9	0	35	24	1			
10–shark	0	48	0	38	42	0	4	53	0		
11–partner	47	3	38	3	0	47	28	0	35	5	
12–dog	8	26	13	26	44	15	12	25	6	23	23
	1	2	3	4	5	6	7	8	9	10	11

Table 2
Data Matrix for Multidimensional Unfolding

2–	**										
3–	**	**									
4–	**	**	**								
5–	**	**	**	**							
6–	**	**	**	**	**						
7–	33	1	27	4	14	45					
8–	0	38	1	42	42	1	**				
9–	44	16	52	9	0	35	**	**			
10–	0	48	0	38	42	0	**	**	**		
11–	47	3	38	3	0	47	**	**	**	**	
12–	8	26	13	26	44	15	**	**	**	**	**
	1	2	3	4	5	6	7	8	9	10	11

but the present application appears to be. Furnas (1980) has generalized the case of multidimensional unfolding to cluster analysis. For those unfamiliar with multidimensional unfolding, Table 2 might prove helpful.

With the case of multidimensional unfolding, one typically has a set of objects that have been rated as to the degree to which they possess certain features of interest to the experimenter. Consequently, the rows in the matrix represent objects and the columns indicate the various features. In the course of a multidimensional unfolding analysis, both the objects and their features

0005-7878/82/020170-06$00.85/0

are placed in an m-dimensional space so that similar objects are located close together, similar features (across objects) are located together, and features that are closely associated with specific objects are placed close together. Clearly, the algorithms (and computer programs) for accomplishing this are not trivial, and the likelihood of a good match between representation and original data is typically not as high as with a comparable case involving multidimensional scaling.

In Table 2, one can imagine that the rows indicate objects and the columns, features. The multidimensional unfolding technique must place the objects and features in a space such that the interobject distances, the interfeature distances, and the object-feature distances are congruent with the rectangular matrix of data shown. Naturally, the missing data indicated by asterisks must be filled in with distances that are appropriate, given the constraints of the model. Multidimensional unfolding essentially fills in the gaps while maintaining a monotonic relationship between the values in the rectangle with those of the corresponding fitted distances.

Consider two extensions to this. First, rather than consider objects and their features, imagine the data to consist of the relatedness of one set of objects with another. Referring to Table 1, this means Objects 1-6 with Objects 7-12. Second, use Furnas' (1980) formulation of unfolding for cluster analysis, which includes the necessary mathematics and computer programs.

The general implication is that there is a reduction in the amount of data that must be collected by roughly one half. The other half can be estimated using the unfolding techniques. Since there is a certain amount of error associated with the estimated entries in the matrix, it is advisable to select the set of rows and columns for any given individual (who participates in the experiment) on a random basis. The pooled data should then dilute the consequences of the estimation procedure.

The cited data were collected prior to the development of this method, so all respondents generated data for all pairs. Consequently, for illustrative purposes, we can "throw away" half of these data, apply the techniques described for estimating the empty cells, and compare the results with those that would have been obtained had all the data been used.

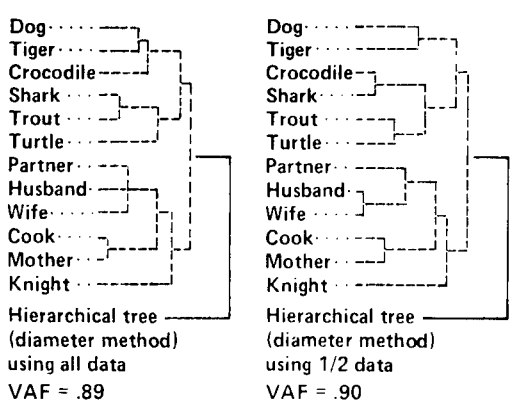

Figure 1. Nonmetric clustering with complete and incomplete data.

Figure 1 shows Johnson (1967) (diameter method) solutions based on all of the original data and on the data for which half of the data have been estimated. The overall trees are quite similar, although there is some disagreement in the placement of four of the animal terms, TROUT, TURTLE, SHARK, and CROCODILE, at the lower levels. The correlation between the sets of distances derived from the two solutions is .94. When Johnson's connectedness method is applied, the results are even better. The correlation is .98. This means that over 95% of the variance in the "true" cluster solution has been captured from only half of the data.

It is worth drawing attention to the variance-accounted-for (VAF) figures. Both solutions account for roughly 90% of the variance. These figures were obtained by assuming that the data are better than ordinal in scale. The distances derivable from the trees were correlated with the original data. Regardless of whether the interval scale assumption is justified, this is a global measure of goodness of fit. It is a fit of the whole tree to the data. It is also reasonable and often preferable to examine the goodness of fit of various levels (or partitions) within the tree. For instance, one might ask how well the two-cluster solution of animal nouns vs. human nouns fits the data.

DATA ANALYSIS

The use of cluster analysis on proximity data generally involves a number of problems. The first is the choice of a clustering method. There are literally hundreds of clustering algorithms, some with available computer programs. The selection of the most appropriate program for any given set of data is not an easy one. Fortunately, there are some fairly thorough comparative studies of some of the more popular methods (e.g., Blashfield, 1976).

The second problem is to some extent associated with the first. The data analyst must decide whether he is interested in only one solution (hopefully, the best one) or whether he would like to consider several. If the second alternative is chosen, the investigator must decide whether the multiple partitions should be nested. The most popular techniques typically produce a hierarchy of nested partitions in which two clusters at one level may merge to form one cluster at a higher level. Frequently, the complete hierarchy has intrinsic meaning beyond that revealed at any one or more individual levels.

Independent of the decisions associated with the number of partitions, one must decide whether or not to allow overlapping clusters. Until recently, it was rarely necessary to make this decision, since most clustering programs always produced nonoverlapping clusters. Shepard and Arabie (1979), however, have

demonstrated that it is sometimes meaningful to consider clusters that overlap to some extent and have proposed a model, called ADCLUS, that produces such solutions and tests them using a linear additive model. Arabie and Carroll (1980b) have extended this method and released a computer program, called MAPCLUS, for public consumption (Arabie & Carroll, 1980a).

There are very few well established significance tests for cluster analysis. Consequently, it is difficult to tell whether one solution is better than another. That is not to say that there are no measures of goodness of fit; it is simply that the practitioner does not know how big a difference must be to be significant.

Cluster analysis is probably best used in behavioral research to convey some conceptual structure in the data. For example, one might want to use people's judgments of the relatedness of members of a set of key words to extract a hierarchical representation that might be used to form the basis of an on-line data base or filing system. Many cluster analysis programs produce one or more hierarchies from the same set of data. Regardless of how well any one of these representations actually fit the data, it is also important that the solution chosen be meaningful. Unless there appears to be some apparent basis for cluster membership (and, consequently, some appropriate label that could be used to describe each cluster), the solution may not be useful, no matter how well it appears to fit in a mathematical sense.

It should be obvious that whatever the proposed method is, it is not going to solve all of the problems. However, it is possible to construct an interactive clustering program with enough flexibility that it is at least possible to explore various alternatives in each problem area. At the same time, there will be adequate opportunities for model testing if that is desired.

The Algorithm

The method adopted here is a hybrid of both new and well established components. It contains the following elements.

Peay/Hubert generalization of nonmetric clustering. Johnson's (1967) paper on hierarchical clustering schemes is perhaps the most influential in introducing cluster analysis to psychologists and other behavioral scientists. Johnson proposed two nonmetric methods of producing a hierarchy of cluster partitions that were monotonically related to the data from which they were generated. Although many synonyms are available, they are generally called the connectedness and the diameter methods. Because they were nonmetric, these methods have wide applicability, since no assumptions must be made about the underlying distribution of the data or whether the data fall on an interval scale.

Hubert (1974) and Peay (1974) have independently shown that the two methods proposed by Johnson (1967) are in fact two extremes of a continuum of such methods. Since there are pros and cons associated with both of the extreme methods, it seems reasonable to examine solutions that are in between.

The Peay/Hubert generalization of the original techniques is based on a graph-theoretical representation of the clustering problem and, consequently, permits an examination of other graph properties in the generation and representation of the data.

Overlapping clusters. Many of the intermediate solutions have levels in the hierarchical representation in which elements are permitted to belong to more than one cluster. Shepard and Arabie (1979) have shown that there is often good reason to expect overlapping clusters and have stressed this in the development of their ADCLUS model. The more recent Arabie and Carroll (1980a, 1980b) realization of the ADCLUS model in the form of MAPCLUS maintains this emphasis.

Model fitting by ADCLUS approach. The MAPCLUS computer program for fitting the ADCLUS model attempts to maximize the goodness of fit between arrangements of objects into clusters and the original data. The model is expressed as follows: $\bar{S} = PWP' + C$, where $\bar{S}$ = n by n symmetric matrix of reconstructed similarities, P = n by m rectangular matrix of binary values, W = m by m diagonal matrix of weights, P′ = m by n transpose of P, and C = n by n matrix with zeros in the diagonal and the fitted additive constant elsewhere.

Table 3 summarizes the ADCLUS model as it is applied to a simple set of data. Four objects are involved in the similarity matrix. A two-cluster solution is proposed with cluster membership in the binary P matrix signified by ones. The ADCLUS model is reexpressed in summation format in the middle of the table. This formula is simplified slightly by letting "q" stand for the products of the "p" terms. For each pair of objects,

Table 3
Example of ADCLUS Model Fitting

Similarity data

$$S_{n \times n} = \begin{bmatrix} - & 6 & 7 & 4 \\ 6 & - & 3 & 5 \\ 7 & 3 & - & 1 \\ 4 & 5 & 1 & - \end{bmatrix}$$

Hypothetical cluster solution

$$P_{n \times m} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

ADCLUS model: $\hat{s}_{ij} = \sum_{k=1}^{m} w_k p_{ik} p_{jk} + c$

$$= \sum_{k=1}^{m} w_k q_{(ij)k} + c$$

(ij)	q_1	q_2	S	
1,2	1	0	6	
1,3	0	0	7	c = .75
1,4	0	0	4	
2,3	0	1	3	
2,4	0	1	5	R^2 = .46
3,4	0	1	1	
w_k =	.08	−.42		

then, there is a binary q-value for each cluster. When we add the values from the S matrix as a final column, we have the classical setup for a multiple-regression analysis. This is the most unique aspect of the ADCLUS model. The goodness-of-fit measure is consequently R^2, and the importance of the individual clusters can be examined by looking at the beta weights.

The present approach embraces this ADCLUS/multiple-regression procedure for fitting any fixed number of clusters to the original data but dispenses with the complex cluster-seeking algorithm employed in MAPCLUS. If it is not already obvious, any clustering method could be used to obtain overlapping or nonoverlapping clusters prior to ADCLUS model fitting. I have made use of the Peay algorithm, which provides a family of hierarchical solutions of a nonmetric nature. Since the solutions are based on the graph-theoretic concept of a clique (as opposed to a "true" cluster), overlap is possible. Some objects may be adequately represented as members of more than one clique. Since it neither forces true partitions nor overlapping sets, the method can be used to test either of these concepts.

The fundamental difference between MAPCLUS and the algorithm employing the Peay method is that the Peay method is based on sound graph-theoretic concepts, whereas MAPCLUS employs a number of iterative approximation techniques that are more difficult to appreciate in a theoretical sense.

Mathematical simplicity. By expressing the clustering problem in graph-theoretic terms, the mathematics are also greatly simplified. There is an elementary quasi-matrix multiplication routine that does most of the work in generating the family of distance matrices from which the clustering solutions between Johnson's (1967) connectedness and diameter methods can be generated.

An on-line interactive implementation is desirable because of the large number of solutions, good and bad, that are possible within the Peay/Hubert family. With an interactive cluster analysis, the practitioner can preset a range of values for several parameters: (1) As with MAPCLUS, you may select the number of clusters you feel should be present in the solution or you may select a range of sizes for consideration. (2) You may specify a reachability value or range of values. Without going into extensive detail, this is a parameter whose value changes as each member of the family of hierarchical solutions emerges between the extreme methods popularized by Johnson (1967) and others. (3) You can control the output to display only those solutions that account for a certain proportion of variance after ADCLUS model fitting.

Naturally, these parameters can be selected in any of a number of combinations. The principal advantage of the interactive approach is that the quality (i.e., meaningfulness or interpretability) of the solutions within the constraints of the parameters selected can be examined and "traded off" against the goodness-of-fit criterion.

With the data from the noun-relatedness experiment, it seemed appropriate to choose a fairly conservative minimum for the VAF criterion. The threshold was set at .81, corresponding to a multiple R of .90 in the ADCLUS/regression model-fitting stage. On the other hand, there did not appear to be any a priori reasons to constrain the range of reachability values; so the full range was permitted. In considering the number of clusters for examination, the program was constrained to display only those solutions with at least two or no more than six (half the number of objects) clusters.

There were four hierarchies generated (i.e., four reachability values) before the algorithm converged on Johnson's (1967) connectedness solution. Within the four hierarchies, nine clique sets met the criteria. When these were displayed on-line, three were identical: the two-cluster solution that separates the animal nouns from the human nouns.

Of the six remaining sets, three had five subsets, two had four subsets, and one had three. With the exception of the two-cluster solution already mentioned, there were overlapping clusters within all of the remaining sets.

It is interesting to observe how the meaningfulness of the cluster sets dissipates as the number of clusters increases (for these data at least).

Figure 2 shows the two-cluster solution. There are no real surprises here. The three-subset solution (Figure 3) was the same as the two-subset solution with

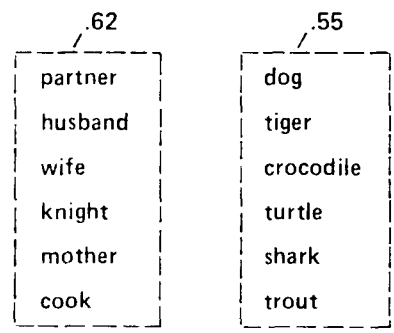

Figure 2. Two-cluster solution.

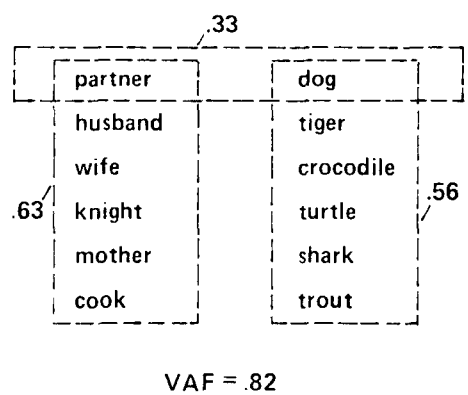

Figure 3. Three-cluster solution.

the exception that there was a clique containing one member from each of the two already discussed, PARTNER and DOG. The pet/owner association is fairly obvious. With one of the four-subset solutions (Figure 4), the same three subsets appeared plus another two-member subset, COOK and TROUT. Again, the association is readily interpretable.

In the other four-subset solution (Figure 5), PARTNER, DOG, and HUSBAND form a separate clique. The "man's best friend" analogy provides a simple interpretation.

With all of the five-subset solutions, the interpretability of the clusters declined noticeably. Figure 6, for instance, shows one of these solutions. Here, there are two clusters that are rather anomalous: DOG, TIGER, KNIGHT and KNIGHT, HUSBAND, PARTNER, MOTHER, DOG. It is worth noting that the weights associated with these two clusters are quite low and that the VAF figure is not any larger than the corresponding value for the four-cluster solutions. Consequently, it seems reasonable to ignore cases like this.

On the basis of this admittedly one-sided view of the data obtained through interactive clustering, the following simple conclusions can be drawn: (1) There is a fundamental perceived separation between the animal- and human-oriented nouns. (2) There are secondary, but interpretable, associations across the basic categories with respect to COOK and TROUT, and PARTNER, DOG, and HUSBAND, the first based on the preparation of food and the second based on a traditional pet-owner alliance. Other properties in the data either do not meet the preset criteria for acceptability or fail to exhibit substantial interpretability.

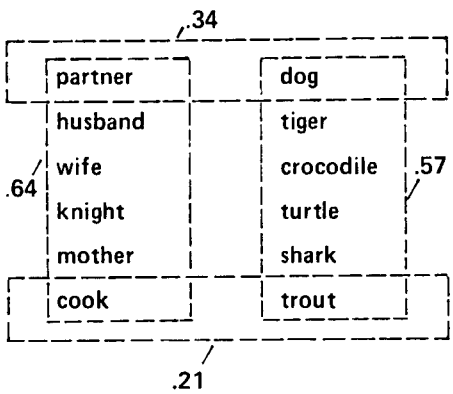

Figure 4. Four-cluster solution–A.

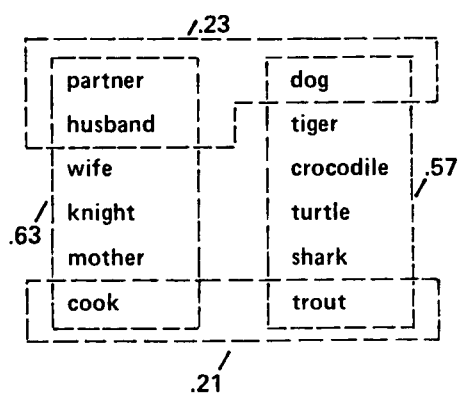

Figure 5. Four-cluster solution–B.

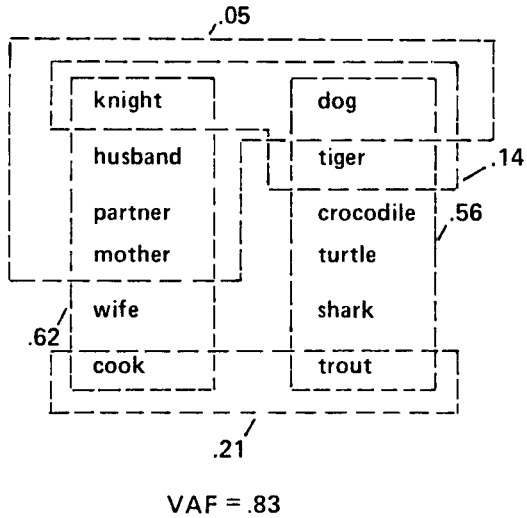

Figure 6. Five-cluster solution.

By way of demonstrating that the flexibility of the current method does not carry with it additional computational costs, the following exercise was carried out. First, the new program based on the Peay algorithm and ADCLUS model fitting was allowed to generate all hierarchical cluster sets over the full range of all parameters and test them with the ADCLUS model. Sixty-eight nontrivial cluster sets (solutions) were produced in all. As a separate exercise, the MAPCLUS program was used on the same data to obtain only the best-fitting two-cluster solution for the data.[2]

It was possible to generate and test the 68 clustering solutions with the ADCLUS regression model in 25% of the CPU time required by MAPCLUS to obtain the single two-cluster solution.

SUMMARY

By way of summarizing, the general advantages of the approach described here are as follows.

First, data collection time can be cut in half with remarkably little loss in accuracy when the estimation techniques proposed by Furnas (1980) are used. One simply generalizes the standard objects-by-features case discussed by Furnas to the case of objects-by-objects (in which the two object sets are different).

Second, the clustering routine, especially when used in an exploratory way, eliminates a large number of partitions of the stimulus set (or clique sets in graph-

theoretic terminology) that have no intrinsic meaning, or which account for so little of the variance in the data as to render them useless.

Due to the graph-theoretic approach to the clustering problem, there is a family of hierarchies and cluster sets within hierarchies from which the analyst may choose. These solutions are natural extensions of the most popular nonmetric clustering methods, and yet they are susceptible to testing via the ADCLUS model-fitting approach.

Finally, because the data analyst can monitor the process as it is occurring, he or she can select one or more solutions that optimize the balance between meaningfulness and goodness of fit.

The ultimate conclusion is that over and above the flexibility of the current approach and the many advantages it offers, the data analysis portion is cheaper by far in terms of both computer memory and processing time than rival methods based on mathematical programming such as MAPCLUS (Arabie & Carroll, 1980a).

REFERENCES

ARABIE, P., & CARROLL, J. D. *How to use MAPCLUS, a computer program for fitting the ADCLUS model.* Murray Hill, N.J: Bell Telephone Laboratories, 1980. (a)

ARABIE, P., & CARROLL, J. D. MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika,* 1980, **45**, 211-235. (b)

BLASHFIELD, R. K. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin,* 1976, **83**, 377-388.

CLIFF, N., GIRARD, R., GREEN, R. S., KEHOE, J. F., & DOHERTY, L. M. INTERSCAL: A TSO FORTRAN IV program for subject computer interactive multidimensional scaling. *Educational and Psychological Measurement,* 1977, **37**, 185-188.

FURNAS, G. W. *Objects and their features: The metric representation of two-class data.* Unpublished doctoral dissertation, Stanford University, 1980.

HUBERT, L. J. Some applications of graph theory to clustering. *Psychometrika,* 1974, **39**, 283-309.

JOHNSON, S. C. Hierarchical clustering schemes. *Psychometrika,* 1967, **32**, 241-254.

PEAY, E. R. Hierarchical clique structures. *Sociometry,* 1974, **37**, 54-65.

SATTATH, S., & TVERSKY, A. Additive similarity trees. *Psychometrika,* 1977, **42**, 319-345.

SHEPARD, R. N., & ARABIE, P. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review,* 1979, **86**, 87-123.

WHALEY, C. P. Collecting paired comparison data with a sorting algorithm. *Behavior Research Methods & Instrumentation,* 1979, **12**, 147-150.

WHALEY, C. P. Computer-augmented decision making. *Behavior Research Methods & Instrumentation,* 1981, **13**, 294-297.

ZURIF, E. B., CARAMAZZA, A., MYERSON, R., & GALVIN, J. Semantic feature representations for normal and aphasic language. *Brain and Language,* 1974, **1**, 167-187.

NOTES

1. These data were collected in 1974 while the author was at the University of Waterloo. The nouns were taken from a study by Zurif, Caramazza, Myerson, and Galvin (1974) to facilitate comparisons between the two studies.

2. MAPCLUS will only permit the user to specify a single number of clusters. Separate runs must be made to look at different numbers of clusters.