# SOME STATISTICAL PROCEDURES TO AID IN THE EVALUATION OF A CLUSTER ANALYSIS

By

JAMES MICHAEL NORTON

Bachelor of Science in Mathematics
University of Vermont
Burlington, Vermont
1970

Master of Arts
University of Vermont
Burlington, Vermont
1972

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
December, 1975

SOME STATISTICAL PROCEDURES TO AID IN THE

EVALUATION OF A CLUSTER ANALYSIS

Thesis Approved:

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to thank Dr. William D. Warde for his guidance in the preparation of this thesis. I would especially like to thank him for enduring my repeated slaughters of the "King's English".

I would also like to thank Professors J. Leroy Folks, Barbara Weiner and Shair Ahmad for serving on my advisory committee and for offering many useful suggestions in the preparation of this thesis.

I would like to thank Dr. David L. Weeks for serving on my advisory committee and for his employment of me on the Eglin Research Contract for the previous two and a half years.

I would like to thank my parents, Norma and James Norton, for their constant encouragement and many sacrifices during my lifetime.

I would also like to thank Dr. David E. Bee for kindling my interest in statistics, and Dr. David L. Sylwester for encouraging me to pursue this degree, when it appeared to me to be impossible.

There are many faculty and students, too numerous to mention with whom I have spent many enjoyable hours here in Stillwater and I wish to thank them all.

I would like to express my thanks to Jan Jones whose efforts in helping prepare this thesis should earn her a degree.

Lastly, I would like to thank my wife, Marguerite, for all the wonderful contributions to my life she has made in the last seven years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

# CHAPTER I

# INTRODUCTION AND REVIEW OF THE LITERATURE

## General Statement of the Problem

There are at least two distinct types of problems for which cluster analysis has been utilized. The first type of clustering will be referred to as mathematical cluster analysis, which may be appropriate when data has been collected on each of $n$ units and a scientist wishes to make inferences about only those $n$ units. The existence of "clusters" is determined by the definition of a cluster in this situation. For some definitions of a cluster there are algorithms which will locate all the clusters in the data. Since there are so many possible definitions of a cluster, it is unlikely that for a specific definition an algorithm exists to locate all the clusters according to that definition. There are many algorithms in use today that find approximate clusters, while using relatively little computer time. The evaluation of this type of clustering will have to center on the adequacies of the definitions of clusters and with how similar the approximate clusters are to the real clusters. Until a single satisfactory definition of a cluster is accepted, many aspects of the evaluation of mathematical clustering will remain arbitrary.

The second type of clustering will be referred to as inferential clustering, which may be appropriate when data has been collected on each of $n$ units and a scientist wishes to make inference to more

than the n units he has observed. The n units are regarded as being a sample from one or more populations, whereas mathematical clustering regards the n units as being one or more populations.

There has been a great deal of confusion with respect to the evaluation of these two types of clustering, because many of the same algorithms are used with both types of clustering. Clustering algorithms attempt to group units into "clusters" such that the units within a cluster are homogeneous and such that units in different clusters are heterogeneous. A major problem is that the algorithm does not know whether it is operating on a population or on a sample from a population.

The problem that this study is concerned with, simply stated, is to discover practical methods to aid the user of cluster analysis in the evaluation of his cluster analysis.

## General Review of Some of the Clustering Algorithms in Existence

Currently there exist a large number of clustering algorithms which operate directly or indirectly on a data matrix. All of the algorithms can be applied to either populations or samples, as mentioned in the previous section, with "clusters" resulting. The class of clustering algorithms that will be of primary importance for this paper will be the sequential, agglomerative, hierarchal algorithms. Most agglomerative procedures are sequential, and begin at stage one with all n observations regarded as clusters each containing a single observation. The agglomerative procedures compute a matrix of pairwise similarities (correlations, etc.) or dissimilarities (distances, etc.).

The two units which are most similar (or least dissimilar) are grouped together to be regarded as a single unit at stage two. The similarities or dissimilarities matrix is then recomputed regarding the two previously grouped observations as one, so that the dimension of the new similarities matrix will be $(n-1)x(n-1)$. This procedure is repeated until all $n$ observations are grouped together into a single unit at stage $n-1$. One of the differences between the sequential, agglomerative, hierarchal algorithms is the linkage method used. The concept of linkage is associated with the need to compute the distance (or similarity) between group $L$ and the group $(J \cup K)$, which was formed at the previous stage as the union of groups $J$ and $K$. Lance and Williams (1967) have developed a formula which will compute the distance between group $L$ and group $(J \cup K)$ for many of the common linkage methods by variation of the four parameters in the formula

$$U_{(J \cup K),L} = \alpha_J U_{J,L} + \alpha_K U_{K,L} + \beta U_{J,K} + \gamma \left| U_{J,L} - U_{K,L} \right| \quad (1)$$

where $\alpha_J$, $\alpha_K$, $\beta$ and $\gamma$ are determined by the linkage method used, $U_{(J \cup K),L}$ is the distance between group $L$ and group $J \cup K$ at stage $i$, $U_{M,N}$ is the distance between group $M$ and group $N$ at stage $i-1$. There is some confusion about whether Euclidean or Squared Euclidean distance is most appropriately used with this formula. This question will be considered in Chapter III. If we substitute $\alpha_J = 1/2$, $\alpha_K = 1/2$, $\beta = 0$, $\gamma = -1/2$ the linkage method is called single linkage. If we substitute $\alpha_J = 1/2$, $\alpha_K = 1/2$, $\beta = 0$, $\gamma = 1/2$ then the linkage method is called complete linkage.

Regardless of the method selected, the results may be graphically displayed by a tree (dendogram) or by a contour map. If the number of

groups or clusters is known, then the scientist uses the appropriate stage of the algorithm to indicate which observations have been grouped together. If the number of groups or clusters is not known, then there are theoretically $n$ choices for the number of groups to best represent the data. In either case, it is not clear how the results are to be interpreted. Some discussion and various applications of sequential agglomerative algorithms and sequential agglomerative algorithms with stopping rules may be found in Lance and Williams (1967), Sneath (1957), Sokal and Michener (1958), Sokal and Sneath (1963) and Sneath and Sokal (1973).

A second major class of clustering algorithms is the divisive algorithms, which unlike the agglomerative procedures does not compute a similarities or distance matrix. The divisive algorithms are sequential and they do impose a hierarchal structure on the data. After beginning at stage 1 with all $n$ observations being regarded as one cluster, the next step is to divide the observations into two groups in such a way as to optimize some pre-determined criterion. One such criterion is the ratio of between groups sum of squares to within group sum of squares, in the univariate case (Edwards and Cavalli-Sforza, 1965). The result is of course two groups at stage two with $n_1$ and $n_2$ observations in each group. This process is repeated on each of the resulting clusters until all clusters have two or less observations (it is impossible to split a group having two observations or less using this criterion), or until the procedure is halted by a stopping rule. Other criteria which may be considered as generalizations of the sum of squares criterion, are used with multivariate data. Some of these criteria are: Minimum $\prod_{i=1}^{g} |W_i|^{n_i}$, Minimum $|W|$, Min $g^2|W|$ and

Min trace $W$, where $g$ is the number of groups, $W$ is the pooled within group covariance matrix, $W_i$ is the within group covariance matrix for group $i$ and $n_i$ is the number of units in group $i$. As with the agglomerative procedures, a tree or contour map may be used to graphically display the results. The same problem remains of interpreting the results, even if the number of groups is known.

Various optimization criteria and arguments for and against the usage of divisive algorithms can be found in Williams and Dale (1965), Edwards and Cavalli-Sforza (1965), Orloci (1967), Ling (1971), Sneath and Sokal (1973), Scott and Symon (1971), Everitt (1974).

A third major class of clustering algorithms might be called simultaneous techniques. There is no hierarchal structure imposed because the number of groups, $K$, is required as input to the algorithm. The observations are then partitioned into $K$ groups according to some pre-determined criterion, such as the ratio of between group sum of squares to within group sum of squares, in the univariate case. The most common practice appears to be running the procedure several times with a different $K$ each time and then selecting the result which subjectively seems best. Techniques of this type are presented and/or discussed in Fisher (1958), Ward (1963), Rubin (1967), Friedman and Rubin (1967), Edwards and Cavalli-Sforza (1965) and Ling (1971).

There are many more techniques which do not strictly belong in any of these categories, but these are similar to the above techniques in many respects.

All clustering algorithms have one feature in common; they produce groups or clusters of observations. The vast majority of the literature until very recently has been concerned with the comparison

of techniques or the proposal of new techniques and their application to specific problems. These efforts appear to be directed toward mathematical clustering because there are no distributional results given, but the algorithms can of course be used with inferential clustering. The inferential clustering problem has received more attention recently and will be very important in this study. As pointed out previously, a clustering algorithm will cluster any set of data that is input to the algorithm. If a scientist collects a sample from a single population and a clustering algorithm "discovers" several clusters, then the scientist is likely to be misled concerning the structure of the population from which the data was sampled. The literature and mathematical problems associated with inferential clustering will be discussed in some detail in the next chapter.

## A Numerical Example of a Simple Agglomerative Clustering Procedure

A specific example will be given to illustrate the details of a clustering algorithm. The algorithm selected is a sequential, agglomerative, hierarchal procedure using squared Euclidean distance as a measure of dissimilarity. The linkage method used will be complete linkage, which corresponds to $\alpha_J = 1/2$, $\alpha_K = 1/2$, $\beta = 0$, $\gamma = 1/2$, in the Lance and Williams (1967) formula mentioned previously (1).

Following much the same notation as Mrachek (1972) and Warde (1975), let $X$ be a $p \times n$ data matrix where each of the $n$ columns represents a p-variate response vector.

Let $X_i$, $i = 1, \ldots, n$ represent one of the $n$ observation vectors.

Let $M_1$ be a matrix, nxn, of pairwise squared Euclidean distances between the columns of X. Let $d_{ij}$, i, j = 1, ..., n represent the elements of $M_1$. $M_1$ will be symmetric with zeroes on the main diagonal.

The following iterative procedure is used to cluster a data set using the complete linkage criterion. The procedure begins with i = 1 and terminates when i = n-1.

1. Select the minimum distance from the $\binom{n+1-i}{2}$ elements above the main diagonal of $M_i$.

2. The two vectors which were separated by the minimum distance in 1 above are to be regarded as a single group for further computational purposes.

3. Recompute the distance matrix which will now have dimension (n-i) x (n-i) using the complete linkage criterion. Call this new matrix $M_{i+1}$. Note that only distances between the new group and the other vectors need be computed.

4. If i is less than n-1, then add one to i and go back to step 1. If i is equal to n-1 then stop.

Let $M_i$* be the matrix $M_i$ with row and column labels adjoined for clarity.

Example: Suppose it is desired to cluster the following bivariate observations:

$$\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix}.$$

Using a sequential, agglomerative, hierarchal method with complete linkage and squared Euclidean distance, we would have:

$$X = \begin{bmatrix} 0 & 0 & 1 & 4 & 5 \\ -1 & 1 & 1 & 4 & 5 \end{bmatrix}$$

$M_1^* =$

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | 0 | 4 | 5 | 41 | 61 |
| $X_2$ | 4 | 0 | 1 | 25 | 41 |
| $X_3$ | 5 | 1 | 0 | 18 | 32 |
| $X_4$ | 41 | 25 | 18 | 0 | 2 |
| $X_5$ | 61 | 41 | 32 | 2 | 0 |

The execution of the procedure outlined above will be illustrated.

At step 1 of iteration 1, the minimum distance above the main diagonal of $M_1$ is 1, hence $X_2$ and $X_3$ are grouped together. At step 3 of iteration 1 the new distances are: $d_{1,(2\cup3)} = \text{Max}(4,5) = 5$, $d_{4,(2\cup3)} = \text{Max}(25,18) = 25$, and $d_{5,(2\cup3)} = \text{Max}(41,32) = 41$. Hence

$M_2^* =$

| | $X_1$ | $X_2 \cup X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| $X_1$ | 0 | 5 | 41 | 61 |
| $X_2 \cup X_3$ | 5 | 0 | 25 | 41 |
| $X_4$ | 41 | 25 | 0 | 2 |
| $X_5$ | 61 | 41 | 2 | 0 |

At step 4 of iteration 1, $i = 1$ which is less than $i = 4$, so $i$ is set equal to 2.

At step 1 of iteration 2, the minimum distance above the main diagonal of $M_2$ is 2, hence $X_4$ and $X_5$ are grouped together. At step 3 of iteration 2 the new distances are $d_{1,(4\cup5)} = \text{Max}(41,61) = 61$, $d_{(2\cup3),(4\cup5)} = \text{Max}(d_{4,(2\cup3)}, d_{5,(2\cup3)}) = \text{Max}(25,41) = 41$. Hence

$$M_3^* = \begin{array}{c|ccc} & X_1 & X_2 \cup X_3 & X_4 \cup X_5 \\ \hline X_1 & 0 & 5 & 61 \\ X_2 \cup X_3 & 5 & 0 & 41 \\ X_4 \cup X_5 & 61 & 41 & 0 \end{array}$$

At step 4 of iteration 2, $i = 2$ which is less than $i = 4$, so $i$ is set equal to 3.

At step 1 of iteration 3, the minimum distance above the main diagonal of $M_3$ is 5, hence $X_1$ is grouped with $X_2 \cup X_3$. At step 3 of iteration 3 the new distances are $d_{(4\cup5),(1\cup2\cup3)} = \text{Max}(d_{(4\cup5),1}, d_{(4\cup5),(2\cup3)}) = \text{Max}(61,41) = 61$. Hence

$$M_4^* = \begin{array}{c|cc} & X_1 \cup X_2 \cup X_3 & X_4 \cup X_5 \\ \hline X_1 \cup X_2 \cup X_3 & 0 & 61 \\ X_4 \cup X_5 & 61 & 0 \end{array}$$

At step 4 of iteration 3, $i = 3$ which is less than $i = 4$, so $i$ is set equal to 4.

At step 1 of iteration 4, the only remaining distance above the main diagonal is 61, hence $X_4 \cup X_5$ is grouped with $X_1 \cup X_2 \cup X_3$. At step 3 of iteration 4 there are no new distances. Hence

$$M_5^* = \begin{array}{c|c} & X_1 \cup X_2 \cup X_3 \cup X_4 \cup X_5 \\ \hline X_1 \cup X_2 \cup X_3 \cup X_4 \cup X_5 & 0 \end{array}$$

At step 4 of iteration 4, $i = 4$ which is not less than 4, so that the procedure is terminated. The data may be summarized by a tree diagram (or dendogram) (see Figure 1).

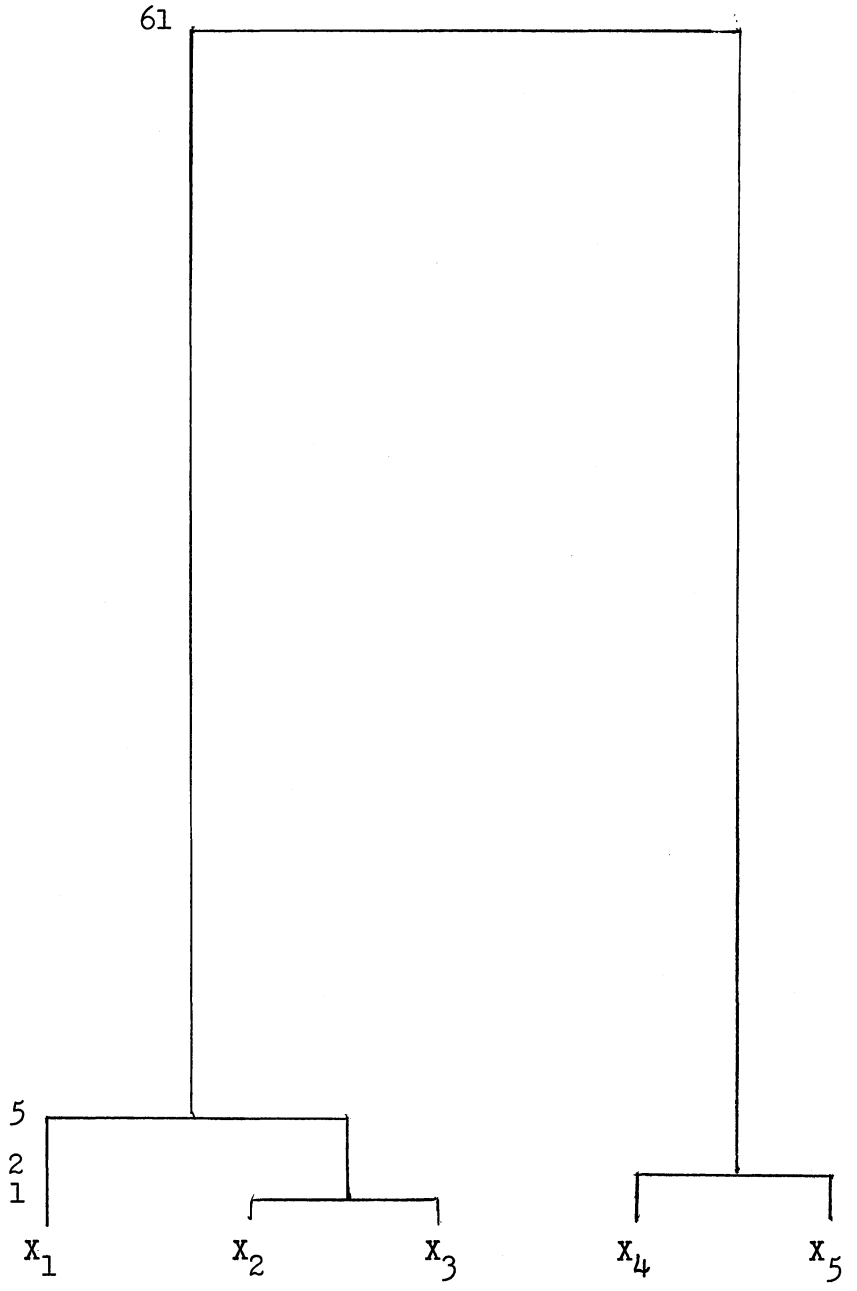


Figure 1. Dendogram (or Tree Diagram) for Data Given in Section 3

## Objectives and Scope of This Study

The first objectives of this study are concerned with the needs of the potential user of cluster analysis. First, the user of cluster analysis needs to be aware of the type of evidence that a cluster analysis can provide, in order to avoid confusion and misinterpretation of his data. Second, when a test is appropriate, the potential user of cluster analysis needs a test that can be used immediately. This test procedure should be reasonably easy to use and should be inexpensive to use. This study strives to provide a conservative test that can be used immediately, and also gives estimated percentage points of a level $\alpha$ test. Since the percentage points are estimated the test may not in reality be a level $\alpha$ test. The percentage points are estimated for as many sample sizes as cost will allow. Third, the potential user needs help in deciding whether or not he should spend his time and money on a cluster analysis. To help the potential user decide whether a cluster analysis will be worthwhile, this study will provide some estimates of the power of the proposed procedures assuming various, "representative" alternatives.

The other objective is to provide some guidance to future researchers who will be concerned with statistical and mathematical problems in cluster analysis, by pointing out where some of the main problems lie. The author will also suggest some test procedures that appear promising, but which the author has been unable to pursue in depth.

It is the author's opinion that the simple cases of clustering must be understood before more ambitious goals are attempted, so the

single variable cases have been selected as a logical starting point. It is also appropriate to begin with single variable cases because little progress has been made in this area. The generalization of single variable results to multivariate cases is not immediate because we lose the ability to rank observations and there are more unknown parameters present.

This study will be limited to the four major sequential, agglomerative, hierarchal clustering procedures that are in use today. The four procedures selected include boundary type algorithms and "representative" type algorithms, which require little enough computer time to be considered practical.

Many empirical results are presented in this paper, because the complexity of the mathematics prohibited other types of solution. The computer programs, written in FORTRAN, which were used to generate the results contained in this study are listed in the Appendix. They are ready to use to generate estimated percentage points for additional sample sizes and estimates of the power of these procedures, with only minor changes in the parameters being necessary.

The application of mathematical clustering will be illustrated by considering several approaches to grouping target elements from a complex target element configuration supplied by the Air Force.

# CHAPTER II

## SOME MATHEMATICAL AND STATISTICAL PROBLEMS

### The Problem of Definition of a Cluster

Although the concept of a cluster is central to the field of cluster analysis, an adequate definition still remains elusive. One seemingly logical and straightforward approach would be to devise a mathematical definition of a cluster which would incorporate the concepts of homogeneity (or closeness) within clusters and heterogeneity (or separateness) of points in different clusters. A simple example illustrates that a single definition cannot be adequate for all purposes. Consider a collection of houses located in Stillwater and another collection located in Perry. The houses in Stillwater might be considered a cluster since the houses in Stillwater are relatively close to each other and are relatively far from the houses in Perry. Suppose we next consider a collection of houses in Brandon, Vermont, in addition to the others. Now it appears that the houses in Stillwater and Perry may be considered as part of the same cluster because the distances between houses in Oklahoma are relatively close compared to the distances between houses in Oklahoma and in Vermont. Ling (1971) recognizes this problem and makes some progress toward a definition of different levels of clustering. His approach will be discussed in more detail in Chapter IV.

Let us assume for the sake of argument that an adequate definition of a cluster is established for some given problem. It would be possible to find all the clusters according to this definition, although this might involve a lengthy search through all possible partitions of the data. It may be rather cumbersome and costly to search through all the partitions of the data, so it may be advantageous to use an algorithm such as a complete linkage agglomerative procedure to find "approximate" clusters. Emphasis must be placed on "approximate" because the units the algorithm groups together may not be clusters by the definition established, and in addition the algorithm may not succeed in finding all the clusters even if it does find some of them. Some work (Rand, 1969) has been done to determine how well certain algorithms "retrieve" clusters according to several definitions of clusters.

Another type of definition is an operational one, which does not seem as desirable from a philosophical point of view. This definition is: The algorithm generates clusters at each stage of the clustering procedures and the clusters are of different levels. This last definition although less than satisfactory is probably the most frequently used definition in practice.

All the previous discussion has been restricted to cases in which the inferences are to be made only about the structure of the units observed. Suppose a scientist wishes to make inferences about more units than the $n$ units observed. The same definition may be used with this inferential clustering as was used with mathematical clustering, with the resulting clusters being called sample clusters or approximate sample clusters to distinguish them from the mathematical

clusters. The choice of the terminology sample cluster may be unfortunate if it implies that it is estimating a "population cluster". A sample cluster is merely a collection of observations which are "close", and all the sample clusters together provide evidence to help the scientist decide whether the sample was from one parent population or from two or more parent populations. This study will allow four algorithms to generate clusters, and will attempt to discover which of them is most sensitive in detecting the presence of two or more normal populations. The basic question of whether or not there is enough evidence to conclude the sample was not drawn from a single normal population will be explored in more mathematical detail in the next section.

## The Problem the Cluster Analyst Would like to be Able to Solve in the Univariate Case

The scientist would like to be able to take a sample of size $n$ and have a test available to test whether all the observations were from the same population, and he wants the test to be sensitive to any departure from his null hypothesis. This is essentially a one-way classification problem with no replicates.

In order to state the above hypothesis in parametric form, assume the observations are all from normal populations with equal but unknown variance and possibly different means.

Let $X_i \sim N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$, $\sigma^2$ unknown.

The desired test hypothesis is:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_n = \mu \quad \text{versus}$$

$$H_A: \text{Not } H_0 .$$

When a simple null hypothesis is to be tested against a simple alternative hypothesis, the Neyman-Pearson Lemma states that the best level $\alpha$ test is given by the likelihood ratio test. When a test is desired for a composite null versus a composite alternate, one method of constructing a test is the Generalized Likelihood ratio test. The Generalized Likelihood Ratio (GLR) test has no guaranteed optimality properties, but it has been useful in the construction of many good test procedures.

Consider the following attempt to construct a GLR test for the aforementioned hypotheses. Let $L(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)$ be the likelihood function. Let $L_{H_0}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)$ be the likelihood function restricted by the null hypothesis. Let $L_{H_A}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)$ be the likelihood function restricted by the alternative hypothesis.

The GLR test will reject $H_0$ in favor of $H_A$ at level $\alpha$ whenever

$$\frac{\underset{(\mu_1, \mu_2, \ldots, \mu_n, \sigma^2)}{\text{MAX}} L_{H_A}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)}{\underset{(\mu, \sigma^2)}{\text{MAX}} L_{H_0}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)} > K_\alpha$$

where $K_\alpha$ is a constant determined by the level of the test, $\alpha$.

$$L_{H_0}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu^2)}.$$

In order to find the maximum of the above function with respect to choices of $\mu$ and $\sigma^2$ it is sufficient to find the maximum of the logarithm of the function, because a logarithmic transformation is monotonic.

$$\ln L_{H_0}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 .$$

$$\frac{\partial}{\partial\mu}\ln L_{H_0}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{\partial}{\partial\sigma^2}\ln L_{H_0}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 .$$

The partial derivatives above are both set equal to zero, and the resulting equations are solved simultaneously. Thus,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

or

$$\sum_{i=1}^{n}x_i - n\mu = 0$$

which yields

$$\hat{\mu} = \frac{\sum_{i=1}^{n}x_i}{n} .$$

Similarly,

$$\frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

or

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

which yields

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} .$$

The simultaneous solution of these equations is:

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} .$$

These two values do lead to a maximum of the likelihood function, with that maximum value being given by:

$$(2\pi)^{-n/2}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n}\right)^{-n/2} e^{-n/2} .$$

There are no problems yet since these results are well known.

Next consider the maximization of $L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2)$ with respect to choices of $\mu_1, \mu_2, \ldots, \mu_n, \sigma^2$.

$$L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_i)^2} .$$

Again it is sufficient to maximize the logarithm of the function.

$$\ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_i)^2$$

$$\frac{\partial}{\partial\mu_i}\ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = \frac{1}{\sigma^2}(x_i-\mu_i) \quad \text{for } i=1, \ldots, n.$$

$$\frac{\partial}{\partial\sigma^2}\ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = \frac{-n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu_i)^2$$

The partial derivatives above are all set equal to zero.

$$\frac{1}{\sigma^2}(x_i-\mu_i) = 0 \qquad \text{for } i = 1, \ldots, n.$$

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu_i)^2 = 0 \quad .$$

The simultaneous solution of these equations is:

$$\hat{\mu}_i = x_i \qquad \text{for } i = 1, \ldots, n$$
$$\hat{\sigma}^2 = 0 \quad .$$

Substitution of these values into the likelihood function does not lead to a maximum, because there is no maximum. The GLR criterion has failed to lead to any test, good or otherwise. One of the problems seems to be that of obtaining a good estimate of $\sigma^2$, but that should not be surprising since there are only $n$ observations and $n+1$ unknown parameters.

The author attempted to formulate tests on the basis of intuition and other subjective methods. The following procedure will illustrate the difficulties involved.

Let $x_1, \ldots, x_n$ be the random sample of size $n$. Let $y_i = x_i - x_n$ for $i = 1, 2, \ldots, n-1$. Then it follows that:

$$\sum_{i=1}^{n-1} y_i = \sum_{i=1}^{n-1} x_i - (n-1)x_n$$

$$\overline{Y} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} - x_n \quad .$$

Subject to the null hypothesis, that the $x$'s are a random sample from $N(\mu, \sigma^2)$ it follows that $\overline{Y} \sim N(0, \frac{n}{n-1}\sigma^2)$.

Let $$X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_{n-1} \end{bmatrix} \quad .$$

Let

$$S_Y^2 = \frac{Y'\left[I - \frac{1}{n-1}J_{n-1}^{n-1}\right]Y}{n-2}$$

where $I$ is the $(n-1) \times (n-1)$ identity matrix and $J_{n-1}^{n-1}$ is an $(n-1) \times (n-1)$ matrix whose every element is a 1.

$\overline{Y}$ and $S_Y^2$ can be written in terms of $X$ instead of $Y$.

$$\overline{Y} = \frac{(1, 1, 1, \ldots, -(n-1))X}{n-1}$$

$$S_Y^2 = \frac{X'\left[\begin{array}{c|c} I - \frac{1}{n-1}J_{n-1}^{n-1} & O \\ \hline O & O \end{array}\right]X}{n-2}$$

From Theorem 3 of Searle (1971, p. 59), $\bar{Y}$ and $S^2_Y$ are distributed independently, under the null hypothesis. From Theorem 2 of Searle (1971):

$$\frac{X'\begin{bmatrix} I - \frac{1}{n-1}J_{n-1} & 0 \\ 0 & 0 \end{bmatrix}X}{\sigma^2} \quad \text{is distributed as } \chi^2(n-2)$$

where the degrees of freedom are determined by the rank of

$$\begin{bmatrix} I - \frac{1}{n-1}J_{n-1} & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} I - \frac{1}{n-1}J_{n-1} & 0 \\ 0 & 0 \end{bmatrix}$$

is an idempotent matrix and the rank of an idempotent matrix is its trace. The trace of the above matrix is $n-2$, hence the degrees of freedom of the $\chi^2$ variable are $n-2$.

It then follows that under the assumptions of the null hypothesis that:

$$\frac{\bar{Y}}{\sqrt{S^2_{\bar{Y}}}} \sim t(n-2),$$

where $t$ represents the Student's $t$ distribution.

There are some obvious deficiencies with this procedure: The test has "acceptable" power if the true alternative is $\mu_1 = \mu_2 = \ldots = \mu_{n-1} \neq \mu_n$. However, if the true alternative is such that

$$\frac{\mu_1 + \mu_2 + \ldots + \mu_{n-1}}{n-1} = \mu_n \quad \text{where } \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \neq \mu_{n-1},$$

then the test can have power less than $\alpha$. Thus, the test is biased with respect to some alternatives.

If there exists an unbiased test for this general alternative, no one, including the present author, has been successful in finding it.

The next logical step it would seem, would be to search for a test which will be sensitive to the most interesting or most important alternatives, instead of being sensitive to all possible alternatives. There are several approaches of this type in the next section.

## Some Procedures Based on the Likelihood Ratio

Since one of the objectives of a cluster analysis is to reduce or condense data, a scientist given the choice of a procedure which is sensitive to large numbers of clusters, or a procedure which is sensitive to small numbers of clusters, will probably choose the latter. Engelman and Hartigan (1969) formulate the problem in terms of a two cluster alternative.

$H_0$: $x_1, \ldots, x_n$ are a random sample from $N(\mu, \sigma^2)$

versus

$H_A$: For some partition of $x_1, \ldots, x_n$, the cluster $x_{11}, x_{12}, \ldots, x_{1n_1}$ is a sample from $N(\mu_1, \sigma^2)$ and the cluster $x_{21}, x_{22}, \ldots, x_{2n_2}$ is a sample from $N(\mu_2, \sigma^2)$, where $n_1 \geq 1$, $n_2 \geq 1$, $n_1 + n_2 = n$, $n > 2$.

The concept of a partition has been introduced in the Engelman and Hartigan alternative hypothesis. A partition is an assignment of the sample values $x_1, \ldots, x_n$ to a known number of groups. The true partition is the correct assignment of the observations to their respective groups.

An example may help illustrate the point. Suppose there are three observations A, B, C and suppose that observations A and C came from population 1, while observation B came from population 2. The observations may be partitioned into 2 groups in the following way:

| Partition | Group 1 | Group 2 |
|---|---|---|
| 1 | A | BC |
| 2 | B | AC |
| 3 | C | AB |
| 4 | BC | A |
| 5 | AC | B |
| 6 | AB | C |

Partition 5 is the true partition.

In most statistical problems the true partition is known; the "usual" two group t test is an example. Under the alternative, there are $n_1$ observations from population 1, and $n_2$ observations from population 2, where $n_1$ and $n_2$ are known. It is also known which observations came from population 1 and which observations came from population 2.

For notational purposes, let the partition be denoted by H and let $x_{11}, x_{12}, \ldots, x_{1n_1}$ be the observations assigned to group 1; let $x_{21}, x_{22}, \ldots, x_{2n_2}$ be the observations assigned to group 2, etc.

Engelman and Hartigan (1969) used the Generalized Likelihood Ratio criterion to formulate a test procedure.

Reject $H_0$ in favor of $H_A$ when

$$\frac{\underset{(\mu_1,\mu_2,\ldots,\mu_n,\sigma^2,\ H)}{\text{MAX}} L_{H_A}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2, H)}{\underset{(\mu,\ \sigma^2)}{\text{MAX}_2} \qquad L_{H_0}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2)} > K_\alpha$$

where $K_\alpha$ is a constant determined by the level of the test $\alpha$.

The likelihood function restricted by the null hypothesis is:

$$L_{H_0}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

The maximum of this function with respect to $\mu$ and $\sigma^2$ is the same as in the previous section.

$$\hat{\mu} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

are the values of $\mu$ and $\sigma^2$ which maximize the likelihood function. The maximum is:

$$(2\pi)^{-n/2}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}\right)^{-n/2}\exp\left(-\frac{n}{2}\right) .$$

The likelihood function restricted by the alternative hypothesis is:

$$L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2, H) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}$$
$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right)\right) .$$

It is necessary to find the maximum of this function with respect to $\mu_1$, $\mu_2$, $\sigma^2$ and the partition, H, of the sample into two groups.

Fix the partition, H, and maximize with respect to $\mu_1$, $\mu_2$, and $\sigma^2$. As before it is sufficient to maximize the logarithm of the likelihood function.

$$\ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2,\ H) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2)$$

$$-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right] .$$

Hence

$$\frac{\partial}{\partial\mu_1} \ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2,\ H) =$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu_1),$$

$$\frac{\partial}{\partial\mu_2} \ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2,\ H) =$$

$$\frac{1}{\sigma^2}\sum_{j=1}^{n_2}(x_{2j}-\mu_2),$$

and

$$\frac{\partial}{\partial\sigma^2} \ln L_{H_A}(x_1,x_2,\ldots,x_n,\mu_1,\mu_2,\ldots,\mu_n,\sigma^2,\ H) =$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right] .$$

Setting these three partial derivatives equal to 0 and solving simultaneously, the values of $\mu_1$, $\mu_2$ and $\sigma^2$ that maximize the likelihood for a fixed partition are:

$$\hat{\mu}_1 = \bar{x}_1,\ \hat{\mu}_2 = \bar{x}_2,\ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2}{n} .$$

The maximum is then given by:

$$\underset{H}{MAX} \quad (2\pi)^{-n/2} \left( \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2}{n} \right)^{-n/2} \exp(-n/2) \quad .$$

Hence, the likelihood ratio criterion rejects $H_0$ in favor of $H_A$ whenever

$$\underset{H}{MAX} \frac{(2\pi)^{-n/2} \left( \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2 \right)^{-n/2} n^{n/2} \exp(-n/2)}{(2\pi)^{-n/2} \left( \sum_{i=1}^{n}(x_i - \bar{x})^2 \right)^{-n/2} n^{n/2} \exp(-n/2)} > K_\alpha$$

or

$$\underset{H}{MAX} \quad \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2} > K_{1\alpha} \quad .$$

For a fixed partition this is equivalent to

$$\frac{B + W}{W} > K_{1\alpha}$$

or

$\frac{B}{W} > K_{2\alpha}$ where $B$ is the between cluster sum of squares and $W$ the pooled within cluster sum of squares.

The final form of the test as derived by Engelman and Hartigan (1969) is:

Reject $H_0$ in favor of $H_A$ whenever

$\underset{H}{MAX} \frac{B}{W} > K_2$ where $K_2$ is a constant determined by the level of the test, $\alpha$.

The authors then generate percentage points for this test based on 100,000 replications. This test procedure is most appropriate for use with the divisive algorithm that divides the observations into two groups in such a way as to maximize the ratio of between cluster sum of squares to within cluster sum of squares. Another use will be made of this test, which will be discussed in a later chapter.

Although not given in their paper the above procedure is easy to generalize to 3, 4, ..., n-1 cluster alternatives. To illustrate this consider the three cluster alternative formulated in the following way:

$H_0$: $x_1, \ldots, x_n$ are a random sample from $N(\mu, \sigma^2)$

versus

$H_A$: For some partition of $x_1, \ldots, x_n$ into three clusters, the cluster $x_{11}, x_{12}, \ldots, x_{1n_1}$ is a sample from $N(\mu_1, \sigma^2)$, the cluster $x_{21}, x_{22}, \ldots, x_{2n_2}$ is a sample from $N(\mu_2, \sigma^2)$, and the cluster $x_{31}, x_{32}, \ldots, x_{3n_3}$ is a sample from $N(\mu_3, \sigma^2)$ where $n_1 \geq 1$, $n_2 \geq 1$, $n_3 \geq 1$, $n_1 + n_2 + n_3 = n$, and $MAX(n_1, n_2, n_3) \geq 2$.

The maximum of the likelihood restricted by $H_0$ is the same as before, namely:

$$(2\pi)^{-n/2} \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \right)^{-n/2} \exp(-n/2).$$

The likelihood function restricted by $H_A$ is given by:

$$L_{H_A}(x_1, x_2, \ldots, x_n, \mu_1, \mu_2, \ldots, \mu_n, \sigma^2, H) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}$$

$$\times \quad \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2 + \sum_{k=1}^{n_3}(x_{3k}-\mu_3)^2\right]\right).$$

It is necessary to find the maximum of this function with respect to the partition, H, $\mu_1$, $\mu_2$, $\mu_3$, $\sigma^2$. Fix the partition and maximize with respect to $\mu_1$, $\mu_2$, $\mu_3$ and $\sigma^2$. The algebra is straightforward but tedious. The maximum is given by:

$$\underset{H}{MAX} \quad (2\pi)^{-n/2}\left(\frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2 + \sum_{k=1}^{n_3}(x_{3k}-\bar{x}_3)^2}{n}\right)^{-n/2}\exp(-n/2).$$

The likelihood ratio test is to reject $H_0$ in favor of $H_A$ whenever

$\underset{H}{MAX}\ \frac{B}{W} > C_\alpha$ where $C_\alpha$ is a constant determined by the level of the test $\alpha$.

In a similar fashion the generalized likelihood ratio test for all the observations from one normal population versus the alternative that for some partition of the observations into $k$ clusters, where $2 \leq k \leq n-1$,

$x_{11}, x_{12}, \ldots, x_{1n_1}$ is a sample from $N(\mu_1, \sigma^2)$

$x_{21}, x_{22}, \ldots, x_{2n_2}$ is a sample from $N(\mu_2, \sigma^2)$

$\vdots$

$x_{k1}, x_{k2}, \ldots, x_{kn_k}$ is a sample from $N(\mu_k, \sigma^2)$

is given by rejection of $H_0$ in favor of $H_A$ whenever:

$\underset{H}{MAX} \frac{B}{W} > C_\alpha$, $C_\alpha$ is a constant determined by the level of the test.

These tests cannot be used yet because there are no percentage points available for them. It is unlikely that empirical tables of percentage points for these tests will be developed since the number of possible partitions is $\binom{n-1}{i-1}$ where $i$ is the number of groups in the alternative. This number is smaller for the univariate case than for the general multivariate case, but it does increase rapidly with sample size. An example is for $n = 50$ to find the maximum $\frac{B}{W}$ for 5 clusters requires a search through more than 211,000 partitions.

Lee (1974) has formulated another likelihood ratio test under a slightly different set of assumptions. Instead of regarding $n_1$, the number of observations in cluster 1, as being a fixed but unknown constant as Engelman and Hartigan must have, Lee (1974) assumes that $n_1$ is a random variable whose probability distribution is determined by $n$ and $p$, where $p$ is an unknown mixing parameter. The assumption that $n_1$ is random appears to be a good one, but there are several confusing aspects of Lee's (1974) formulation of a test statistic. Lee (1974) lets $\underset{\sim}{X}$ represent the $n$ sample observations and lets $L(\underset{\sim}{X} \mid n_1, n)$ represent the likelihood function under the alternative suggested by Engelman and Hartigan (1969). He states that Engelman and Hartigan (1969) maximize this likelihood by considering $\underset{n_1}{MAX} \; \underset{\underset{\sim}{X}}{MAX} \; L(\underset{\sim}{X} \mid n_1, n)$. It is not clear what Lee (1974) means by $\underset{\underset{\sim}{X}}{MAX}$, but it does appear to the present author that it was the maximization Engelman and Hartigan had in mind. It seems more logical to assume that

maximization was $\underset{H}{MAX}\ \underset{\mu_1,\mu_2,\sigma^2}{MAX} L(\underset{\sim}{X} \mid n_1, n)$. The maximum of this likelihood function is sought with respect to $\mu_1$, $\mu_2$, $\sigma^2$, p and H.

The values of p, $\mu_1$, $\mu_2$, $\sigma^2$ which maximize the likelihood for a fixed partition are

$$\hat{p} = \frac{n_1}{n}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} = \bar{X}_1$$

$$\hat{\mu}_2 = \frac{\sum_{j=1}^{n-n_1} X_{2j}}{n - n_1} = \bar{X}_2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{n-n_1} (X_{2j} - \bar{X}_2)^2}{n} .$$

The likelihood ratio test is then to reject $H_0$ in favor of $H_A$ whenever

$$\underset{H}{MAX} \left[ \binom{n}{n_1} \left(\frac{n_1}{n}\right)^{n_1} \left(1 - \frac{n_1}{n}\right)^{n-n_1} \right]^{2/n} \frac{B}{W} > K_\alpha .$$

Given the partition, then $n_1$ is known. It is not clear why Lee (1974) attempted to maximize the likelihood over $n_1$ with his assumptions. Lee (1974) then tries to compare the power of this statistic, $C_1^*$ with the power of the Engelman and Hartigan (1969) statistic, C for some values of p the mixing parameter. This type of comparison is not really appropriate because the Engelman-Hartigan procedure is not

formulated in terms of a mixing parameter p. This is the same problem encountered when nonparametric and parametric tests are compared when the parametric assumptions are assumed. It is the author's opinion that Lee could have made a good case for using $C_1^*$ rather than $C$ based purely on the underlying assumptions.

This procedure was formulated with respect to a two group alternative. It is easy to see the generalization to the case where under the alternative there are k clusters with different means and k-1 mixing parameters $p_1, p_2, \ldots, p_{k-1}$, where $2 \leq k \leq n-1$. Without displaying the algebra, the test for the k group alternative would be:

Reject $H_0$ in favor of $H_A$ whenever

$$\underset{H}{MAX}\left[\binom{n}{n_1,n_2,\ldots,n_k}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\cdots\left(\frac{n_k}{n}\right)^{n_k}\right]^{2/n} \frac{B}{W} > K_\alpha \quad .$$

It is unfortunate that no tables of percentage points exist for any of these tests, because they seem to be better in terms of assumptions than the Engelman and Hartigan type tests. It is unlikely that percentage points will be generated because of the great expense in examining all the partitions required.

One of the advantages that the likelihood ratio tests have is that the tests for clusters do not depend on any measures of association or distance computed between units. One of the disadvantages is that many partitions have to be examined, which is costly in terms of computer time to the user and also to the researcher trying to generate empirical results. The next section will review problems in formulating a test which is dependent on the measure of distance.

## Some Procedures Based on the Dendogram

Suppose a sequential agglomerative hierarchal algorithm is used to form clusters at each of $n-1$ stages. Suppose that Squared Euclidean distance is used as a measure of distance. As mentioned previously $n-1$ distance matrices are computed in order to "cluster" the data. Let $M_1$ be the first distance matrix computed, $M_2$ the second, and so on with $M_{n-1}$ being the last matrix (2 x 2) computed. Let $f_1$ be the minimum element above the main diagonal of $M_1$, $f_2$ the minimum element above the main diagonal of $M_2$ and so on with $f_{n-1}$ the minimum element above the main diagonal of $M_{n-1}$. Can the observed values of the f's be used to determine whether all the original observations were sampled from the same normal population? A simple bivariate example will be used to illustrate the difficulties with this type of approach.

Let $x_i$, $i = 1, 2, 3$ be randomly sampled from a bivariate normal distribution with parameters $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, $\rho = 0$.

$$f_1 = \text{minimum } (d_{12}, d_{13}, d_{23}) \qquad \text{where}$$

$$d_{12} = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2$$

$$d_{13} = (x_{11} - x_{31})^2 + (x_{12} - x_{32})^2$$

$$d_{23} = (x_{21} - x_{31})^2 + (x_{22} - x_{32})^2$$

and $x_{ij}$ are components of $x_i$.

Can the distribution of $f_1$ be found in some useful form? First the distribution of $d_{12}$ will be found. Let

$$X_{ij} \sim N(0, 1) \qquad \text{for } i = 1, 2, 3 \quad j = 1, 2.$$

$X_{(i,j)}$ and $X_{(m,n)}$ are independent whenever $(i,j) \neq (m,n)$.

Hence $$X_{11} - X_{21} \sim N(0, 2),$$

$$X_{12} - X_{22} \sim N(0, 2)$$

and thus

$$\frac{X_{11} - X_{21}}{\sqrt{2}} \sim N(0, 1)$$

and

$$\frac{X_{12} - X_{22}}{\sqrt{2}} \sim N(0, 1).$$

Thus

$$\left(\frac{X_{11} - X_{21}}{\sqrt{2}}\right)^2 \sim \chi^2(1)$$

and

$$\left(\frac{X_{12} - X_{22}}{\sqrt{2}}\right)^2 \sim \chi^2(1).$$

The sum of two independent Chi-square variables with $n_1$ and $n_2$ degrees of freedom is distributed as a Chi-square variable with $n_1 + n_2$ degrees of freedom, so since

$$d_{12} = (X_{11} - X_{21})^2 + (X_{12} - X_{22})^2 \quad .$$

Hence,

$$\frac{d_{12}}{2} \sim \chi^2(2)$$

and we conclude that $d_{12} \sim 2\chi^2(2)$.

From this result it follows immediately that $d_{13}$ is distributed as $2\chi^2(2)$ and also $d_{23}$ is distributed as $2\chi^2(2)$.

These three distances are not independent since

$$\begin{aligned}\mathrm{COV}(d_{12},d_{13}) &= \mathrm{COV}[(X_{11}-X_{21})^2+(X_{12}-X_{22})^2,\ (X_{11}-X_{31})^2\\ &\quad +(X_{12}-X_{32})^2] = \mathrm{COV}[(X_{11}-X_{21})^2,\ (X_{11}-X_{31})^2]\\ &\quad +\mathrm{COV}[(X_{12}-X_{22})^2,\ (X_{12}-X_{32})^2]\\ &= \mathrm{COV}(X_{11}^2,\ X_{11}^2)+\mathrm{COV}\,(X_{12}^2,\ X_{12}^2)\\ &= \mathrm{VAR}(X_{11}^2)+\mathrm{VAR}(X_{12}^2) = 4.\end{aligned}$$

Hence,

$$\mathrm{CORR}(d_{12},\ d_{13}) = \frac{4}{\sqrt{16\cdot 16}} = .25.$$

Now,

$$f_1 = \mathrm{Min}\,(d_{12},\ d_{13},\ d_{23}),$$

and so the cumulative distribution function for $f_1$ is given by:

$$\begin{aligned}F(t) &= \Pr(d_{12}<t \text{ or } d_{13}<t \text{ or } d_{23}<t)\\ &= 1-\Pr(d_{12}>t,\ d_{13}>t,\ d_{23}>t).\end{aligned}$$

$$\Pr(d_{12}>t,\ d_{13}>t,\ d_{23}>t) = \Pr(\sqrt{d_{12}}>\sqrt{t},\ \sqrt{d_{13}}>\sqrt{t},\ \sqrt{d_{23}}>\sqrt{t}).$$

$$\begin{aligned}&=\int\int\int\int \Pr(\sqrt{d_{12}}>\sqrt{t},\ \sqrt{d_{13}}>\sqrt{t},\sqrt{d_{23}}>\sqrt{t}\ \big|\ X_{11}=x_{11},\ X_{12}=x_{12},\ \ldots,\\ &\qquad X_{21}=x_{21},\ X_{22}=x_{22})\cdot\phi(x_{11})\phi(x_{12})\phi(x_{21})\phi(x_{22})\\ &\qquad dx_{22}\,dx_{21}\,dx_{12}\,dx_{11}\end{aligned}$$

$$\begin{aligned}&=\int\int\int\int_{R_1} \Pr(\sqrt{d_{12}}>\sqrt{t},\ \sqrt{d_{13}}>\sqrt{t},\sqrt{d_{23}}>\sqrt{t}\ \big|\ X_{11}=x_{11},\ X_{12}=x_{12},\\ &\qquad X_{21}=x_{21},\ \ldots,\ X_{22}=x_{22})\ \phi(x_{11})\phi(x_{12})\phi(x_{21})\phi(x_{22})\\ &\qquad dx_{22}\,dx_{21}\,dx_{12}\,dx_{11}+\end{aligned}$$

$$\iiiint_{R_2} \Pr\left[\sqrt{d_{12}} > \sqrt{t}, \sqrt{d_{13}} > \sqrt{t}, \sqrt{d_{23}} > \sqrt{t} \,\middle|\, X_{11} = x_{11}, X_{12} = x_{12}, X_{21} = x_{21}, X_{22} = x_{22}\right] \phi(x_{11})\phi(x_{12})\phi(x_{21})\phi(x_{22})\, dx_{22}dx_{21}dx_{12}dx_{11}$$

where $R_1$ = Region where $\sqrt{(X_{11} - X_{21})^2+(X_{12} - X_{22})^2} \geq t$,

$R_2$ = Region where $\sqrt{(X_{11} - X_{21})^2+(X_{12} - X_{22})^2} < t$,

and

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad .$$

Thus,

$$F(t) = \iiiint P(t|2, r_1)\phi(x_{11})\phi(x_{22})\phi(x_{12})\phi(x_{21})\, dx_{11}dx_{22}dx_{12}dx_{21}$$

$$+ \iiiint P(t|2, r_2)\phi(x_{11})\phi(x_{22})\phi(x_{12})\phi(x_{21})\, dx_{11}dx_{22}dx_{12}dx_{21}$$

$$+ \iiiint \Pr\left[\sqrt{d_{12}} > \sqrt{t}, \sqrt{d_{13}} > \sqrt{t}, \sqrt{d_{23}} > \sqrt{t} \,\middle|\, X_{11} = x_{11}, X_{12} = x_{12}, X_{21} = x_{21}, \ldots, X_{22} = x_{22}\right] \phi(x_{11})\phi(x_{12})\phi(x_{21})\phi(x_{22})$$

$$dx_{22}dx_{21}dx_{12}dx_{11}$$

where

$P(t|2, r)$ is the cumulative distribution function of the non-central $\chi^2$ distribution with 2 degrees of freedom and non-centrality parameter, $r$, $r_1 = x_{11}^2 + x_{12}^2$, $r_2 = x_{21}^2 + x_{22}^2$.

The conditional probability in the last integral can be found by integrating over two intersecting circles in the plane with respect to a bivariate normal distribution.

The region where the two circles intersect is the worst problem. One approximation would be to assume the region where the two circles intersected was an ellipse and follow procedures set forth by Snow and Ryan (1970) for its numerical evaluation.

Numerical integration of multiple integrals is often very costly and is subject to many rounding errors during computation. The problem is further complicated by the fact that the cumulative distribution function of the non-central Chi-square distribution is not known in closed form. It is the author's conclusion that numerical evaluation of the above expressions is not feasible, especially since the above example is perhaps the most simple case of clustering that can be considered.

Another approach to this problem is to assume that $d_{12}$, $d_{13}$, $d_{23}$ are independent in calculating the distribution of $f_1$. It was previously demonstrated that this is not true. The result is to be compared with an empirical distribution of $f_1$ based on a large number of observations. The result may indicate that the assumption of independence is tolerable.

It was previously demonstrated that $d_{12} \sim 2\chi^2(2)$. The $\chi^2(2)$ density is:

$$f(x) = \begin{cases} (1/2\ e^{-x/2}) & \text{where } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

which is the exponential density with parameter 1/2. Then by a simple change of variable the density of $d_{12}$ is:

$$f_{12}(y) = \begin{cases} 1/4\ e^{-y/4} & \text{where } 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

which is exponential with parameter $1/4$.

The cumulative distribution function is then:

$$F_{12}(y) = \int_0^y 1/4\ e^{-t/4}\ dt = 1 - e^{-y/4} \quad .$$

The density of the minimum of three independent variables each having this same density is given by:

$$f_{min}(y) = \begin{cases} \frac{3!}{0!2!} [F(y)]^0 [1-F(y)]^2 f(y) & \text{for } 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} (3/4)e^{-3y/4} & \text{for } 0 < y < \infty \\ 0 & \text{otherwise} \quad . \end{cases}$$

This is exponential with parameter $3/4$.

10,000 values of $f_1$ were generated, and an empirical frequency distribution was computed (see Table I). An estimate of the density function was found for the midpoint of each class interval. This estimate $y$ is given by $\frac{\text{observed \% frequency of the class}}{\text{length of the class interval}}$. If it is assumed that $f_1$ has an exponential density with parameter $\lambda$, then $f_1(x) = \lambda\ e^{-\lambda x}$ for $x > 0$. It follows that $\log(y)$ and $x$ would be approximately linearly related, and a least squares procedure might be useful in estimating the parameters of the line. However, for some of the larger values of $x$, the observed frequency is zero, hence the logarithm cannot be taken. An arbitrary decision was made to disregard all the class intervals greater than 10.1, and to estimate $\lambda$ from only the slope of the regression line. The result is a somewhat crude estimate of the density of $f_1$. The estimate

TABLE I

EMPIRICAL FREQUENCY DISTRIBUTION FOR
$f_1$, n = 3

| Class Interval | Frequency | Class Interval | Frequency |
|---|---|---|---|
| 0 - .20 | 1363 | 5.00 - 5.20 | 32 |
| .20 - .40 | 1168 | 5.20 - 5.40 | 25 |
| .40 - .60 | 1013 | 5.40 - 5.60 | 29 |
| .60 - .80 | 902 | 5.60 - 5.80 | 25 |
| .80 - 1.00 | 756 | 5.80 - 6.00 | 29 |
| 1.00 - 1.20 | 651 | 5.00 - 6.20 | 18 |
| 1.20 - 1.40 | 528 | 6.20 - 6.40 | 26 |
| 1.40 - 1.60 | 484 | 6.40 - 6.60 | 10 |
| 1.60 - 1.80 | 415 | 6.60 - 6.80 | 17 |
| 1.80 - 2.00 | 358 | 6.80 - 7.00 | 7 |
| 2.00 - 2.20 | 321 | 7.00 - 7.20 | 8 |
| 2.20 - 2.40 | 245 | 7.20 - 7.40 | 10 |
| 2.40 - 2.60 | 198 | 7.40 - 7.60 | 10 |
| 2.60 - 2.80 | 196 | 7.60 - 7.80 | 9 |
| 2.80 - 3.00 | 186 | 7.80 - 8.00 | 5 |
| 3.00 - 3.20 | 165 | 8.00 - 8.20 | 7 |
| 3.20 - 3.40 | 149 | 8.20 - 8.40 | 5 |
| 3.40 - 3.60 | 119 | 8.40 - 8.60 | 4 |
| 3.60 - 3.80 | 109 | 8.60 - 8.80 | 13 |
| 3.80 - 4.00 | 72 | 8.80 - 9.00 | 6 |
| 4.00 - 4.20 | 72 | 9.00 - 9.20 | 2 |
| 4.20 - 4.40 | 72 | 9.20 - 9.40 | 3 |
| 4.40 - 4.60 | 45 | 9.40 - 9.60 | 1 |
| 4.60 - 4.80 | 52 | 9.60 - 9.80 | 3 |
| 4.80 - 5.00 | 35 | 9.80 -10.00 | 2 |
| | | 10.00 -10.20 | 2 |
| | | 10.20 - $+\infty$ | 18 |

of the density is: $f_1(x) \doteq .66\, e^{.66x} \quad x > 0.$

Suppose that instead of a sample of size 3 from the bivariate normal distribution previously mentioned, a sample of ten points is available. Since some of the 45 distances are pairwise independent, will the distribution of $f_1$ assuming all the distances are independent be any "better" than the previous case where all the distances were pairwise dependent?

Assuming all 45 distances are independent, the density of $f_1$ would be:

$$f_1(y) = \begin{cases} 45[F(y)]^0\,[1-F(y)]^{44}\, f(y) & \text{for } 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} (45/4)e^{-45y/4} & \text{for } 0 < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

10,000 values of $f_1$ were generated, and an empirical frequency distribution was computed (see Table II). Following the same procedure as in the previous example, the estimate of the density was found to be:

$$f_1(x) \doteq \begin{cases} 10.99\, e^{-10.99x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

It is to be noted that both of the estimated densities are based on the assumption that $f_1$ has an exponential distribution. It would be difficult to objectively evaluate the approximation of the distribution of $f_1$, based on the assumption of independence of the distances, even if the true distribution of $f_1$ was known. The author will not attempt to do so, since it is a subjective judgment.

TABLE II

EMPIRICAL FREQUENCY DISTRIBUTION FOR
$f_1$, n = 10

| Class Interval | Frequency | Class Interval | Frequency |
|---|---|---|---|
| 0 - .01 | 1053 | .26 - .27 | 55 |
| .01 - .02 | 924 | .27 - .28 | 54 |
| .02 - .03 | 852 | .28 - .29 | 51 |
| .03 - .04 | 728 | .29 - .30 | 46 |
| .04 - .05 | 677 | .30 - .31 | 42 |
| .05 - .06 | 578 | .31 - .32 | 29 |
| .06 - .07 | 547 | .32 - .33 | 32 |
| .07 - .08 | 477 | .33 - .34 | 25 |
| .08 - .09 | 489 | .34 - .35 | 29 |
| .09 - .10 | 353 | .35 - .36 | 26 |
| .10 - .11 | 368 | .36 - .37 | 19 |
| .11 - .12 | 314 | .37 - .38 | 23 |
| .12 - .13 | 259 | .38 - .39 | 11 |
| .13 - .14 | 243 | .39 - .40 | 19 |
| .14 - .15 | 211 | .40 - .41 | 14 |
| .15 - .16 | 207 | .41 - .42 | 11 |
| .16 - .17 | 168 | .42 - .43 | 11 |
| .17 - .18 | 167 | .43 - .44 | 9 |
| .18 - .19 | 146 | .44 - .45 | 6 |
| .19 - .20 | 118 | .45 - .46 | 8 |
| .20 - .21 | 101 | .46 - .47 | 5 |
| .21 - .22 | 97 | .47 - .48 | 5 |
| .22 - .23 | 103 | .48 - .49 | 5 |
| .23 - .24 | 94 | .49 - .50 | 10 |
| .24 - .25 | 76 | .50 - .51 | 3 |
| .25 - .26 | 60 | .51 - $+\infty$ | 42 |

The distribution of $f_2$ is conditional on the result at stage 1, and is therefore much more complex than that of $f_1$. Besides the complexity of the mathematics, the author believes that this type of approach will not lead to many useful results for reasons to be mentioned at the end of this section.

It has been worthwhile studying this type of approach because the theoretical difficulties are illustrated in a simplified form, which allows one to avoid difficulties such as those encountered by Rohlf (1975). Rohlf (1975) considers a procedure which computes $g_i = f_i - f_{i-1}$, (where $f_0$ is defined as 0). The algorithm used is the single linkage, sequential agglomerative hierarchal procedure with squared Euclidean distance as the distance measure. The Minimum Spanning Tree (MST), is a minimally connected graph with no circuits (Gower and Ross, 1969). Rohlf (1975) bases several theoretical derivations on the false assumption that the edges in the MST are independent (see Norton and Warde, 1975). Judgment on this study must be reserved until the empirical results Rohlf used as a basis for his conclusions are examined.

There are two major difficulties that the author sees with these dendogram-based procedures; first, the distributional results will depend on the measure of similarity or distance between units that the algorithm is using; and secondly, there is no indication that the procedures would be independent of $\sigma^2$. It does not seem reasonable that $\sigma^2$ be assumed known in clustering problems. It is also to be noted that standardization of a variable by its estimated standard deviation is not justification for assuming $\sigma^2 = 1$.

## Formulation of the Clustering Problem as a Mixture of Normal Distributions

In addition to Lee (1974), other authors, such as Day (1969), have considered formulating the clustering problem as a mixture of normal distributions. There are so many difficulties still present in solving the normal mixtures problem that solution of the clustering problem has hardly been more than mentioned, with the exception of Lee (1974).

A formulation of the problem in a manner in which there is a reasonable chance for solution is:

Let $x_1, \ldots, x_n$ be a random sample from a population $f$ where

$$f = \sum_{i=1}^{k} P_i f_i$$

with $f_i \sim N(\mu_i, \sigma^2)$, $P_1, P_2, \ldots, P_k$ are mixing parameters $0 \leq P_i \leq 1$, $i = 1, \ldots, k$, $\sum_{i=1}^{k} P_i = 1$, and $k$ is known $2 \leq k \leq n/2$.

The procedures that are used most often to find point estimators of the unknown parameters are the method of moments (which goes back to Pearson, 1894) and the method of maximum likelihood. In the clustering context, it would be desirable to set confidence intervals on the $P_i$'s. In order to set confidence intervals on the $P_i$, $i = 1, \ldots, k$, the sampling distributions of the estimates $\hat{P}_i$, $i = 1, \ldots, k$, are needed, and they are not presently available. To further complicate the issue, the maximum likelihood estimators are usually found by iterative procedures, which occasionally converge to an incorrect value.

The author's conclusion is that the formulation of the clustering problem in terms of mixtures of normal distributions is most tractable using the Lee (1974) approach. Further work on this approach may lead to some advances in the decomposition of normal mixtures.

## Other Miscellaneous Test Procedures

Ling (1971) selects a single linkage sequential agglomerative hierarchal algorithm as a procedure that will yield "exact" distributional results. A matrix of pairwise distances is assumed to be available, but the actual distances above the main diagonal are replaced with the ranks of these distances. The single linkage algorithm is then used in the usual way on this matrix of ranks, to cluster the data. Ling (1971) tries to formulate a null hypothesis of "randomness". Based on this null hypothesis, he finds exact probability distributions for the number of units belonging to clusters of 2 or more at each stage of the algorithm..

Assuming there are $n$ responses to be clustered, then there are $n(n-1)/2$ ranks above the main diagonal of the distance matrix. Ling (1971) assumes that all $[n(n-1)/2]!$ permutations are equally likely. He points out that the assumption of equally likely permutations of ranks makes the mathematics possible to work with, but admits that this assumption is impossible for some lower dimensional spaces. He avoids mentioning that the problems arise when the number of units to be clustered is larger than the number of response variables measured on each unit. The most important and useful results from clustering data are obtained when there are more units than response variables.

Example: Suppose four univariate observations are collected and the six pairwise distances are computed, then the following relative ordering is impossible:

$$d_{12} < d_{24} < d_{23} < d_{34} < d_{13} < d_{14} \quad .$$

The present author has no confidence in results based on this assumption of equally likely permutations of ranks, regardless of how well the mathematics works. In many cases, the null hypothesis could be rejected without bothering to look at the data.

Mrachek (1972) proposed two tests. The first test assumed at the outset that

$$E(X_i) = \mu_i$$

and

$$V(X_i) = \sigma^2 I \quad .$$

It is not stated that $\sigma^2$ is known, but later computation of the test statistic seems to require $\sigma^2$ to be known. Let $D' = (d_{12}, d_{13}, \ldots, d_{1n}, d_{23}, \ldots, d_{2n}, \ldots, d_{(n-1)n})$. Some argument is given to justify assuming $D$ has approximately a multivariate normal distribution.

In the type of situations where cluster analysis is most appropriate it seems unreasonable to assume that $\sigma^2$ will be known. Mrachek (1972) makes the comment that if $\sigma^2$ is not known then the data matrix should be standardized and $\sigma^2$ set equal to one in the computation of the test statistic. Standardizing variables by their estimated standard deviation does not justify assuming $\sigma^2 = 1$, although several authors in the literature seem to believe that this is possible. If the data matrix is standardized by estimated standard deviations the distribution of $D$ will be affected, but no mention

is made of this problem. It is also to be noted that Mrachek's (1972) arguments concerning the normality of D are based on allowing the number of components of the response vector to become large. This seems to make the test quite questionable when there are a relatively small number of components in the response vector. For application purposes, this comment seems especially important in view of Mrachek's conclusion that the addition of uninformative variables hinders the performance of clustering procedures.

The second test is based on Jacknife estimators and it also assumes $\sigma^2$ is known. The test statistic which is computed has a Hotelling's $T^2$ distribution. This is based on a conjecture by Tukey (1962) that individual Jacknife estimators of distances are approximately normally distributed. From this conjecture Mrachek (1972) concludes that the vector of distances is approximately normally distributed. This appears to be rather weak "evidence" on which to base a test.

In the next chapter, some test procedures based on empirical results will be suggested as practical alternatives to the procedures previously mentioned in this chapter.

# CHAPTER III

## A MONTE CARLO INVESTIGATION OF FOUR MAJOR AGGLOMERATIVE CLUSTERING PROCEDURES

### Justification for the Selection of These Four Procedures

There is at least some theoretical justification for using the divisive algorithm which partitions the data into clusters using the sum of squares criterion discussed previously, with regard to the Engelman and Hartigan (1969) test. Since the divisive algorithms require a relatively large amount of computing time, they are not extensively used in applications of cluster analysis. The author sought an algorithm that would approximate the partitioning given by the aforementioned divisive algorithm, which would be extensively used at present and hence readily available to potential users. An algorithm which is relatively inexpensive to use was also sought because more cases could be considered by empirical methods. Since the sum of squared deviations of a set of observations about their mean is smaller than the sum of squared deviations about any other number, it was intuitively reasoned that the weighted average algorithm or centroid algorithm would closely approximate the partitioning given by the divisive algorithm. In order to be able to make evaluations of some of the algorithms not specifically considered, it was decided to include the two algorithms, which are considered as "boundary" algorithms.

More discussion of "boundary" type algorithms will be given in Chapter IV. After the above choices of algorithms had been made and many of the results collected, Baker and Hubert (1975) published an article which evaluated the power of some of these same algorithms, but with respect to a different class of null and alternative hypotheses. This makes the above choice of algorithms all the more appropriate for comparison purposes.

## Statement of the Proposed Statistical Procedures

In order to test for a known number of clusters, k, where $2 \leq k \leq n-1$, the user may select one of the four algorithms and cluster the data. At stage $n - k + 1$, the algorithm has assigned the $n$ observations to $k$ clusters. The ratio, between cluster sum of squares divided by within cluster sum of squares (B/W), must be computed and referred to the appropriate table of percentage points.

There is at present a practical problem using the above procedure, namely, tables of percentage points have been generated for only a very limited number of sample sizes. In order to have some test available until more tables can be generated, a conservative test for the two cluster alternative can be devised based on the presently tabled Engelman and Hartigan (1969) percentage points. To perform this conservative test, the user selects one of the four agglomerative algorithms and as before clusters the data. At stage $n-1$, the algorithm has assigned the observations to two clusters. The ratio B/W is computed and referred to the table of percentage points of Engelman and Hartigan (1969). The test is conservative because the percentage points

of the Engelman and Hartigan (1969) test are based on the computation of the maximum B/W, and there is no guarantee that the assignment of observations to clusters by the agglomerative procedures will lead to the maximum B/W.

Selected percentage points of the null distributions are given in Tables III through XVI. For notational purposes a * to the right of an entry will indicate that the entry is theoretically too large, making the test conservative. If an entry is larger than 4,000,000 it will be replaced by ** in the table, and if an entry is larger than 8,000,000 it will be replaced by *** in the table. E-H will represent the Engelman and Hartigan percentage points.

## Approximate Confidence Intervals for the Estimated Percentage Points

Let $F_{B/W}(x)$ be the cumulative distribution function of an unknown distribution. Let $C_\alpha$ be the point such that $F_{B/W}(C_\alpha) = \alpha$. The problem in general is to take a sample from this distribution, and to find an approximate confidence interval for $C_\alpha$.

More specifically, 1000 observations were selected at random from $F_{B/W}$, and it is desired to find approximate 95% confidence intervals for $C_{.50}, C_{.75}, C_{.90}, C_{.95}, C_{.99}$.

Consider the problem of finding a confidence interval for $C_{.50}$. A point estimate of $C_{.50}$ is given by the value of the 500th order statistic from the sample of 1000. The density of the 500th order statistic from a random sample of size 1000 is given by:

$$f_{B/W}(x) = \begin{cases} \frac{1000!}{(499!)(500!)} [F_{B/W}(x)]^{499} [1 - F_{B/W}(x)]^{500} f_{B/W}(x) & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

TABLE III

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE TWO GROUP ALTERNATIVE
WITH n = 6

| | E-H | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | 3.88 | 3.525 | 3.85 | 3.85 | 3.85 |
| .75 | 5.96 | 5.95 | 5.925 | 5.95 | 5.925 |
| .90 | 9.84 | 9.30 | 9.30 | 9.30 | 9.20 |
| .95 | 14.1 | 13.825 | 13.825 | 13.825 | 13.55 |
| .99 | 33.1 | 30.8 | 30.8 | 30.8 | 28.225 |

TABLE IV

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE THREE GROUP ALTERNATIVE
WITH n = 6

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 19.267 | 21.0 | 20.867 | 20.867 |
| .75 | 36.80 | 37.6 | 37.600 | 37.600 |
| .90 | 71.60 | 72.8 | 72.800 | 72.800 |
| .95 | 118.00 | 118.0 | 118.000 | 118.000 |
| .99 | 418.40 | 418.4 | 418.400 | 418.400 |

TABLE V

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE FOUR GROUP ALTERNATIVE
WITH n = 6

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 110.40 | 111.15 | 111.15 | 111.15 |
| .75 | 234.75 | 234.75 | 234.75 | 234.75 |
| .90 | 530.55 | 537.90 | 537.90 | 537.90 |
| .95 | 1485.15 | 1485.15 | 1485.15 | 1485.15 |
| .99 | 4006.95 | 4006.95 | 4006.95 | 4006.95 |

TABLE VI

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE FIVE GROUP ALTERNATIVE
WITH n = 6

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 1132.4 | 1132.4 | 1132.4 | 1132.4 |
| .75 | 5461.6 | 5461.6 | 5461.6 | 5461.6 |
| .90 | 38795.6 | 38795.6 | 38795.6 | 38795.6 |
| .95 | 215374.8 | 215374.8 | 215374.8 | 215374.8 |
| .99 | ** | ** | ** | ** |

** > 4,000,000

TABLE VII

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE TWO GROUP ALTERNATIVE
WITH n = 10, RUN ONE

| | E-H | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | 2.76 | 1.5625 | 2.525 | 2.5625 | 2.5625 |
| .75 | 3.76 | 3.0625 | 3.525 | 3.55 | 3.5625 |
| .90 | 5.14 | 5.30* | 5.3125* | 5.40* | 5.40* |
| .95 | 6.34 | 6.575* | 6.575* | 6.575* | 6.575* |
| .99 | 9.89 | 10.2875* | 10.2875* | 10.2875* | 10.2875* |

TABLE VIII

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE TWO GROUP ALTERNATIVE
WITH n = 10, RUN TWO

| | E-H | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | 2.76 | 1.5625 | 2.525 | 2.475 | 2.475 |
| .75 | 3.76 | 3.075 | 3.6375 | 3.6375 | 3.6125 |
| .90 | 5.14 | 4.7875 | 4.8125 | 4.8125 | 4.8125 |
| .95 | 6.34 | 5.775 | 5.775 | 5.775 | 5.775 |
| .99 | 9.89 | 10.650 | 10.650* | 10.650* | 10.650* |

TABLE IX

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE TWO-GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | E-H | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | 2.76 | 1.5625 | 2.525 | 2.525 | 2.5125 |
| .75 | 3.76 | 3.0750 | 3.5625 | 3.6125 | 3.6125 |
| .90 | 5.14 | 5.0125 | 5.0500 | 5.0625 | 5.0625 |
| .95 | 6.34 | 6.2250 | 6.1875 | 6.2250 | 6.2250 |
| .99 | 9.89 | 10.650* | 10.650* | 10.650* | 10.650* |

TABLE X

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE THREE GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 7.343 | 9.800 | 9.771 | 9.686 |
| .75 | 12.657 | 14.143 | 13.857 | 13.857 |
| .90 | 18.857 | 19.600 | 19.571 | 19.514 |
| .95 | 25.086 | 25.943 | 26.257 | 26.257 |
| .99 | 41.40 | 41.40 | 41.400 | 41.400 |

TABLE XI

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE FOUR GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 23.3 | 28.35 | 28.35 | 28.15 |
| .75 | 37.2 | 41.45 | 41.15 | 40.85 |
| .90 | 58.75 | 64.00 | 62.65 | 62.30 |
| .95 | 78.4 | 83.95 | 82.60 | 82.55 |
| .99 | 136.8 | 151.05 | 152.05 | 151.05 |

TABLE XII

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE FIVE GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 65.04 | 74.80 | 73.20 | 72.80 |
| .75 | 108.72 | 122.24 | 120.40 | 120.40 |
| .90 | 180.16 | 187.44 | 187.44 | 186.56 |
| .95 | 234.24 | 255.60 | 244.32 | 244.32 |
| .99 | 431.68 | 431.68 | 431.68 | 435.16 |

TABLE XIII

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE SIX GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 189.0 | 209.0 | 208.5 | 208.5 |
| .75 | 334.135 | 348.125 | 344.5 | 344.125 |
| .90 | 626.875 | 642.25 | 637.75 | 644.750 |
| .95 | 869.500 | 883.25 | 883.250 | 883.250 |
| .99 | 1509.500 | 1509.500 | 1509.500 | 1509.50 |

TABLE XIV

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE SEVEN GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 562.8 | 582.4 | 585.4 | 585.4 |
| .75 | 1115.4 | 1176.2 | 1163.6 | 1163.6 |
| .90 | 2488.8 | 2509.4 | 2509.4 | 2509.4 |
| .95 | 4019.0 | 4019.0 | 4019.0 | 4019.0 |
| .99 | 10027.0 | 10027.0 | 10027.0 | 10027.0 |

TABLE XV

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE EIGHT GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 2310.35 | 2363.2 | 2359.7 | 2359.7 |
| .75 | 6737.50 | 6737.5 | 6737.5 | 6737.5 |
| .90 | 18572.40 | 18885.3 | 18885.3 | 18885.3 |
| .95 | 43214.50 | 43214.5 | 43214.5 | 43214.5 |
| .99 | 182917.0 | 182917.0 | 182917.0 | 182917.0 |

TABLE XVI

SELECTED PERCENTAGE POINTS OF THE B/W TESTS
FOR THE NINE GROUP ALTERNATIVE
WITH n = 10, COMBINED RUNS

| | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|
| $C_\alpha$ = .50 | 24152.8 | 24152.8 | 24152.8 | 24152.8 |
| .75 | 132056.8 | 132056.8 | 132056.8 | 132056.8 |
| .90 | 810454.4 | 810454.4 | 810454.4 | 810454.4 |
| .95 | 5380584.8 | 5380584.8 | 5380584.8 | 5380584.8 |
| .99 | *** | *** | *** | *** |

*** > 8,000,000

This expression contains the unknown distribution of $F_{B/W}$, hence an exact confidence interval cannot be found.

The following approximate procedure can be used:

1. Plot a smoothed empirical distribution function.
2. Regard the observed value of the 500th order statistic as a fixed constant, the true median of $F_{B/W}$.
3. Find a point estimate of the probability that a randomly chosen observation will be less than this observed value of the 500th order statistic (the point estimate is .5 of course, for this sample).
4. Find an approximate distribution for $\hat{P}$. In this case, it is assumed $\hat{P} \sim N(.5, \frac{(.5)(.5)}{1000})$. In general it is assumed $\hat{P} \sim N(P, \frac{P(1-P)}{N})$ where $n$ is the sample size.
5. Find a confidence interval on $P$. Using the distributional assumptions in (4) above, a 95% confidence interval on $P$ is given by:
$$(.5 - 1.96\sqrt{\frac{.25}{1000}}, \quad .5 + 1.96\sqrt{\frac{.25}{1000}}).$$
6. This interval is inverted graphically to approximate a 95% confidence interval about $C_{.5}$.

This procedure is illustrated and carried out in Figures 2 through 5. The results are collected in Table XVII.

## The Alternative Hypotheses Selected for this Study

The test procedures suggested in one of the previous sections could be used for detection of two different types of alternative. The first type of alternative is to consider the number of observations from each

TABLE XVII

APPROXIMATE CONFIDENCE INTERVALS ON $C_\alpha$ FOR THE B/W TESTS FOR THE TWO GROUP ALTERNATIVE WITH n = 10

| | |
|---|---|
| Single Linkage | $1.46 < C_{.50} < 1.69$ |
| | $2.89 < C_{.75} < 3.34$ |
| | $4.75 < C_{.90} < 5.29$ |
| | $5.74 < C_{.95} < 6.94$ |
| | $9.56 < C_{.99} < 13.13$ |
| Complete Linkage | $2.475 < C_{.50} < 2.63$ |
| | $3.34 < C_{.75} < 3.69$ |
| | $4.81 < C_{.90} < 5.50$ |
| | $6.00 < C_{.95} < 7.19$ |
| | $9.56 < C_{.99} < 13.13$ |
| Weighted Average Linkage | $2.40 < C_{.50} < 2.63$ |
| | $3.41 < C_{.75} < 3.68$ |
| | $4.73 < C_{.90} < 5.35$ |
| | $5.63 < C_{.95} < 6.85$ |
| | $9.56 < C_{.99} < 13.13$ |
| Centroid Linkage | $2.34 < C_{.50} < 2.59$ |
| | $3.44 < C_{.75} < 3.81$ |
| | $4.76 < C_{.90} < 5.35$ |
| | $5.96 < C_{.95} < 7.13$ |
| | $9.56 < C_{.99} < 13.13$ |

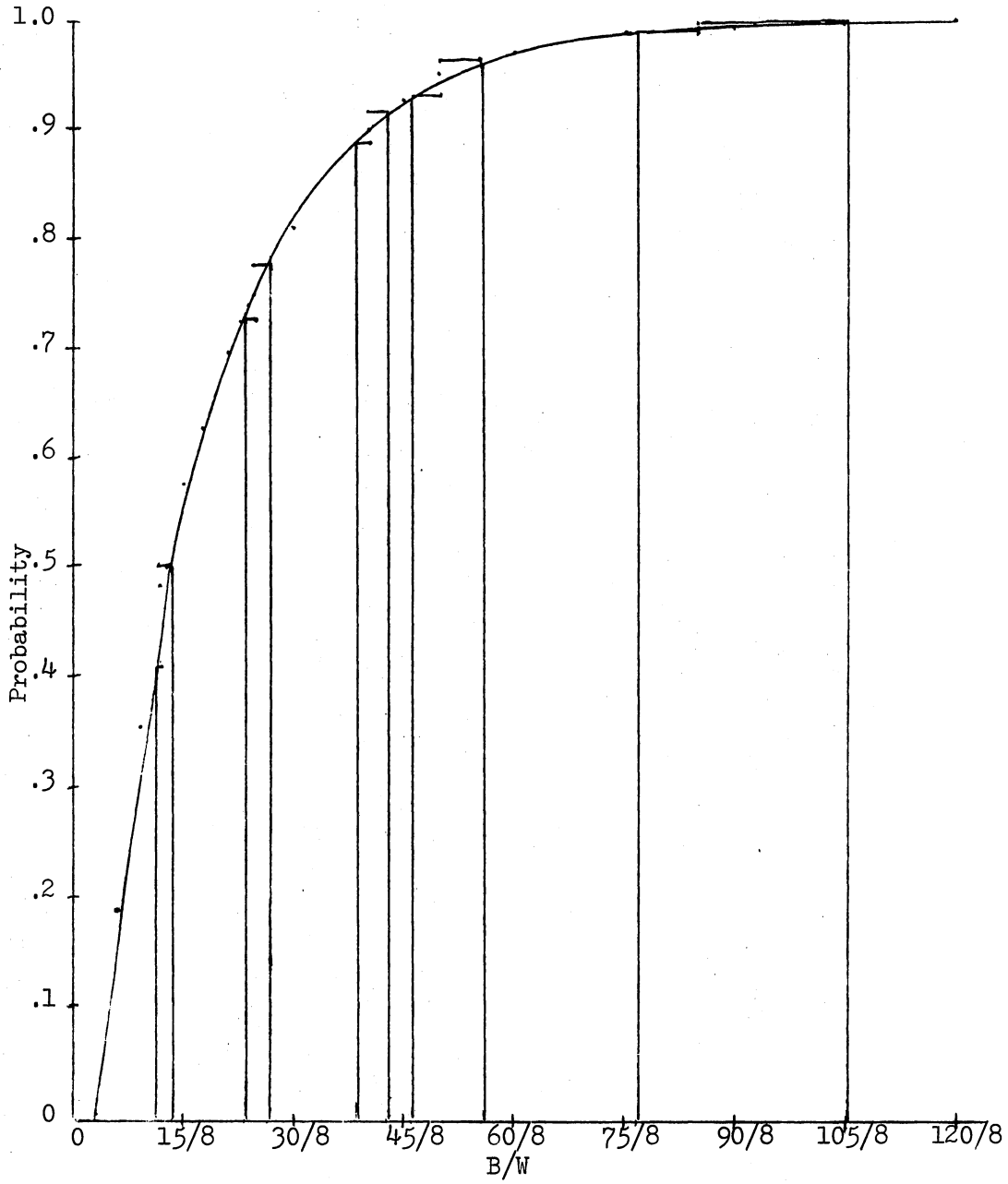


Figure 2. Approximate 95% Confidence Intervals for the $C_\alpha$'s Derived from the Single Linkage Test

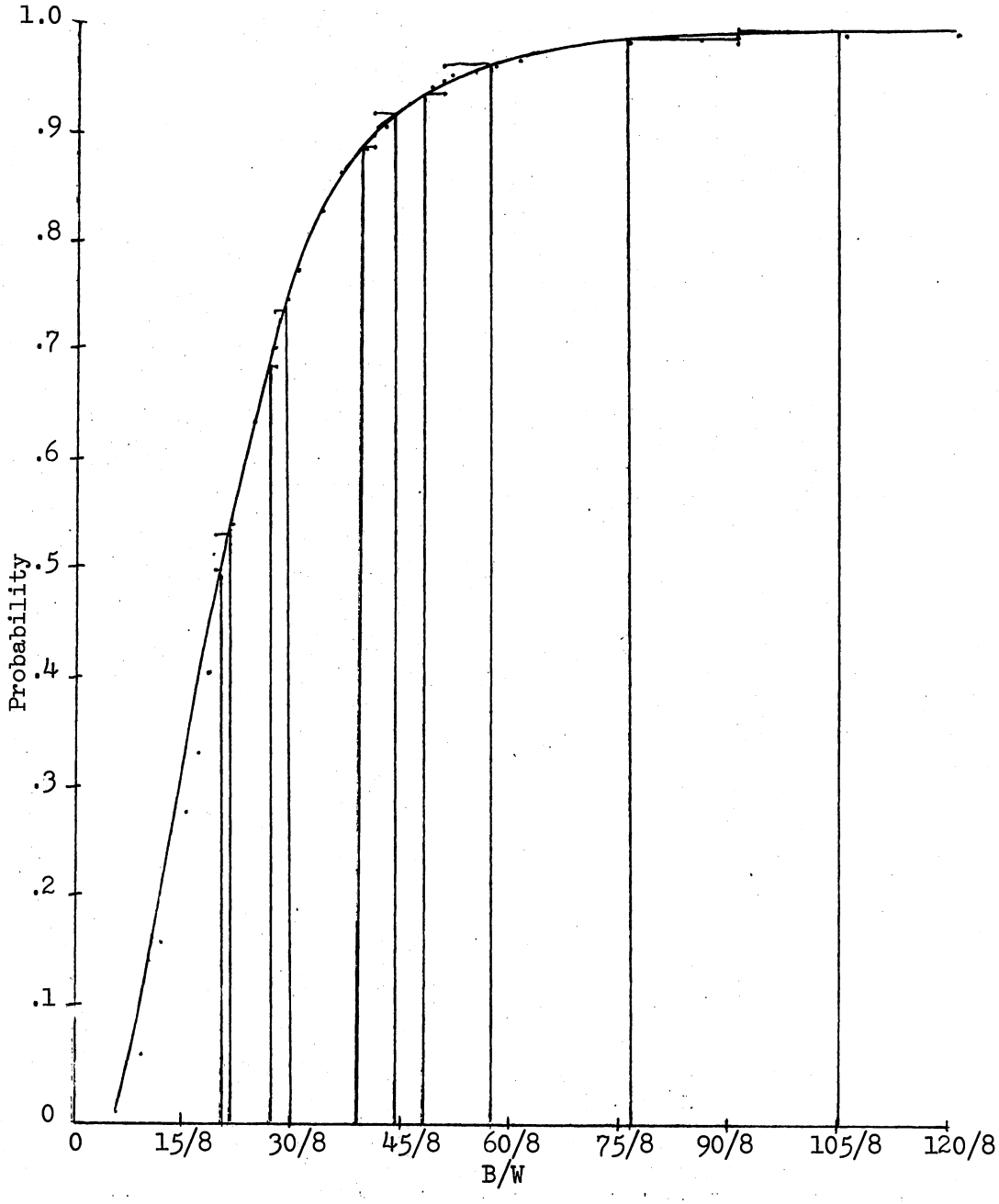


Figure 3. Approximate 95% Confidence Intervals for the $C_\alpha$'s Derived from the Complete Linkage Test

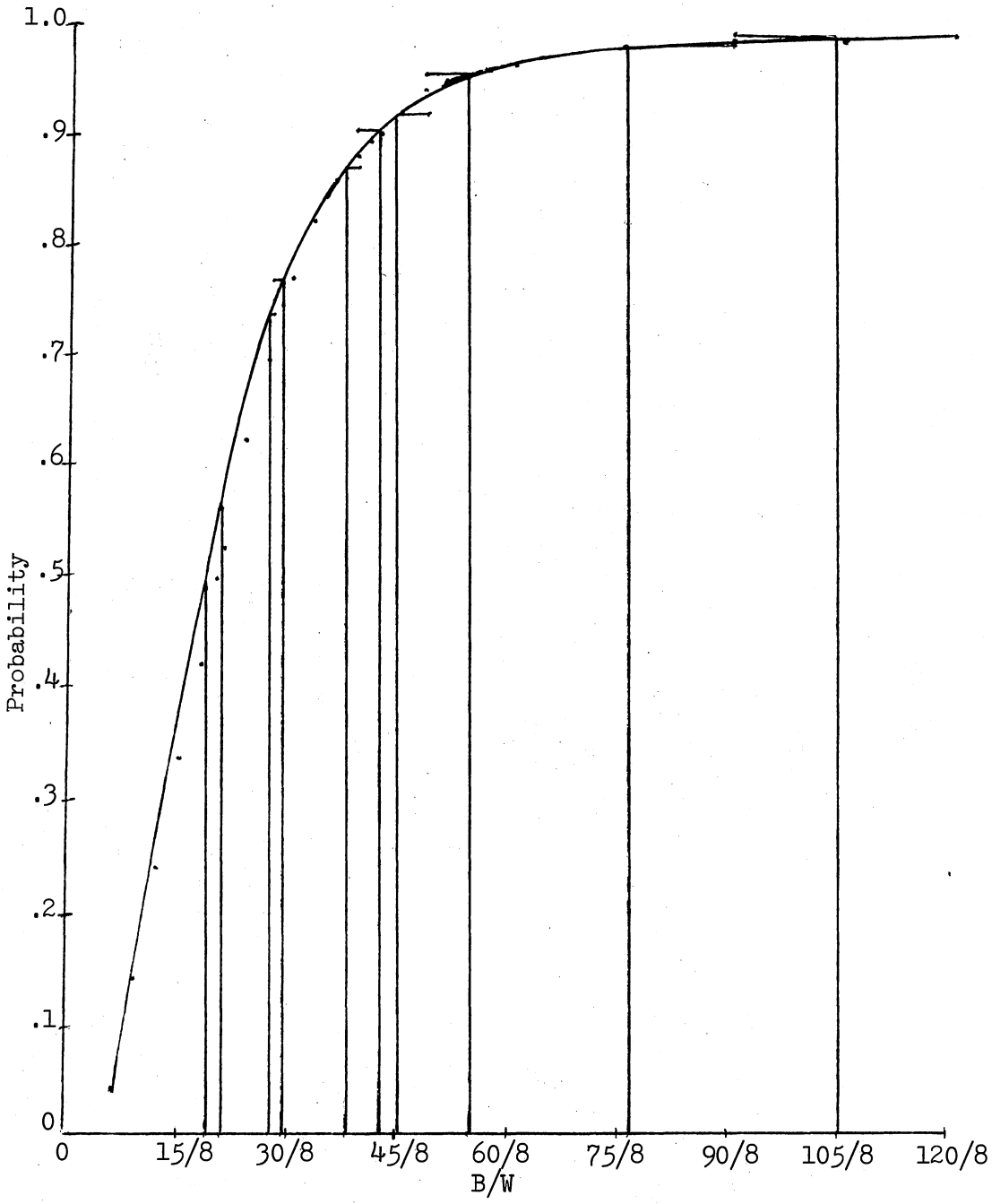


Figure 4. Approximate 95% Confidence Intervals for the $C_\alpha$'s Derived from the Weighted Average Linkage Test

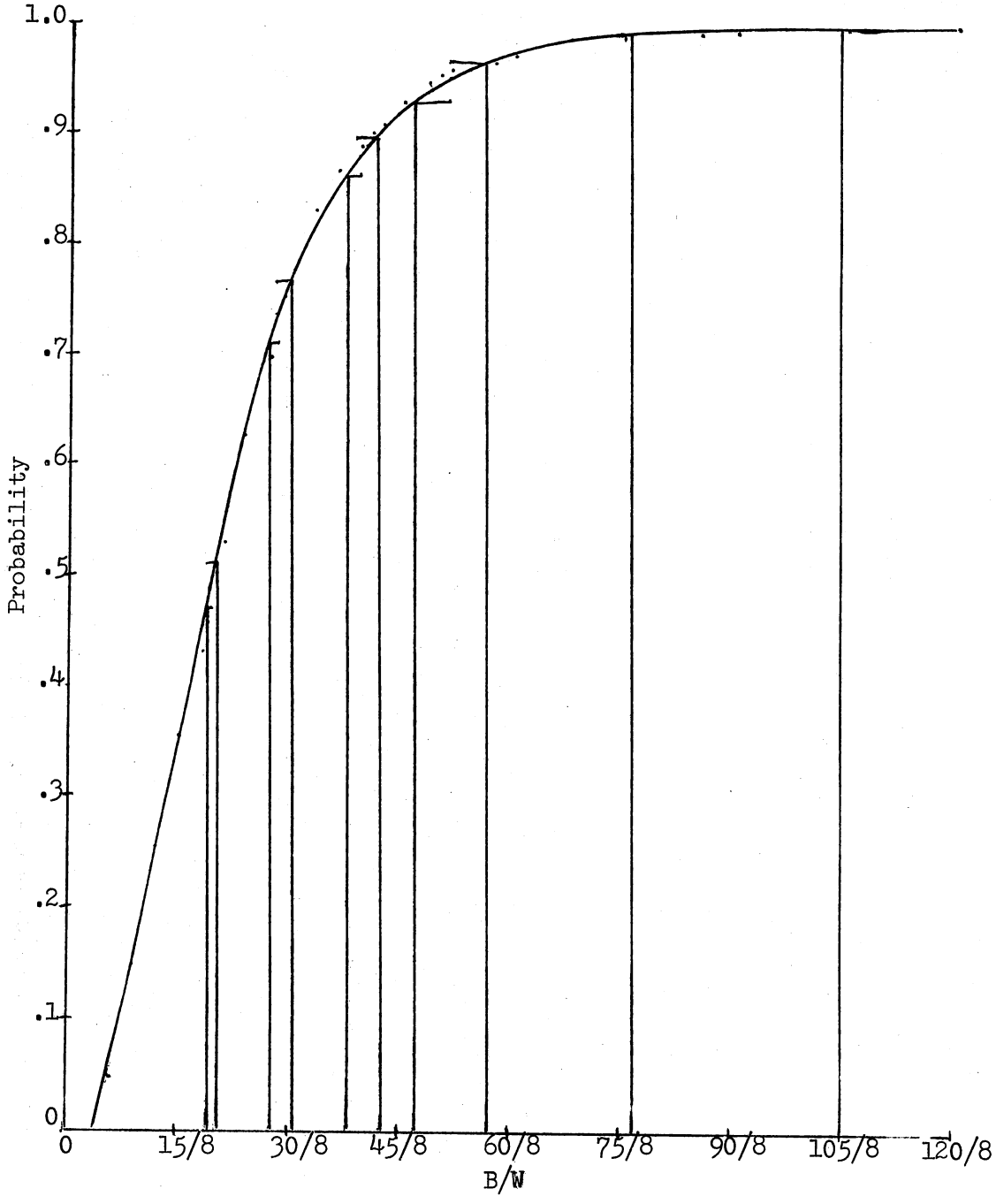


Figure 5. Approximate 95% Confidence Intervals for the $C_\alpha$'s Derived from the Centroid Linkage Test

population as fixed, but unknown as with the Engelman and Hartigan (1969) formulation. The second type of alternative is to consider the number of observations from each population as being random, with their probability distribution being determined by unknown mixing parameters. Tyis is the type of formulation of the alternative considered by Lee (1974).

Some of the procedures proposed will be evaluated by considering their power against the alternatives selected.

There are no published tables of percentage points for the Lee (1974) test, hence to compare the power of the previously suggested procedures with the Lee (1974) test would require generation of percentage points for that test. However, there are already percentage points available for the Engelman and Hartigan (1969) test; therefore, alternatives of the Engelman and Hartigan (1969) type will be considered in this study.

For $n = 10$, 6 different two population alternatives are considered: $(.5\sigma, 5\text{-}5)$, $(2\sigma, 5\text{-}5)$, $(4\sigma, 5\text{-}5)$, $(4\sigma, 7\text{-}3)$, $(4\sigma, 9\text{-}1)$, $(6\sigma, 5\text{-}5)$, where the first coordinate represents the separation of the means, in units of $\sigma$, of the two normal populations and the second coordinate represents the numbers of observations from each of the two normal populations. The results are presented in Tables XVIII through XXIII. If an estimate of power has been based on a tabulated value which is theoretically too large, then a * is placed to the right of the estimate to indicate that the estimate of power is conservative. A less conservative estimate of power is given in parentheses. This less conservative estimate is obtained by rejecting the null hypothesis when the Engelman-Hartigan critical point is exceeded.

TABLE XVIII

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, (.5$\sigma$, 5-5) ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | .502 | .542 | .518 | .516 | .519 |
| .75 | .240 | .292 | .260 | .246 | .246 |
| .90 | .084 | .102 | .104 | .104 | .106 |
| .95 | .046 | .064 | .066 | .068 | .070 |
| .99 | .014 | .008* | .008* | .008* | .008* |
| | | (.010) | (.010) | (.010) | (.010) |

TABLE XIX

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, $(2\sigma, 5\text{-}5)$ ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha = .50$ | .592 | .556 | .628 | .640 | .638 |
| .75 | .340 | .344 | .380 | .388 | .386 |
| .90 | .166 | .132 | .152 | .154 | .158 |
| .95 | .090 | .082 | .082 | .082 | .082 |
| .99 | .020 | .012* | .012* | .012* | .012* |
| | | (.018) | (.018) | (.018) | (.018) |

TABLE XX

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, ($4\sigma$, 5-5) ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | .966 | .844 | .944 | .942 | .938 |
| .75 | .820 | .778 | .800 | .810 | .808 |
| .90 | .614 | .600 | .578 | .598 | .602 |
| .95 | .474 | .450 | .442 | .452 | .452 |
| .99 | .218 | .144* | .144* | .144* | .144* |
| | | (.174) | (.174) | (.174) | (.174) |

TABLE XXI

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, $(4\sigma, 7\text{-}3)$ ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha = .50$ | .916 | .868 | .904 | .892 | .892 |
| .75 | .738 | .748 | .742 | .738 | .738 |
| .90 | .448 | .486 | .480 | .494 | .496 |
| .95 | .306 | .364 | .364 | .364 | .366 |
| .99 | .118 | .110* | .110* | .110* | .110* |
| | | (.136) | (.136) | (.136) | (.136) |

TABLE XXII

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, $(4\sigma, 9\text{-}1)$ ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha$ = .50 | .492 | .719 | .502 | .492 | .485 |
| .75 | .270 | .317 | .282 | .268 | .263 |
| .90 | .130 | .122 | .130 | .130 | .128 |
| .95 | .078 | .072 | .072 | .072 | .072 |
| .99 | .018 | .016* | .016* | .016* | .016* |
| | | (.018) | (.018) | (.018) | (.018) |

TABLE XXIII

ESTIMATED POWER OF THE B/W TESTS FOR THE TWO GROUP, $(6\sigma, 5\text{-}5)$ ALTERNATIVE

| | E-H Power | Single | Complete | Weighted Average | Centroid |
|---|---|---|---|---|---|
| $C_\alpha = .50$ | .998 | .990 | .996 | .998 | .998 |
| .75 | .996 | .986 | .986 | .990 | .992 |
| .90 | .956 | .962 | .960 | .962 | .962 |
| .95 | .896 | .908 | .906 | .908 | .908 |
| .99 | .648 | .562* | .562* | .562* | .562* |
| | | (.624) | (.624) | (.624) | (.624) |

## Power Comparisons Among the Proposed Procedures

For each of the six alternatives considered, 500 data sets were generated. The four agglomerative procedures were allowed to cluster the same 500 data sets. For each data set the ratio, B/W, was computed for each algorithm at stage n-1, then the observed value of B/W was referred to the table of estimated percentage points for that algorithm. The Engelman-Hartigan (1969) test statistic, MAX B/W, was computed at a later time for each of the 500 data sets that were generated for each of the six alternatives. MAX B/W was computed from different generated observations than the B/W's that were computed by the agglomerative algorithms. The percentage of rejections of the null hypothesis is an estimate of the power of the procedure.

All the procedures were more powerful when there were an equal number of observations from each of the two populations. This result is not unexpected since in many standard allocation problems the optimal allocation of samples to two groups is proportional to the variances of the groups. In the present case, the variances were assumed to be equal; hence, the data sets were generated from populations having equal variances.

Based on samples of size 500 there does not seem to be any evidence of differences in power between the Engelman-Hartigan test and the procedures based on the four agglomerative clustering algorithms expecially at the $\alpha$ levels most commonly used, $\alpha = .10, .05, .01$. The interpretation may be that in order to have significance at a "low" value of $\alpha$, the observations must be so distinctly separated that all the agglomerative procedures cluster them in the same way. The

agreement of the agglomerative clustering procedures will be discussed further in the next section.

It is interesting to note that the conclusion that Baker and Hubert (1975) reached was that the single linkage method was not a good procedure for most cases. It is to be noted that the null hypothesis that Baker and Hubert (1975) are testing, is the null hypothesis formulated by Ling (1971) which the present author found objectionable in a previous chapter. The size of the power as a function of the separation of the two populations is worthy of note. The power of the tests, at $\alpha = .10$, for the two population alternative where the population means are separated by $.5\sigma$ units is little better than .10 (the estimate of the power of the Engelman-Hartigan (1969) test is even less than .10), which means for this alternative the test is little better than ignoring the data completely and adopting a test which randomly rejects $H_0$ at level $\alpha$. When the separation of the population means is $2\sigma$ units the power increases to about .16, which is somewhat better than ignoring the data. A separation of the two population means by $4\sigma$ units finally leads to a reasonably good power, which is about .6. It should not be too surprising that the power is low for some separations of the populations, considering the almost total lack of assumptions that are imposed, and also considering the small sample size.

## Agreement of the Four Agglomerative Clustering Procedures

The data presented in the previous sections indicate that the four agglomerative procedures "agree" in many cases. Two procedures will be

defined to "agree" if the ratio, B/W, computed from the algorithm assignment of observations to two clusters is the same. Therefore, if two procedures cluster the data in the same way they will "agree", or if they cluster the data in a different manner that leads to the same ratio, B/W, they will be defined to agree. It is to be noted that different clusterings which lead to the same ratio, B/W, for the two group alternative are rare.

The methods that the different algorithms use to measure distances between clusters will be examined as the first step toward understanding why the procedures "agree" as often as is observed. The coefficients that are used in the Lance and Williams (1967) formula, to represent the four agglomerative procedures are given in Table XXIV, where $NG_J$ represents the number of units in cluster J, and $NG_K$ represents the number of units in cluster K.

The following simple example will illustrate several important points. Suppose there are three units to be clustered, A, B, C, that are located at +1, +2, and +4, respectively. The measure of distance is to be Euclidean or squared Euclidean distance. A and B are the closest, and are grouped together first, by all of the procedures. The next step is to compute the distance between C and the group $(A \cup B)$. This distance is computed for each of the four algorithms and for each of the two measures of distance (see Table XXV).

It can be seen that there is some lack of correspondence between the names attached to the procedures and the geometrical interpretation implied by those names. For example, the Euclidean distance from C to the centroid of the cluster $(A \cup B)$ is 2.5 units, but according to the centroid linkage method the distance is computed to be 2.25. Also

TABLE XXIV

COEFFICIENTS IN THE LANCE AND WILLIAMS FORMULA
FOR FOUR AGGLOMERATIVE ALGORITHMS

| | $\alpha_J$ | $\alpha_K$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Single Linkage | .5 | .5 | 0 | -.5 |
| Complete Linkage | .5 | .5 | 0 | .5 |
| Weighted Average Linkage | $\frac{NG_J}{NG_J + NG_K}$ | $\frac{NG_K}{NG_J + NG_K}$ | 0 | 0 |
| Centroid Linkage | $\frac{NG_J}{NG_J + NG_K}$ | $\frac{NG_K}{NG_J + NG_K}$ | $\frac{-(NG_J)(NG_K)}{(NG_J + NG_K)^2}$ | 0 |

TABLE XXV

DISTANCES BETWEEN CLUSTERS AS MEASURED
BY DIFFERENT ALGORITHMS

| | Linkage Method | Euclidean Distance | Squared Euclidean Distance |
|---|---|---|---|
| $d_{(A \cup B),C}$ | Single | 2 | 4 |
| $d_{(A \cup B),C}$ | Complete | 3 | 9 |
| $d_{(A \cup B),C}$ | Weighted Average | 2.5 | 6.5 |
| $d_{(A \cup B),C}$ | Centroid | 2.25 | 6.25 |

the squared Euclidean distance between C and the average of the units in cluster $(A \cup B)$ is 6.25, but the weighted average linkage method computes the distance as 6.5, which is the average of the squared distances, $(AC)^2$, $(BC)^2$. The user should be aware of how the algorithm he selects computes distances and the consequences this selection has on the resulting "clusters". This problem is being studied in detail by DuBien (1975).

A fourth point, X, will be assumed, whose value is greater than that of unit C. If the distance from C to X is greater than the distance from C to the cluster $(A \cup B)$, then C will be added to the cluster $(A \cup B)$. However, if the distance from C to X is less than the distance from C to $(A \cup B)$, then the resultant grouping into two clusters will be $(A \cup B)$, $(C \cup X)$. The behavior of the four procedures in grouping the four points into two clusters will be studied as a function of the position of X. The results are in Table XXVI.

Let $\chi$ denote the coordinate of X. The single linkage algorithm was the algorithm that divided the observations ABC, X for the smallest value of $\chi$, while the complete linkage algorithm was the algorithm that divided the observations, ABC, X for the largest value of $\chi$. These two algorithms are sometimes referred to as boundary algorithms because of their diverse ways of measuring distances. It is also to be noted that the weighted average linkage method is the only one for which the choice of distance measure altered the resultant clustering.

For comparison purposes the maximum, B/W, criterion will cluster the units as follows:

$$4 < \chi < 4 + \frac{5\sqrt{3}}{3} \qquad \text{AB, CX}$$

$$\chi > 4 + \frac{5\sqrt{3}}{3} \qquad \text{ABC, X}$$

TABLE XXVI

CLUSTERING OF FOUR POINTS BY DIFFERENT ALGORITHMS

| Value of | Linkage Method | Distance Measure | Resultant Clustering |
|---|---|---|---|
| $4 < X < 6$ | Single | Euclidean | AB, CX |
| $X > 6$ | Single | Euclidean | ABC, X |
| $4 < X < 6$ | Single | Squared Euclidean | AB, CX |
| $X > 6$ | Single | Squared Euclidean | ABC, X |
| $4 < X < 7$ | Complete | Euclidean | AB, CX |
| $X > 7$ | Complete | Euclidean | ABC, X |
| $4 < X < 7$ | Complete | Squared Euclidean | AB, CX |
| $X > 7$ | Complete | Squared Euclidean | ABC, X |
| $4 < X < 6.5$ | Weighted Average | Euclidean | AB, CX |
| $X > 6.5$ | Weighted Average | Euclidean | ABC, X |
| $4 < X < 4 + \sqrt{6.5}$ | Weighted Average | Squared Euclidean | AB, CX |
| $X > 4 + \sqrt{6.5}$ | Weighted Average | Squared Euclidean | ABC, X |
| $4 < X < 6.25$ | Centroid | Euclidean | AB, CX |
| $X > 6.25$ | Centroid | Euclidean | ABC, X |
| $4 < X < 6.25$ | Centroid | Squared Euclidean | AB, CX |
| $X > 6.25$ | Centroid | Squared Euclidean | ABC, X |

The value of $X$, where the maximum $B/W$ criterion changes from the clustering $AB, CX$ to the clustering $ABC, X$ is found by equating the within cluster $(A \cup B \cup C)$ sum of squares to the pooled within cluster sum of squares computed from the two clusters $(A \cup B)$ and $(C \cup X)$, then solving for $X$.

It is to be noted that all the agglomerative procedures and the maximum $B/W$ criterion agree as to the clustering except for values of $X$ such that $6 < X < 7$. In this interval the agglomerative procedures all disagree as to the proper clustering, and without knowledge about how $X$ is distributed in this interval it is impossible to predict which procedure would most often be in agreement with the maximum $B/W$ clustering. It should be pointed out that this example is only one possible configuration of four points, so the generalizations are somewhat limited. More generally, how much agreement should be expected from the four agglomerative algorithms when they are assigning observations to two groups? The answer will surely depend on the sample size, but in an effort to gain partial information about the agreement of the algorithms, 200 data sets were generated from a single normal population. Each of the agglomerative algorithms clustered each of the data sets and the ratio $B/W$ was computed as usual. The agreement of the procedures was tabulated (see Table XXVII). In almost 75% of the 200 cases, all four procedures produced the same $B/W$ ratio. This percentage is probably a lower bound for $n = 6$, since under any of the two population alternatives the observations are more likely to be separated by larger distances.

For small sample sizes the above results suggest the following "rule of thumb" for the two group alternative: If the four agglomerative

TABLE XXVII

THE AGREEMENT OF FOUR AGGLOMERATIVE ALGORITHMS

| | Number of Agreements |
|---|---|
| All 4 in agreement | 149 |
| When not all 4 agree then | |
| Single and Complete | 1 |
| Single and Weighted Average | 15 |
| Single and Centroid | 19 |
| Complete and Weighted Average | 32 |
| Complete and Centroid | 28 |
| Weighted Average and Centroid | 47 |

procedures agree as to the clustering then continue and perform the tests suggested, but if the agglomerative procedures are not in agreement about the assignment of the observations to two groups then stop. It is unlikely that the observed significance level will be .10 or less if the procedures disagree. This procedure needs further investigation.

## The Role of Clustering Tests as Statistical Tools

Should researchers in various fields be encouraged to use clustering tests? If the purpose of the clustering is to make inferences to units other than the units observed, and if the researcher is concerned about being misled by observations which happened to be "close" by chance, then he should consider a clustering test. Very few assumptions are required to use the test procedures proposed in the previous sections, only that the number of populations under the alternative be known. If more assumptions can be made about the data, then there are probably much better methods than clustering tests to detect the presence of more than one normal population. This suggests that clustering tests are probably most useful in situations where little is known about the data. This might logically occur at the beginning of research in a new field. As with almost any statistical test, as the separation between the populations becomes larger and as the sample size becomes larger the power of the test increases. However the power of these tests is not as large as many researchers tend to think it is. As was pointed out previously with a sample size of 10, the test for 2 populations requires

the populations be separated by almost $4\sigma$ units and have equal (5) observations from each population in order to have power of about .6.

It is the author's opinion that many of the articles in the literature whose conclusions were based on the results of a cluster analysis with no test of any type would not be there if the probability of a type I error was known. It is also the author's opinion that clustering tests will help to screen many spurious conclusions. In addition, it is felt that clustering results have been used as "proof" or evidence in many studies where other more powerful procedures might have reasonably been used. In general, a great deal more care needs to be taken in deciding whether a cluster analysis will be beneficial to the researcher.

# CHAPTER IV

# MATHEMATICAL CLUSTERING

## A Mathematical Definition of a Cluster

The problem of definition of a cluster has been alluded to in a previous chapter. Ling (1971) has proposed a formal definition of a cluster, which incorporates several topological concepts. The first property is connectedness of a set of points. A set of points is r-connected if any two points in the set can be connected by a path traced by line segments between pairs of points of the set, where the line segments are all less than or equal to $r$ in length. If a set of $n$ points is r-connected, then it follows that the set is contained in a sphere of diameter less than or equal to $(n-1)r$. Another important consequence of r-connectedness is that a set of points which is r-connected is also s-connected for any $s > r$. The second property that Ling (1971) considers is bondedness of a set of points. A set of points, $X$ is $(k, r)$ bonded if for every point $x \in X$, a spherical neighborhood of radius $r$ contains $k$ points of $X$ other than $x$. Since a set of points which is $(k, r)$ bonded is not necessarily r-connected a third definition is needed to incorporate both connectedness and bondedness. A set of points is $(k, r)$ connected if it is both r-connected and $(k, r)$ bonded. If $r$ is the minimum value for which a set of points $X$ is $(k, r)$ connected and if $X$ is not a proper subset of any $(k, r)$ connected set, then $X$ is defined to be a $(k, r)$ cluster.

This definition of a cluster is quite flexible; for $k = 1$, rather elongated clusters are possible while for $k = n-1$ all the points of the cluster are contained in a sphere of radius $r$. It is to be noted that although the points are bounded by spheres there are no constraints to require that the shapes of the clusters will be spherical. Using this definition of a cluster, the clusters can be represented by a tree (see Hartigan, 1967).

One method of finding all the $(k, r)$ clusters would be to examine all the possible partitions of the data, but that method is inefficient. Ling (1971) has written a computer program to locate all the $(k, r)$ clusters more efficiently than by searching all partitions.

If one is willing to specify very precisely the type of clusters that are being sought, the clustering problem becomes much the same as the pattern recognition problem in engineering.

## The Role of An Algorithm in Mathematical Clustering

The procedure that Ling (1971) proposes is to define a cluster mathematically, and then write a program to find all the sets of points that satisfy the definition of a cluster. The only role of the algorithm in this situation is to search efficiently through the data for all the clusters.

There is an obvious problem for the researcher who for various reasons cannot accept Ling's (1971) definition of a cluster. The problem is that he would be forced to write a new program to locate all the clusters in the data according to his own definition of a cluster. If the researcher is not willing to write a new program each time the

definition of a cluster changes, he may choose one of the standard clustering algorithms to "approximate" the clusters that he has defined. The new role for the clustering algorithm is then to "approximate" clusters.

Still another role of the algorithm is to define clusters. An algorithm may be chosen to cluster data because of the topological properties of the clusters it produces, such as connectedness. After the algorithm has been selected for use, the data are clustered with the output being operationally defined to be "clusters".

A slight modification of the above procedure leads to still another role for the clustering algorithm. This role may be the most common one in practice. An algorithm or algorithms that happen to be available are used to cluster the data. The resulting "clusters" are regarded as candidates for being real clusters. After an examination of the suggested clusters the researcher formulates the definition of a cluster, although the definition may not be formally stated.

The problems that occur when one tries to evaluate algorithms with respect to their intended roles is discussed in the next section.

## The Problems of Assessing the Performance of Algorithms in the Mathematical Clustering Context

In order to be able to evaluate the performance of an algorithm, the algorithm's intended role must be considered. The role of the algorithm which Ling (1971) devised was to locate all the clusters according to his definition of a cluster. His algorithm does find all the clusters, according to the definitions he has made, so the algorithm works

perfectly with respect to its intended role.

If an algorithm is used to find "approximate" clusters with respect to some definition of a cluster, its performance may be judged on the frequency with which it will locate the "real" clusters in the data, for various configurations of the data points. This is the type of evaluation many researchers seem to have been using to justify the usage of a new algorithm. The researchers will use the new algorithm on the Fisher (1936) Iris data, or the Skull data presented in Sneath and Sokal (1973). Several clusters have been identified by experts for the Iris data and Skull data and the algorithm is tested with respect to its ability to find these "known" clusters. Since there are an infinite number of possible cluster definitions, it is impossible to test the algorithms performance for all cluster definitions.

There is some confusion about the evaluation of algorithms in the mathematical clustering context, because the performance of the algorithm depends on both the properties of the algorithm and the definition of a cluster.

If the purpose or role of an algorithm is to define clusters or to suggest clusters, then there is no objective method with which to evaluate the performance. The algorithm has produced clusters. Several quantities might be computed to help describe the resulting clusters, such as the average squared deviation of points from the centroid of the cluster or the radius of the smallest sphere containing the cluster, but in general there is no set of quantities that can be computed that will contain all the information about an arbitrary cluster.

The conclusion is that if a researcher only wishes to make inferences about the units he has observed, he may choose to compute any

quantity which he feels is informative, but he must justify his selection in the context of his specific problem. There is essentially no statistical problem involved (other than a degenerate one).

# CHAPTER V

# APPLICATION OF MATHEMATICAL CLUSTERING TO A COMPLEX TARGET CONFIGURATION

## Statement of the Problem

A situation which arises in a military context is one in which there are $n$ individual target elements in a complex target configuration. The optimal strategy for attacking the configuration is desired. If the optimal offensive strategy is known, it can be used also for defensive purposes by suggesting modifications to existing installations. If it is assumed that one pass is sufficient to totally destroy a target element, and if it is possible to make $n$ passes, then the optimal strategy is obvious; a pass is attempted at each target element. However, it is unrealistic to assume that $n$ passes can be attempted, if $n$ is at all large, because of the prohibitive cost in terms of personnel and materials. If some of the target elements are close enough together, a single pass may damage several target elements. This suggests that grouping the target elements into a smaller number than $n$ groups may be useful in determining the optimal strategy. If a maximum of $k$ passes can be attempted, can the target elements be assigned to $k$ groups so the maximum damage is inflicted by making a pass at each group? Since clustering algorithms are procedures for assigning observations to groups, can the algorithms be of use to aid in the selection of $k$ optimal aim points? The problem that this

chapter of this study is concerned with is to formulate some strategies based on clustering algorithms in order to provide at least some partial answers to the above question.

## A Simplified Model of a Complex Target Configuration

The first simplifying assumption that will be made is that the complex target configuration is an airfield.

A decision must be made about the kind of inferences that are to be drawn from an examination of an airfield or airfields. The first possible viewpoint is that the specific airfield being examined is a random sample from a population of all possible airfields. Some parametric form might be assumed about the population from which this airfield was sampled. The parameters could then be estimated by examining several airfields. The author has been advised that an "average" airfield has been constructed, which may imply that inferences are to be made in some instances about airfields not actually observed.

Another viewpoint is that the only concern should be with respect to observed airfields and groupings of observed target elements. The argument for this viewpoint is that it seems unreasonable to assume that any military airfield configuration is unknown, anywhere in the world. These airfields, it is argued, are the population of interest, because it is impossible to carry out an attack on an airfield that exists only hypothetically. If inferences are to be made about only the airfields and target elements observed then the mathematical clustering approach seems to be the most reasonable approach.

From maps supplied by the Air Force, a "typical", although fictional, airfield was created arbitrarily by the author and Gibson (1975). Target elements were invented and then were placed where they might reasonably be expected to be found in an airfield complex. A measure of importance was assigned to each target element (see Table XXVIII for a summary of the airfield). There are a number of problems in attempting to model the airfield attack problem. One such problem is the assignment of the importance measure, because in addition to the fact that the importance measures were not being assigned by true experts, there should be a time dependency associated with each element. A surface to air missile emplacement is likely to be very important in the early stages of an attack, with other target elements gaining importance, conditional on nullifying the effectiveness of the missile sites. There was no attempt to build the time dependency into the importance measure, but an attempt was made to subjectively average the importance measure with respect to time.

Another problem is that each target element is represented by a point located at its geometrical center. This representation may cause some difficulties in cases such as the following: Suppose a runway is assigned to group A, while an aircraft located on the end of the runway is assigned to group B. The aircraft and the runway are probably closer together than many of the elements of group A or group B, but they have been assigned to different groups.

The most reasonable assumption about damage within a pattern would be that the damage is most severe near the center of the pattern, the impact point, and that the damage decreases as the outer boundaries of the pattern are approached. In the interests of simplicity it is

TABLE XXVIII

THE AIRFIELD COMPLEX

| Target Element Number | Description | X | Y | Importance |
|---|---|---|---|---|
| 1 | A/C Shelter | 15 | 97 | 6 |
| 2 | A/C Shelter | 11 | 96 | 6 |
| 3 | A/C Shelter | 8 | 98 | 6 |
| 4 | A/C Shelter | 10 | 101 | 6 |
| 5 | A/C Shelter | 14 | 102 | 6 |
| 6 | A/C Repair Hangar | 22 | 87 | 4 |
| 7 | A/C Repair Hangar | 20 | 85 | 4 |
| 8 | A/C Repair Hangar | 18 | 82 | 4 |
| 9 | A/C Repair Hangar | 18 | 86 | 4 |
| 10 | A/C Repair Hangar | 20 | 88 | 4 |
| 11 | Visible Aircraft | 28 | 91 | 13 |
| 12 | Visible Aircraft | 26 | 93 | 13 |
| 13 | Visible Aircraft | 24 | 92 | 13 |
| 14 | Visible Aircraft | 23 | 95 | 13 |
| 15 | Visible Aircraft | 24 | 99 | 13 |
| 16 | Visible Aircraft | 26 | 105 | 13 |
| 17 | Defense Installation | 13 | 113 | 95 |
| 18 | Ordinance | 73 | 104 | 50 |
| 19 | Runway | 51 | 91 | 20 |
| 20 | Runway | 74 | 77 | 20 |
| 21 | Defense Installation | 105 | 52 | 100 |
| 22 | Tower | 109 | 70 | 65 |
| 23 | Command Post | 136 | 89 | 70 |
| 24 | Pumping Station | 151 | 92 | 5 |
| 25 | Fuel Storage | 150 | 94 | 17 |
| 26 | Fuel Storage | 153 | 94 | 17 |
| 27 | Fuel Storage | 154 | 92 | 17 |
| 28 | Runway | 126 | 109 | 15 |
| 29 | Runway | 108 | 102 | 25 |
| 30 | Runway | 104 | 84 | 15 |

TABLE XXVIII (continued)

| Target Element Number | Description | X | Y | Importance |
|---|---|---|---|---|
| 31 | Defense Installation | 118 | 148 | 90 |
| 32 | Runway | 113 | 124 | 15 |
| 33 | A/C Repair Hangar | 115 | 80 | 4 |
| 34 | A/C Repair Hangar | 116 | 83 | 4 |
| 35 | A/C Repair Hangar | 115 | 86 | 4 |
| 36 | A/C Repair Hangar | 115 | 89 | 4 |
| 37 | A/C Repair Hangar | 116 | 92 | 4 |
| 38 | A/C Repair Hangar | 121 | 98 | 4 |
| 39 | A/C Repair Hangar | 123 | 98 | 4 |
| 40 | A/C Repair Hangar | 126 | 97 | 4 |
| 41 | A/C Repair Hangar | 128 | 97 | 4 |
| 42 | Visible Aircraft | 99 | 103 | 13 |
| 43 | Visible Aircraft | 101 | 105 | 13 |
| 44 | Visible Aircraft | 105 | 106 | 13 |
| 45 | A/C Shelters | 94 | 100 | 6 |
| 46 | A/C Shelters | 97 | 106 | 6 |
| 47 | A/C Shelters | 99 | 109 | 6 |
| 48 | A/C Shelters | 100 | 104 | 6 |
| 49 | A/C Shelters | 104 | 106 | 6 |
| 50 | A/C Shelters | 106 | 109 | 6 |

assumed that the damage is uniform within the pattern.

Weapons are differentially effective against target elements; an incendiary weapon may be effective against a petroleum storage area, but highly ineffective against a tank. Once again, to simplify the problem to a manageable level it is assumed the nature of the weapon causing the pattern is irrelevant; that is, it is assumed that all target elements are damaged equally by all weapons.

## Selection of Variables and Scaling

There are many variables that could be measured on each of the target elements, but the inclusion of too many variables might obscure the main issue. The model must be kept simple enough to allow a decision to be made about whether or not clustering algorithms will be useful aids to a decision maker, who wishes to select optimal aim points. If the algorithms appear to be performing satisfactorily for the simple problems, then more variables can be added to the problem, but if the algorithms are not performing well for simple cases using the most important variables, then there is little hope that the algorithms will perform well for more complex cases.

It seems reasonable to assume that the most important variables for grouping target elements together to maximize damage to the group will be variables measuring physical separation. Two target elements that are separated by a "large" distance cannot be damaged in a single pass no matter how similar the elements may be to each other in size, shape, construction material, etc. Two variables which were selected to measure physical separation were rectangular coordinates measured from some arbitrarily defined origin.

The first attempts to cluster the target elements utilized only these coordinates, and since there was no bias desired in any direction, the variables were not scaled. Later attempts to cluster the target elements included the importance measure as a third variable.

Scaling can be used to bias the shapes of resulting clusters, in addition to the usual usage, which is to convert all the measurements on a variable to comparable units. An example of how scaling can be used to bias the shape of clusters is the following: Suppose there are four points A, B, C, D located at (1, 0), (-1, 0), (0, 2), (0, -2) respectively. Without any scaling, points A and B are the closest. Suppose a bias is desired in the direction of the y-axis, then the y co-ordinate is multiplied by a scale factor k, where $0 < k < 1$. For selections of k, $\frac{\sqrt{3}}{2} < k < 1$, A and B are still the closest points, but if $k = \frac{\sqrt{3}}{2}$ the distances AB, AC, AD, BC, BD are all the same. For selections of k, $\frac{\sqrt{3}}{6} < k < \frac{\sqrt{3}}{2}$, AC, AD, BC, BD are equal and are the smallest distances. When $0 < k < \frac{\sqrt{3}}{6}$ the smallest distance is CD, which initially was the largest distance.

In order to use scaling to bias the resulting clusters, it is necessary to know both the magnitude and direction of the bias desired.

## Two Strategies for Selecting Aimpoints

Suppose that a maximum of k passes can be made at an airfield complex of n target elements. One present procedure is to look at a map displaying the target elements and visually select the k best aimpoints. Another procedure is to use a computer program by Gay (1974), which evaluates the damage over a grid of points covering the entire airfield complex, to select the k best aim points. The following

procedures seek to restrict the search for optimal aim points to groups of target elements, which are close enough together to be damaged by a single pass, when such groups exist. Beginning at stage one the target elements are each regarded as a cluster, which is represented by a single point. It is assumed that any size or shape of pattern will cover a single point. At stage two, two points have been joined so there are n-1 clusters. The value of $f_1$, defined previously in Chapter II, is computed in order to measure the distance that the two entities just merged were apart. If the distance is small enough relative to the pattern size, then the clustering is carried out through the next stage. Thus $f_1, f_2, \ldots, f_t$ are examined successively until either k clusters are formed or until the pattern is too small to damage all the elements in the cluster most recently formed. If there are more than k clusters when the procedure is terminated, then the importance measures of the points within each cluster are added, with the k largest importance totals determining the k clusters for aim point purposes. Both strategies are the same until this point in the procedure. The first strategy is to select the centroid from each of the k clusters as the best aimpoint. The second strategy is to use the Gay (1974) program on each of the k clusters to find the best aimpoint within that particular cluster.

It is to be noted that the values of the $f_i$'s provide different information about the clusters when different algorithms are used. A further discussion about how to evaluate the information provided by the $f_i$'s will be presented in the next section.

## Selection of the Algorithms to Execute the Strategies

It would be possible to define the clusters in an airfield complex, for a known size and shape of pattern, as "disjoint sets of points such that the pattern could be delivered in such a way as to damage all the elements in the cluster". Theoretically it would be possible to write a program which would search all possible partitions of the data to locate all the sets of target elements that could be damaged by the pattern given. However, this procedure would likely be as costly, in terms of computing time, as the Gay (1974) program. As an alternative two of the four agglomerative procedures, discussed previously, could be used to approximate the clusters.

As was previously mentioned, the $f_i$'s computed from the different algorithms provide different information about the configuration of the target elements in the clusters. Consider first the single linkage algorithm, in which $f_i$ is computed from the target elements in the two clusters, $J$ and $K$, which were joined at stage $i + 1$. $f_i$ is the minimum distance between two target elements, where one of the target elements is from $J$ and the other target element is from $K$. It is easy to see that if $f_i = t$, then all the clusters at stage $i + 1$ are t-connected. The single linkage algorithm produces a set of monotonic $f_i$'s; that is, $f_1 < f_2 < \ldots < f_{n-1}$, (see Sneath and Sokal, 1973).

It is also easy to see that a circle of diameter $t(s-1)$, where $s$ is the largest number of target elements contained in any cluster at stage $i + 1$, will cover or bound the target elements in any cluster. This algorithm can locate elongated clusters more easily than most other

algorithms. Elongated clusters might be appropriate when the pattern has been produced by a stick of weapons used with a "long" intervalometer setting. There are two possible difficulties; first, the bound may be too large to be practically useful in determining a stopping point in the algorithm. Secondly, the target elements are bounded by a circle, but there is no implication about the shape of the configuration of target elements contained within the circle. For a given value of $f_i$, the elements may be either elongated or more compact. If the complete linkage algorithm is used to cluster the target elements then the $f_i$ is computed from the two target elements in the two clusters, J and K, which were joined at stage $i + 1$. $f_i$ is computed as the maximum distance between two target elements, where one of the target elements is from J and the other target element is from K. It is easy to see that if $f_i = t$, then all the clusters at stage $i + 1$ are t-connected and $(s\text{-}1, \frac{\sqrt{3}}{2}t)$ bonded, where s is the maximum number of target elements contained in any cluster at stage $i + 1$. The complete linkage algorithm also produces a set of monotonic $f_i$'s (see Sneath and Sokal, 1973). It is also easy to see that a circle of diameter $\sqrt{3}t$ will cover or bound the target elements in any cluster at stage $i + 1$. The bound, for a fixed value of $f_i$, given by this algorithm is much smaller than the bound given by the single linkage algorithm.

There are difficulties involved with using the average and centroid linkage methods for the two strategies proposed. It is difficult to obtain "good" bounds for the average linkage method, and in addition to this problem, the centroid method does not necessarily produce a monotonic set of $f_i$'s (see Sneath and Sokal, 1973). However, in the next section all four of the agglomerative algorithms will be used to

cluster the target elements. This will allow a comparison of the clusters generated at each stage of the agglomerative process and may generate suggestions for further study.

## Application of the Algorithms to the Data and Discussion of Results

The data were first clustered by each of four agglomerative algorithms, utilizing the two physical coordinates as the variables. Squared Euclidean distance was selected as the measure of distance. Tables XXIX through XXXII summarize the clustering for these four algorithms. For notational purposes, in the tables if there is more than one target element (TE) in a cluster, the smallest numbered element in the cluster will be used to identify the cluster. Example: If elements 1 and 3 were joined at the first stage, the notation would be element 1 joined element 3. At the next stage if element 4 joined the first two elements the notation would be element 1 joins element 4. Since k, the maximum number of passes that can be made is usually small relative to n, the number of target elements, the clustering was examined most closely for small numbers of clusters. Figures 6 through 13 visually show the resulting clusters for $k = 2, 4, \ldots, 16$, for each of the four algorithms.

There are some specific groupings which are worthy of note. When $k = 16$, the four algorithms agree very closely except that the single linkage algorithm has grouped elements 33-41 together, whereas the other 3 algorithms have divided these elements into two clusters. When $k = 14$, the single linkage algorithm has produced 10 single element clusters, which is the most of any of the procedures. The complete

TABLE XXIX

CLUSTERING OF THE TARGET ELEMENTS USING SINGLE LINKAGE WITH TWO VARIABLES

| Stage | Target Element | Joined Target Element | Max # Elements in any Cluster | $f_i$ |
|---|---|---|---|---|
| 2 | 44 | 49 | 2 | 1.00 |
| 3 | 42 | 48 | 2 | 2.00 |
| 4 | 42 | 43 | 3 | 2.00 |
| 5 | 38 | 39 | 3 | 4.00 |
| 6 | 40 | 41 | 3 | 4.00 |
| 7 | 7 | 9 | 3 | 5.00 |
| 8 | 6 | 10 | 3 | 5.00 |
| 9 | 12 | 13 | 3 | 5.00 |
| 10 | 24 | 25 | 3 | 5.00 |
| 11 | 26 | 27 | 3 | 5.00 |
| 12 | 6 | 7 | 4 | 8.00 |
| 13 | 11 | 12 | 4 | 8.00 |
| 14 | 24 | 26 | 4 | 8.00 |
| 15 | 35 | 36 | 4 | 9.00 |
| 16 | 11 | 14 | 4 | 10.00 |
| 17 | 33 | 34 | 4 | 10.00 |
| 18 | 33 | 35 | 4 | 10.00 |
| 19 | 33 | 37 | 5 | 10.00 |
| 20 | 38 | 40 | 5 | 10.00 |
| 21 | 42 | 44 | 5 | 10.00 |
| 22 | 42 | 50 | 6 | 10.00 |
| 23 | 2 | 3 | 6 | 13.00 |
| 24 | 2 | 4 | 6 | 13.00 |
| 25 | 6 | 8 | 6 | 13.00 |
| 26 | 42 | 46 | 7 | 13.00 |
| 27 | 42 | 47 | 8 | 13.00 |
| 28 | 1 | 2 | 8 | 17.00 |
| 29 | 1 | 5 | 8 | 17.00 |
| 30 | 11 | 15 | 8 | 17.00 |

TABLE XXIX (continued)

| Stage | Target Element | Joined Target Element | Max # Elements in any Cluster | $f_i$ |
|---|---|---|---|---|
| 31 | 29 | 42 | 9 | 25.00 |
| 32 | 6 | 11 | 10 | 29.00 |
| 33 | 29 | 45 | 10 | 34.00 |
| 34 | 6 | 16 | 11 | 40.00 |
| 35 | 33 | 38 | 11 | 61.00 |
| 36 | 1 | 6 | 16 | 68.00 |
| 37 | 1 | 17 | 17 | 122.00 |
| 38 | 30 | 33 | 17 | 125.00 |
| 39 | 23 | 30 | 17 | 128.00 |
| 40 | 23 | 28 | 17 | 130.00 |
| 41 | 22 | 23 | 17 | 136.00 |
| 42 | 22 | 29 | 23 | 164.00 |
| 43 | 22 | 24 | 27 | 221.00 |
| 44 | 22 | 32 | 28 | 274.00 |
| 45 | 21 | 22 | 29 | 340.00 |
| 46 | 18 | 21 | 30 | 457.00 |
| 47 | 1 | 19 | 30 | 529.00 |
| 48 | 18 | 31 | 31 | 601.00 |
| 49 | 1 | 18 | 49 | 653.00 |
| 50 | 1 | 20 | 50 | 725.00 |

TABLE XXX

CLUSTERING OF THE TARGET ELEMENTS USING COMPLETE LINKAGE WITH TWO VARIABLES

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 2 | 44 | 49 | 1.00 |
| 3 | 42 | 48 | 2.00 |
| 4 | 38 | 39 | 4.00 |
| 5 | 40 | 41 | 4.00 |
| 6 | 7 | 9 | 5.00 |
| 7 | 6 | 10 | 5.00 |
| 8 | 12 | 13 | 5.00 |
| 9 | 24 | 25 | 5.00 |
| 10 | 26 | 27 | 5.00 |
| 11 | 42 | 43 | 8.00 |
| 12 | 35 | 36 | 9.00 |
| 13 | 33 | 34 | 10.00 |
| 14 | 2 | 3 | 13.00 |
| 15 | 12 | 14 | 13.00 |
| 16 | 46 | 47 | 13.00 |
| 17 | 44 | 50 | 13.00 |
| 18 | 7 | 8 | 16.00 |
| 19 | 4 | 5 | 17.00 |
| 20 | 24 | 26 | 20.00 |
| 21 | 42 | 46 | 36.00 |
| 22 | 35 | 37 | 37.00 |
| 23 | 15 | 16 | 40.00 |
| 24 | 1 | 4 | 41.00 |
| 25 | 6 | 7 | 41.00 |
| 26 | 11 | 12 | 41.00 |
| 27 | 38 | 40 | 50.00 |
| 28 | 1 | 2 | 52.00 |
| 29 | 29 | 44 | 53.00 |
| 30 | 42 | 45 | 106.00 |

TABLE XXX (continued)

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 31 | 30 | 33 | 145.00 |
| 32 | 28 | 38 | 148.00 |
| 33 | 6 | 11 | 194.00 |
| 34 | 30 | 35 | 208.00 |
| 35 | 29 | 42 | 225.00 |
| 36 | 1 | 17 | 293.00 |
| 37 | 23 | 24 | 333.00 |
| 38 | 21 | 22 | 340.00 |
| 39 | 1 | 15 | 373.00 |
| 40 | 31 | 32 | 601.00 |
| 41 | 18 | 19 | 653.00 |
| 42 | 18 | 20 | 730.00 |
| 43 | 1 | 6 | 986.00 |
| 44 | 29 | 30 | 1097.00 |
| 45 | 23 | 28 | 1125.00 |
| 46 | 21 | 29 | 3285.00 |
| 47 | 21 | 23 | 4068.00 |
| 48 | 1 | 18 | 5017.00 |
| 49 | 21 | 31 | 9385.00 |
| 50 | 1 | 21 | >10,000.00 |

TABLE XXXI

CLUSTERING OF THE TARGET ELEMENTS USING WEIGHTED AVERAGE LINKAGE WITH TWO VARIABLES

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 2 | 44 | 49 | 1.00 |
| 3 | 42 | 48 | 2.00 |
| 4 | 38 | 39 | 4.00 |
| 5 | 40 | 41 | 4.00 |
| 6 | 7 | 9 | 5.00 |
| 7 | 6 | 10 | 5.00 |
| 8 | 12 | 13 | 5.00 |
| 9 | 24 | 25 | 5.00 |
| 10 | 26 | 27 | 5.00 |
| 11 | 42 | 43 | 5.00 |
| 12 | 35 | 36 | 9.00 |
| 13 | 33 | 34 | 10.00 |
| 14 | 6 | 7 | 10.50 |
| 15 | 12 | 14 | 11.50 |
| 16 | 24 | 26 | 11.50 |
| 17 | 44 | 50 | 11.50 |
| 18 | 2 | 3 | 13.00 |
| 19 | 46 | 47 | 13.00 |
| 20 | 4 | 5 | 17.00 |
| 21 | 42 | 46 | 20,83 |
| 22 | 11 | 12 | 22.00 |
| 23 | 35 | 37 | 23.50 |
| 24 | 6 | 8 | 27.50 |
| 25 | 38 | 40 | 28.00 |
| 26 | 1 | 2 | 33.50 |
| 27 | 1 | 4 | 33.83 |
| 28 | 29 | 44 | 36.67 |
| 29 | 15 | 16 | 40.00 |
| 30 | 29 | 42 | 57.40 |

TABLE XXXI (continued

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 31 | 33 | 35 | 65.00 |
| 32 | 6 | 11 | 95.40 |
| 33 | 29 | 45 | 114.33 |
| 34 | 28 | 38 | 142.00 |
| 35 | 30 | 33 | 152.20 |
| 36 | 1 | 15 | 211.80 |
| 37 | 1 | 17 | 232.57 |
| 38 | 23 | 28 | 269.60 |
| 39 | 1 | 6 | 298.06 |
| 40 | 22 | 30 | 299.50 |
| 41 | 22 | 23 | 521.09 |
| 42 | 29 | 32 | 522.50 |
| 43 | 18 | 19 | 653.00 |
| 44 | 18 | 20 | 727.50 |
| 45 | 22 | 29 | 782.08 |
| 46 | 22 | 24 | 1920.58 |
| 47 | 1 | 18 | 2572.04 |
| 48 | 21 | 22 | 2604.43 |
| 49 | 21 | 31 | 3235.89 |
| 40 | 1 | 21 | 9246.42 |

TABLE XXXII

CLUSTERING OF THE TARGET ELEMENTS USING CENTROID LINKAGE WITH TWO VARIABLES

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 2 | 44 | 49 | 1.00 |
| 3 | 42 | 48 | 2.00 |
| 4 | 38 | 39 | 4.00 |
| 5 | 40 | 41 | 4.00 |
| 6 | 42 | 43 | 4.50 |
| 7 | 7 | 9 | 5.00 |
| 8 | 6 | 10 | 5.00 |
| 9 | 12 | 13 | 5.00 |
| 10 | 24 | 25 | 5.00 |
| 11 | 26 | 27 | 5.00 |
| 12 | 6 | 7 | 8.00 |
| 13 | 24 | 26 | 9.00 |
| 14 | 35 | 36 | 9.00 |
| 15 | 33 | 34 | 10.00 |
| 16 | 12 | 14 | 10.25 |
| 17 | 44 | 50 | 11.25 |
| 18 | 42 | 46 | 13.00 |
| 19 | 2 | 3 | 13.00 |
| 20 | 2 | 4 | 16.25 |
| 21 | 11 | 12 | 18.89 |
| 22 | 42 | 47 | 20.31 |
| 23 | 35 | 37 | 21.25 |
| 24 | 6 | 8 | 24.25 |
| 25 | 1 | 5 | 26.00 |
| 26 | 1 | 2 | 24.72 |
| 27 | 38 | 40 | 26.00 |
| 28 | 29 | 44 | 34.00 |
| 29 | 15 | 16 | 40.00 |
| 30 | 29 | 42 | 43.02 |

TABLE XXXII (continued)

| Stage | Target Element | Joined Target Element | $f_i$ |
|---|---|---|---|
| 31 | 33 | 35 | 56.28 |
| 32 | 6 | 11 | 83.04 |
| 33 | 29 | 45 | 96.65 |
| 34 | 30 | 33 | 133.96 |
| 35 | 28 | 38 | 134.50 |
| 36 | 6 | 15 | 183.17 |
| 37 | 1 | 6 | 179.84 |
| 38 | 23 | 28 | 242.08 |
| 39 | 22 | 30 | 265.69 |
| 40 | 22 | 23 | 403.03 |
| 41 | 1 | 17 | 416.10 |
| 42 | 29 | 32 | 497.89 |
| 43 | 22 | 29 | 559.18 |
| 44 | 18 | 19 | 653.00 |
| 45 | 18 | 20 | 564.25 |
| 46 | 22 | 24 | 1662.81 |
| 47 | 21 | 22 | 2182.37 |
| 48 | 1 | 18 | 2241.93 |
| 49 | 21 | 31 | 2755.73 |
| 50 | 1 | 21 | 8290.97 |

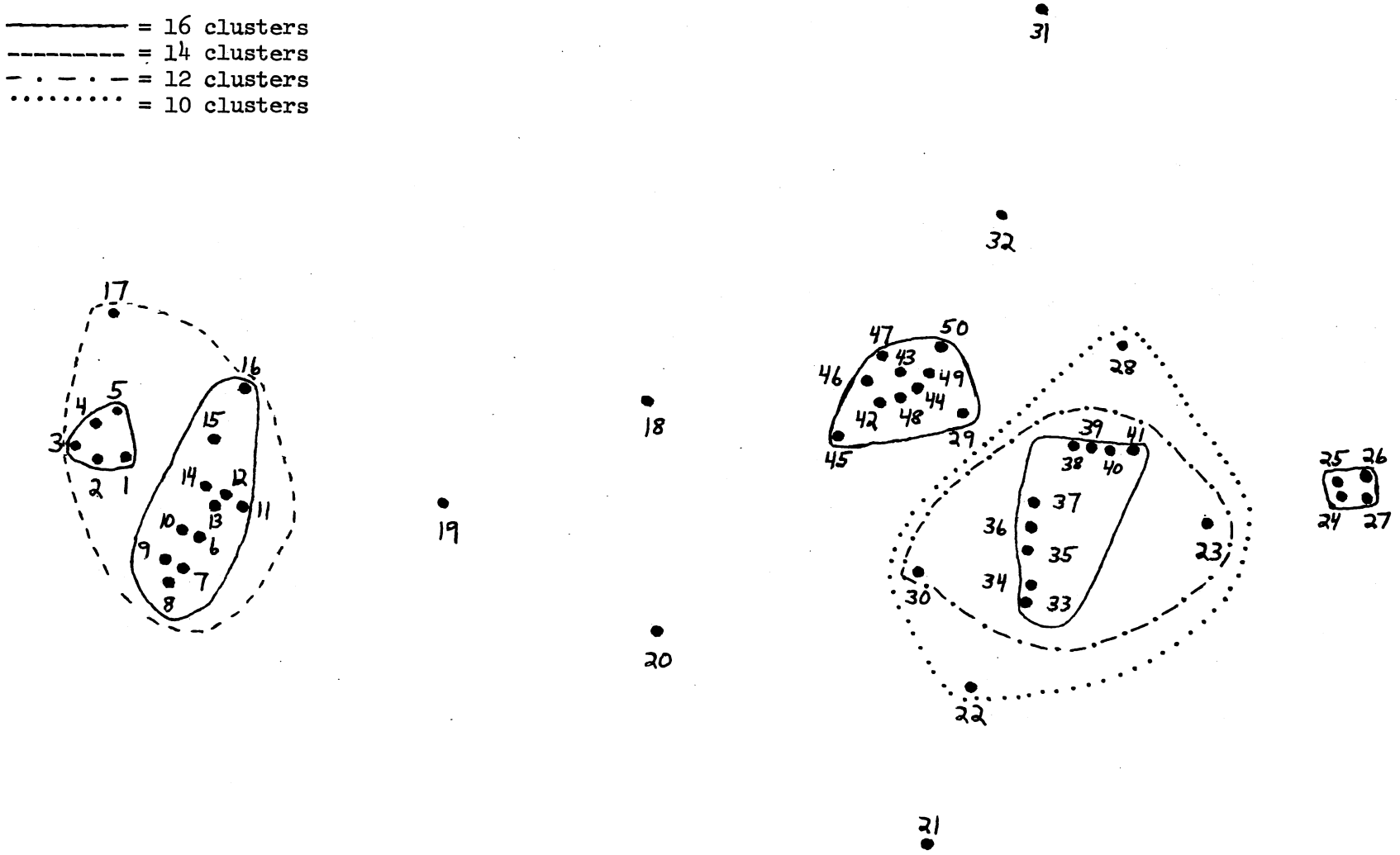


Figure 6. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Single Linkage Algorithm Using Two Variables

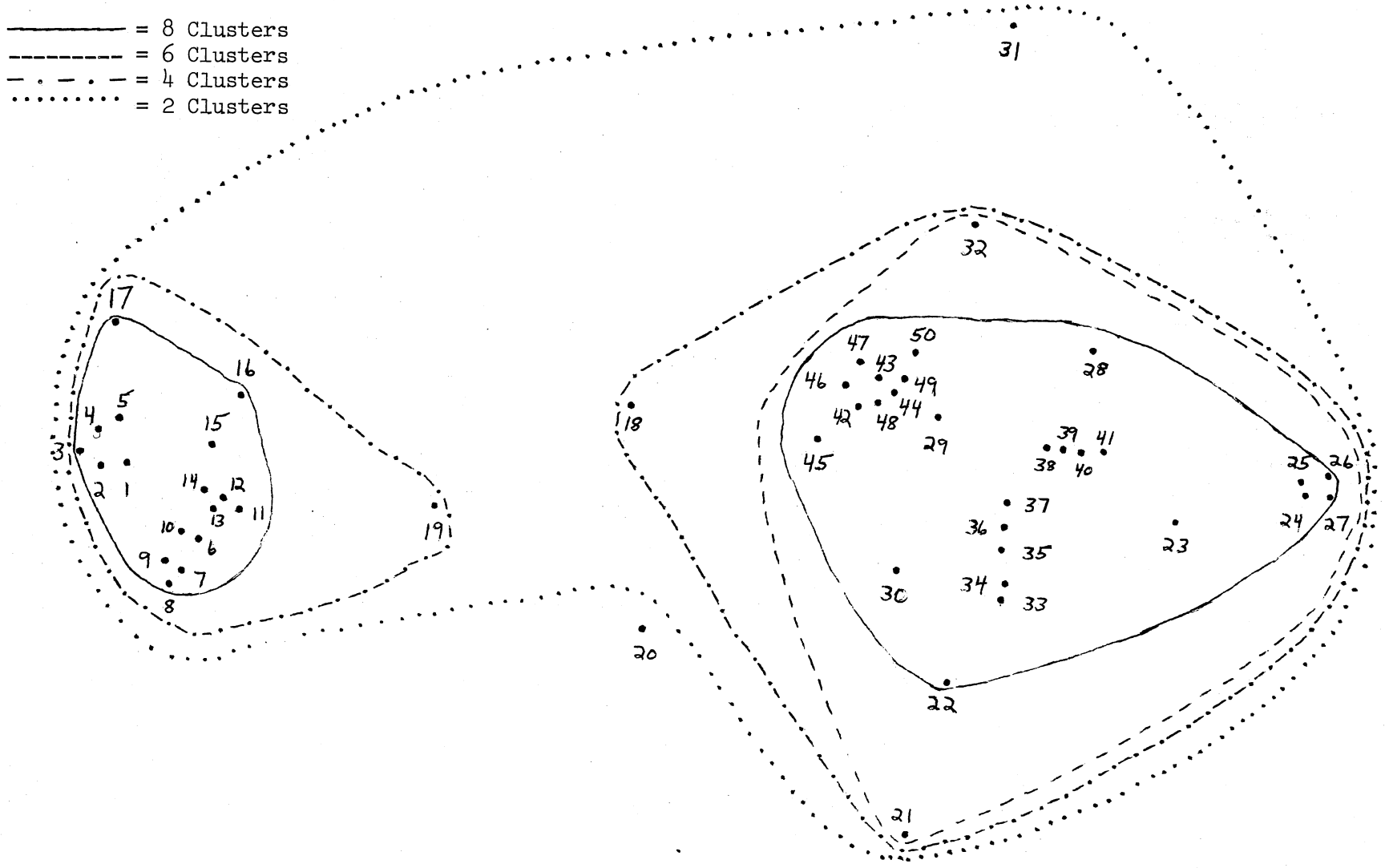


Figure 7. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Single Linkage Algorithm Using Two Variables

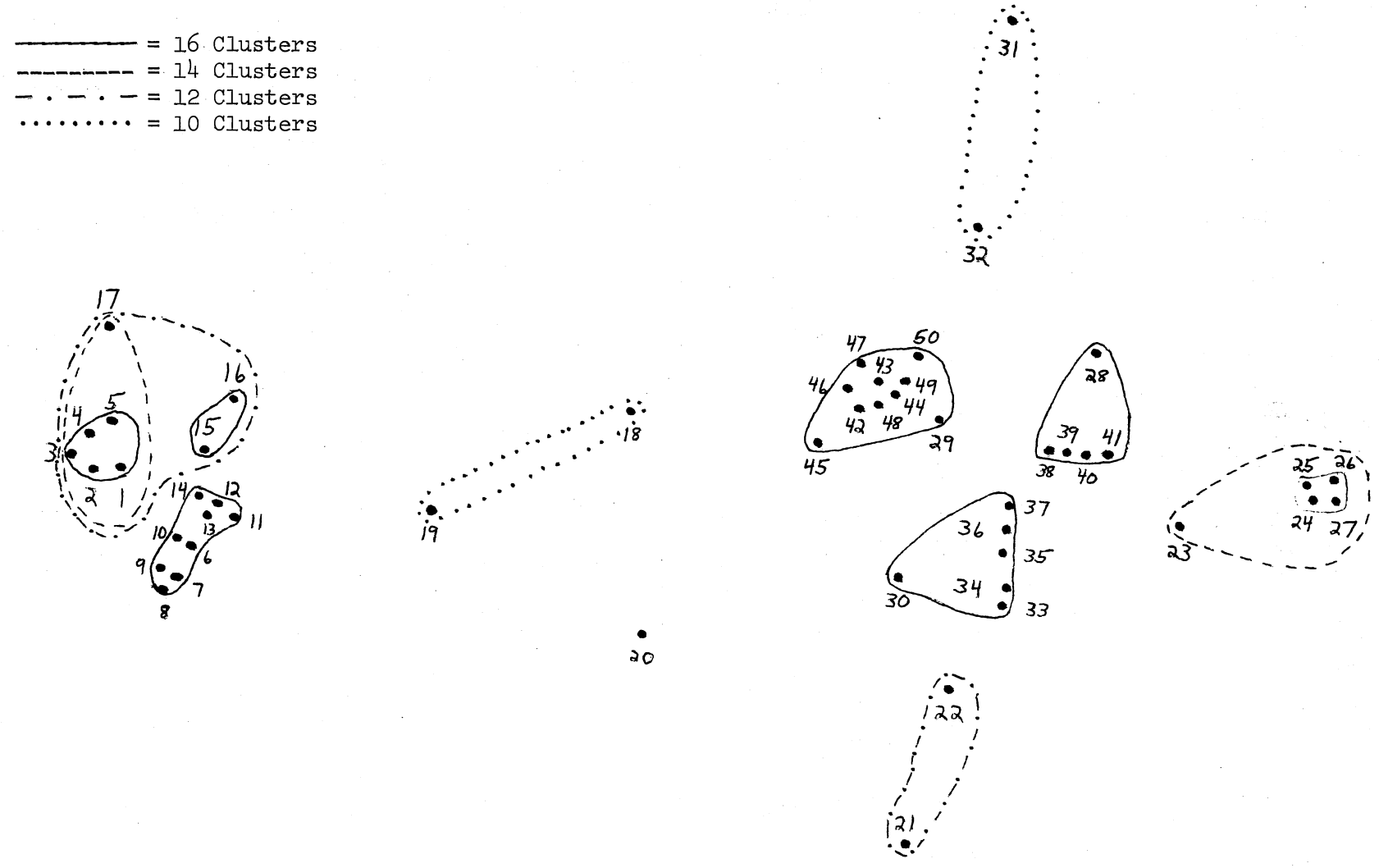


Figure 8. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Complete Linkage Algorithm Using Two Variables

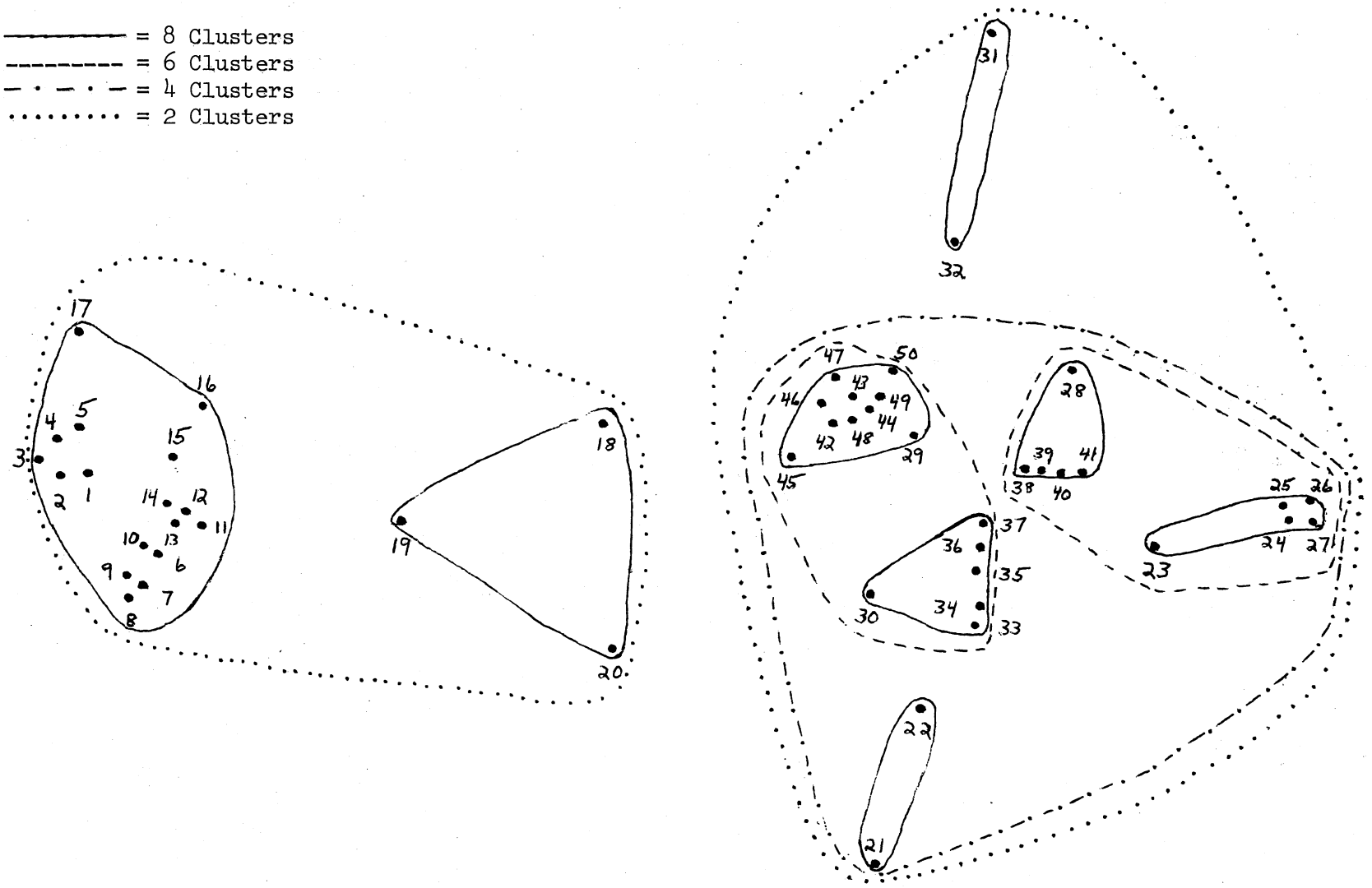


Figure 9. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Complete Linkage Algorithm Using Two Variables

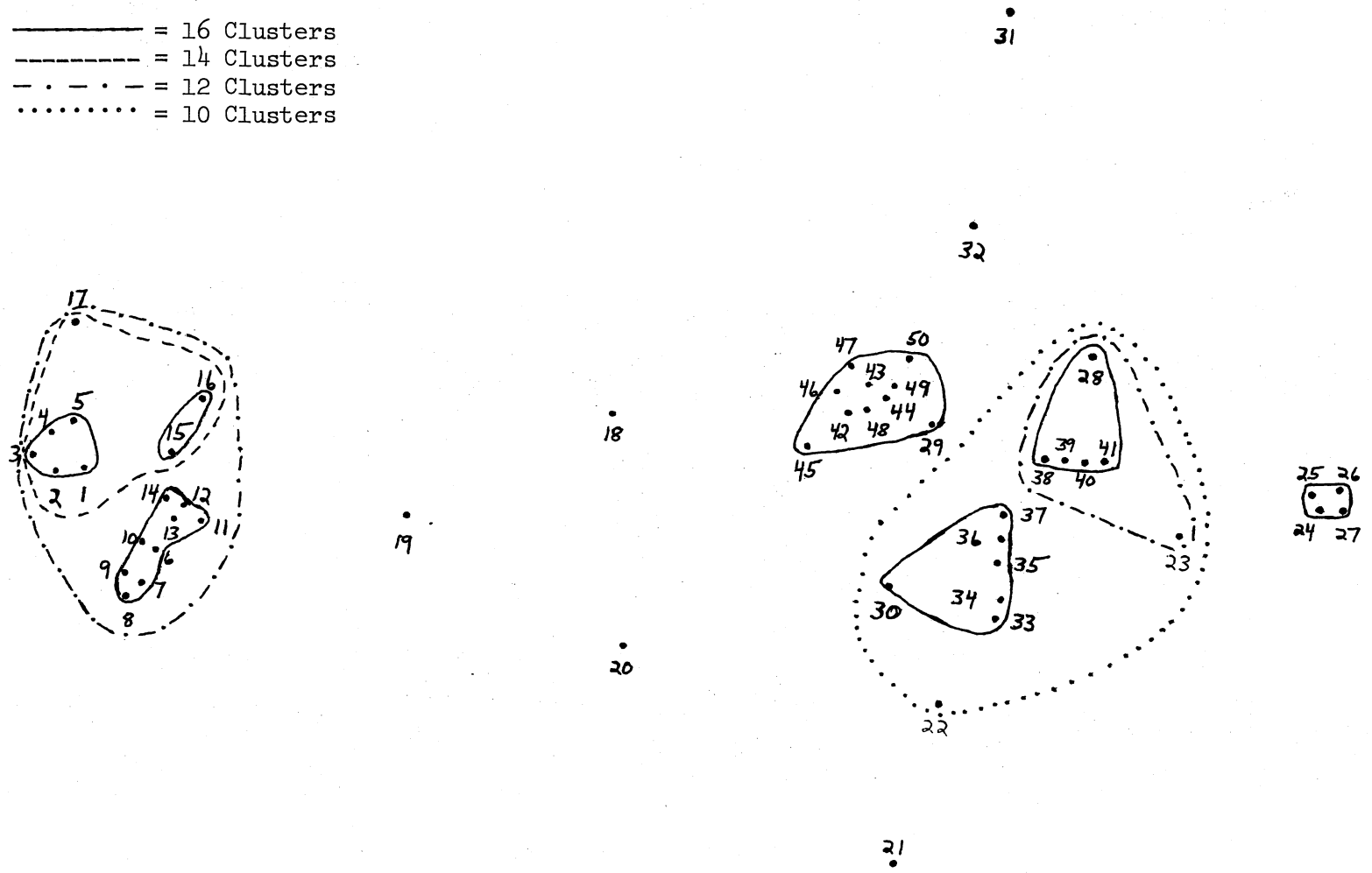


Figure 10. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Weighted Average Linkage Algorithm Using Two Variables

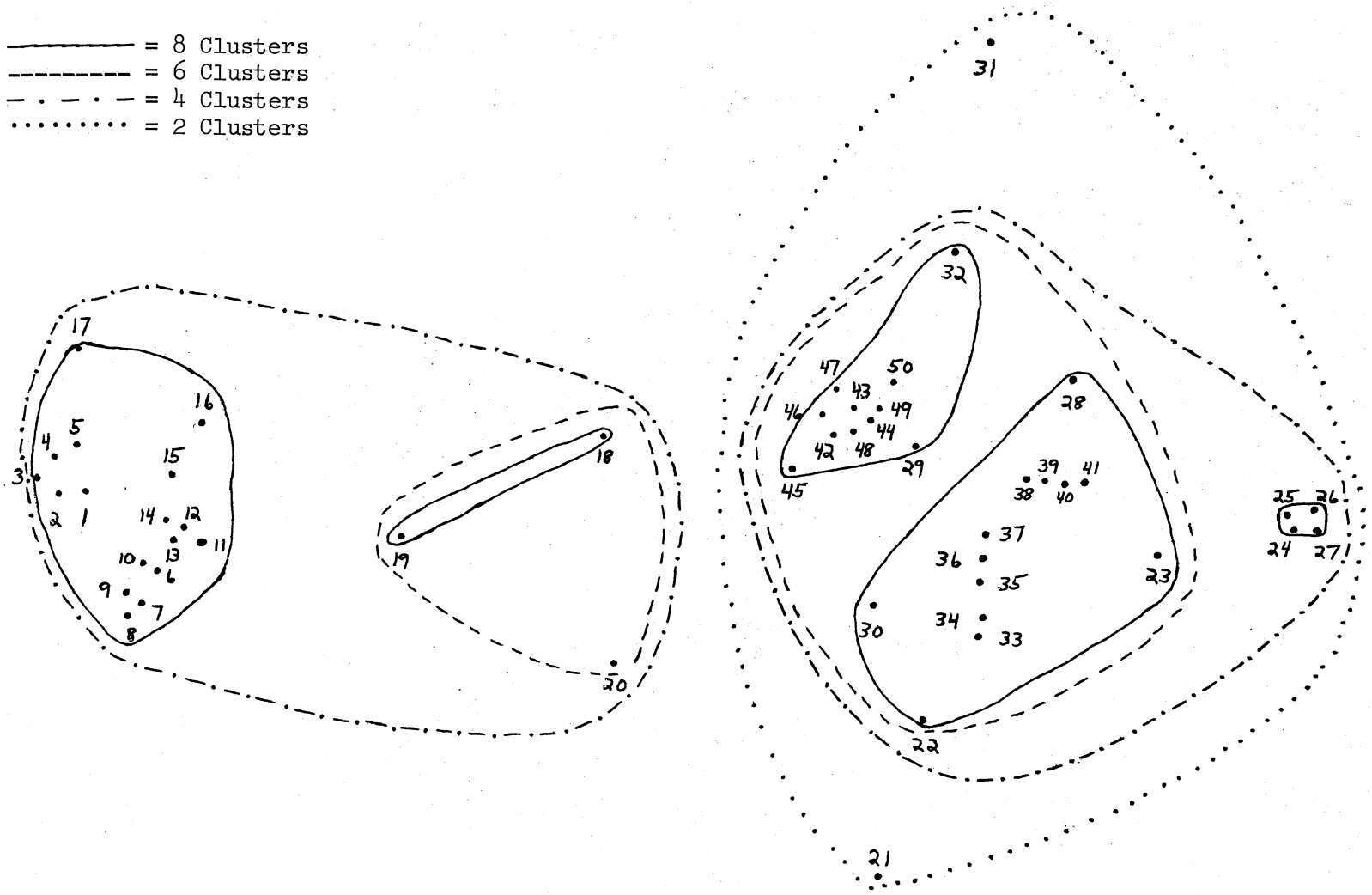


Figure 11. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Weighted Average Linkage Algorithm Using Two Variables

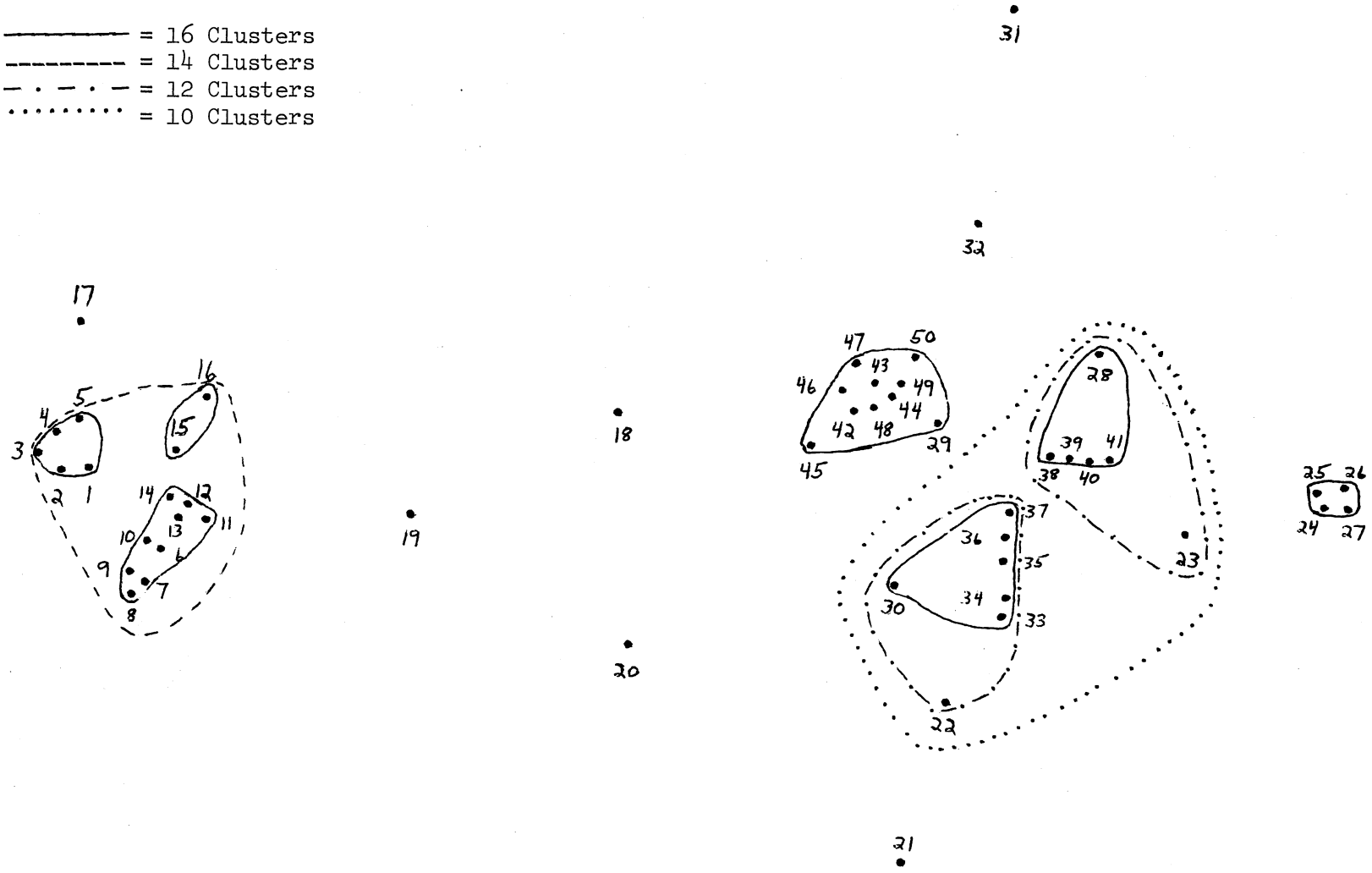


Figure 12. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Centroid Linkage Algorithm Using Two Variables

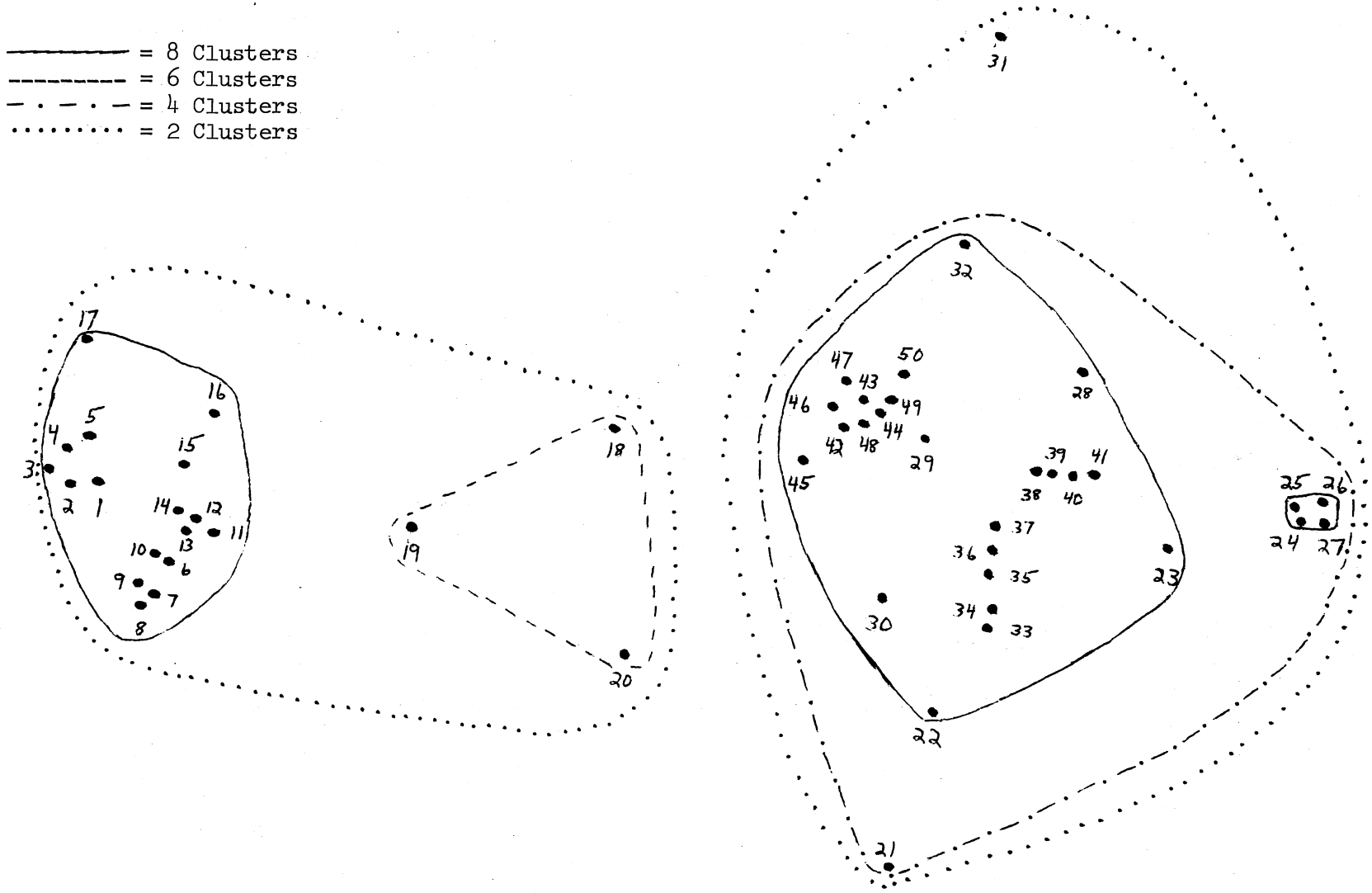


Figure 13. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Centroid Linkage Algorithm Using Two Variables

linkage algorithm has produced only 7 single element clusters, which is the least of any of the procedures. The four algorithms begin to differ more as $k$ becomes smaller. When $k = 10$, the single linkage algorithm has produced 4 large clusters and 6 single element clusters, while the complete linkage algorithm has produced 9 moderate sized clusters and only 1 single element cluster. At this stage of the clustering the single, average and centroid algorithms have produced essentially the same clusters, while the complete linkage algorithm has produced a much different set of clusters. In general the single linkage algorithm has tended to produce a few large clusters and many single element clusters.

The complete linkage algorithm has tended to produce many moderate sized clusters and few single element clusters. The other two algorithms produced clusters which were usually between the extremes produced by the single and complete linkage algorithms.

It was noted previously that the centroid algorithm does not necessarily produce a monotonic set of $f_i$'s, but for this case the $f_i$'s produced were monotonic.

The clusters, which were produced by the algorithms when the importance measure was used as a third variable, were examined for $k = 2, 4, 6, \ldots, 16$. Tables XXXIII through XXXVI summarize the clustering for the four algorithms. The same notation is utilized, except the $f_i$'s are not given because it is not known how to interpret them in these cases. Figures 14 through 21 visually show the resulting clusters for $k = 2, 4, 6, \ldots, 16$.

There are several groupings for these cases which also require special note. When $k = 16$, all of the procedures assign target elements

TABLE XXXIII

CLUSTERING OF THE TARGET ELEMENTS USING SINGLE LINKAGE WITH THREE VARIABLES

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 2 | 38 | 39 |
| 3 | 40 | 41 |
| 4 | 7 | 9 |
| 5 | 6 | 10 |
| 6 | 12 | 13 |
| 7 | 26 | 27 |
| 8 | 6 | 7 |
| 9 | 11 | 12 |
| 10 | 42 | 43 |
| 11 | 25 | 26 |
| 12 | 35 | 36 |
| 13 | 11 | 14 |
| 14 | 33 | 34 |
| 15 | 33 | 35 |
| 16 | 33 | 37 |
| 17 | 38 | 40 |
| 18 | 2 | 3 |
| 19 | 2 | 4 |
| 20 | 6 | 8 |
| 21 | 46 | 47 |
| 22 | 46 | 48 |
| 23 | 49 | 50 |
| 24 | 1 | 2 |
| 25 | 1 | 5 |
| 26 | 11 | 15 |
| 27 | 42 | 44 |
| 28 | 46 | 49 |
| 29 | 11 | 16 |
| 30 | 45 | 46 |

TABLE XXXIII (continued)

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 31 | 42 | 45 |
| 32 | 33 | 38 |
| 33 | 1 | 6 |
| 34 | 1 | 11 |
| 35 | 24 | 25 |
| 36 | 29 | 42 |
| 37 | 30 | 33 |
| 38 | 28 | 30 |
| 39 | 28 | 29 |
| 40 | 28 | 32 |
| 41 | 24 | 28 |
| 42 | 1 | 19 |
| 43 | 1 | 20 |
| 44 | 1 | 24 |
| 45 | 22 | 23 |
| 46 | 1 | 18 |
| 47 | 21 | 22 |
| 48 | 1 | 21 |
| 49 | 1 | 31 |
| 50 | 1 | 17 |

TABLE XXXIV

CLUSTERING OF THE TARGET ELEMENTS USING COMPLETE LINKAGE WITH THREE VARIABLES

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 2 | 38 | 39 |
| 3 | 40 | 41 |
| 4 | 7 | 9 |
| 5 | 6 | 10 |
| 6 | 12 | 13 |
| 7 | 26 | 27 |
| 8 | 42 | 43 |
| 9 | 35 | 36 |
| 10 | 33 | 34 |
| 11 | 2 | 3 |
| 12 | 12 | 14 |
| 13 | 46 | 47 |
| 14 | 49 | 50 |
| 15 | 7 | 8 |
| 16 | 4 | 5 |
| 17 | 25 | 26 |
| 18 | 46 | 48 |
| 19 | 35 | 37 |
| 20 | 15 | 16 |
| 21 | 1 | 4 |
| 22 | 6 | 7 |
| 23 | 11 | 12 |
| 24 | 42 | 44 |
| 25 | 38 | 40 |
| 26 | 1 | 2 |
| 27 | 46 | 49 |
| 28 | 42 | 46 |
| 29 | 33 | 35 |
| 30 | 24 | 25 |

TABLE XXXIV (continued)

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 31 | 11 | 15 |
| 32 | 42 | 45 |
| 33 | 28 | 38 |
| 34 | 30 | 33 |
| 35 | 1 | 6 |
| 36 | 29 | 42 |
| 37 | 1 | 11 |
| 38 | 19 | 20 |
| 39 | 29 | 32 |
| 40 | 28 | 30 |
| 41 | 22 | 23 |
| 42 | 18 | 19 |
| 43 | 28 | 29 |
| 44 | 21 | 22 |
| 45 | 24 | 28 |
| 46 | 1 | 18 |
| 47 | 21 | 31 |
| 48 | 1 | 17 |
| 49 | 21 | 24 |
| 50 | 1 | 21 |

TABLE XXXV

CLUSTERING OF THE TARGET ELEMENTS USING WEIGHTED AVERAGE LINKAGE WITH THREE VARIABLES

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 2 | 38 | 39 |
| 3 | 40 | 41 |
| 4 | 7 | 9 |
| 5 | 6 | 10 |
| 6 | 12 | 13 |
| 7 | 26 | 27 |
| 8 | 42 | 43 |
| 9 | 35 | 36 |
| 10 | 33 | 34 |
| 11 | 6 | 7 |
| 12 | 12 | 14 |
| 13 | 2 | 3 |
| 14 | 46 | 47 |
| 15 | 49 | 50 |
| 16 | 25 | 26 |
| 17 | 4 | 5 |
| 18 | 46 | 48 |
| 19 | 11 | 12 |
| 20 | 35 | 37 |
| 21 | 6 | 8 |
| 22 | 38 | 40 |
| 23 | 42 | 44 |
| 24 | 1 | 2 |
| 25 | 1 | 4 |
| 26 | 15 | 16 |
| 27 | 46 | 49 |
| 28 | 33 | 35 |
| 29 | 42 | 46 |
| 30 | 11 | 15 |

TABLE XXXV (continued)

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 31 | 42 | 45 |
| 32 | 24 | 25 |
| 33 | 33 | 38 |
| 34 | 6 | 11 |
| 35 | 1 | 6 |
| 36 | 29 | 42 |
| 37 | 28 | 32 |
| 38 | 30 | 33 |
| 39 | 28 | 29 |
| 40 | 28 | 30 |
| 41 | 19 | 20 |
| 42 | 22 | 23 |
| 43 | 18 | 19 |
| 44 | 24 | 28 |
| 45 | 21 | 22 |
| 46 | 1 | 18 |
| 47 | 21 | 24 |
| 48 | 1 | 17 |
| 49 | 21 | 31 |
| 50 | 1 | 21 |

TABLE XXXVI

CLUSTERING OF THE TARGET ELEMENTS USING CENTROID LINKAGE WITH THREE VARIABLES

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 2 | 38 | 39 |
| 3 | 40 | 41 |
| 4 | 7 | 9 |
| 5 | 6 | 10 |
| 6 | 12 | 13 |
| 7 | 26 | 27 |
| 8 | 6 | 7 |
| 9 | 42 | 43 |
| 10 | 35 | 36 |
| 11 | 33 | 34 |
| 12 | 12 | 14 |
| 13 | 2 | 3 |
| 14 | 46 | 47 |
| 15 | 49 | 50 |
| 16 | 25 | 26 |
| 17 | 2 | 4 |
| 18 | 46 | 48 |
| 19 | 11 | 12 |
| 20 | 35 | 37 |
| 21 | 6 | 8 |
| 22 | 1 | 5 |
| 23 | 1 | 2 |
| 24 | 38 | 40 |
| 25 | 42 | 44 |
| 26 | 15 | 16 |
| 27 | 46 | 49 |
| 28 | 42 | 46 |
| 29 | 33 | 35 |
| 30 | 11 | 15 |

TABLE XXXVI (continued)

| Stage | Target Element | Joined Target Element |
|---|---|---|
| 31 | 42 | 45 |
| 32 | 24 | 25 |
| 33 | 33 | 38 |
| 34 | 6 | 11 |
| 35 | 1 | 6 |
| 36 | 29 | 42 |
| 37 | 28 | 32 |
| 38 | 30 | 33 |
| 39 | 28 | 29 |
| 40 | 28 | 30 |
| 41 | 19 | 20 |
| 42 | 22 | 23 |
| 43 | 18 | 19 |
| 44 | 24 | 28 |
| 45 | 21 | 22 |
| 46 | 1 | 18 |
| 47 | 21 | 24 |
| 48 | 1 | 17 |
| 49 | 21 | 31 |
| 50 | 1 | 21 |

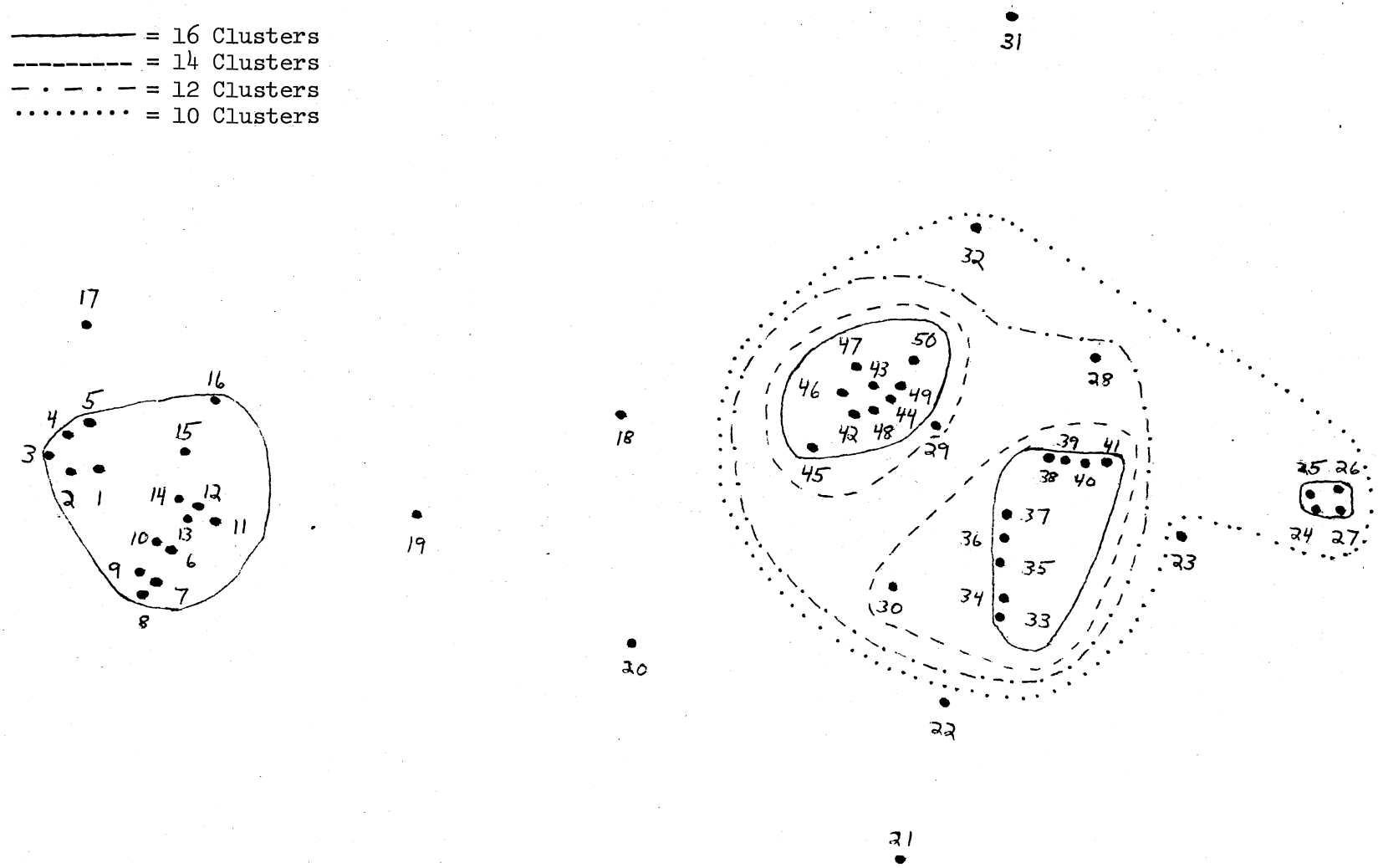


Figure 14. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Single Linkage Algorithm Using Three Variables

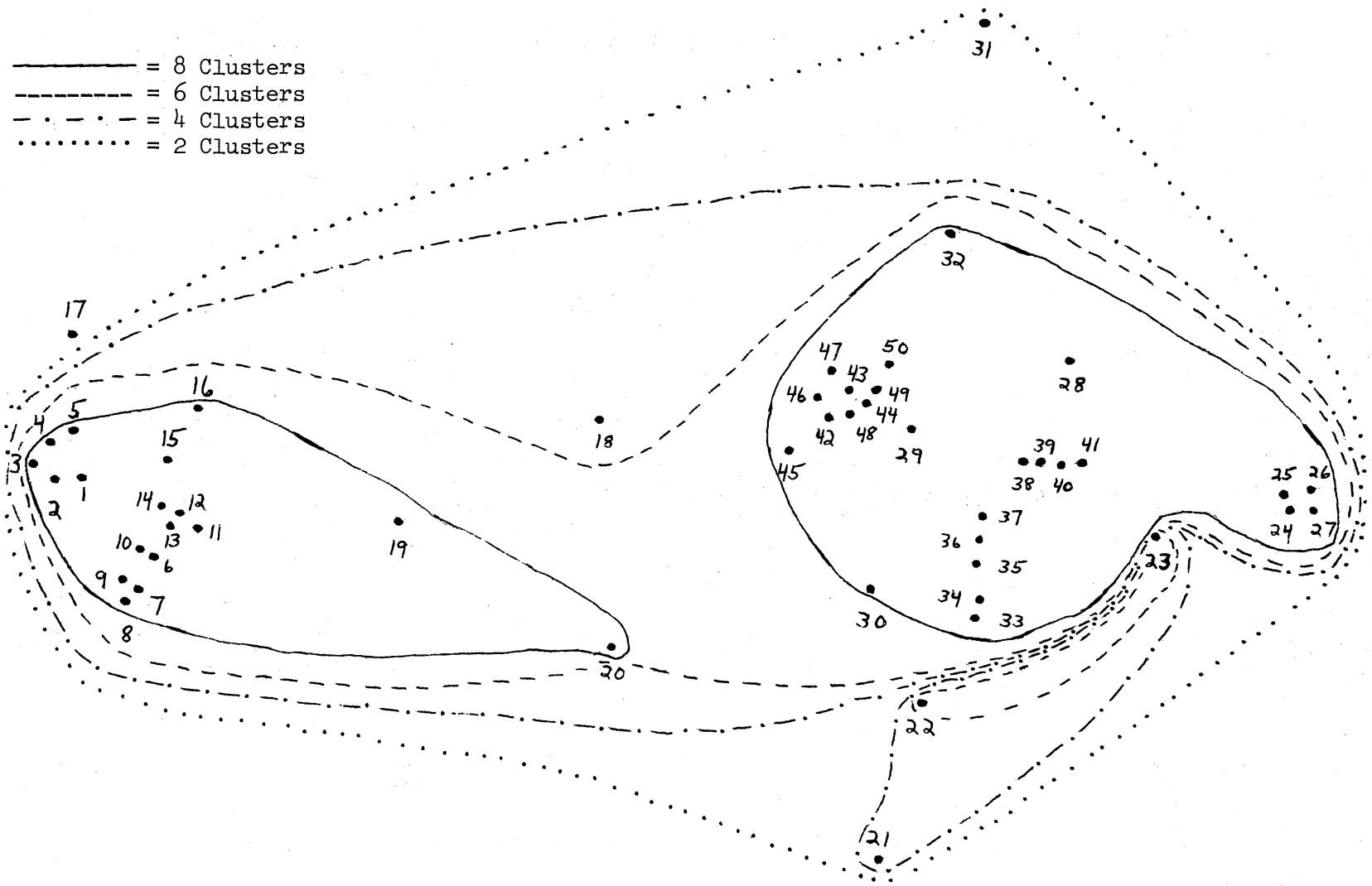


Figure 15. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Single Linkage Algorithm Using Three Variables

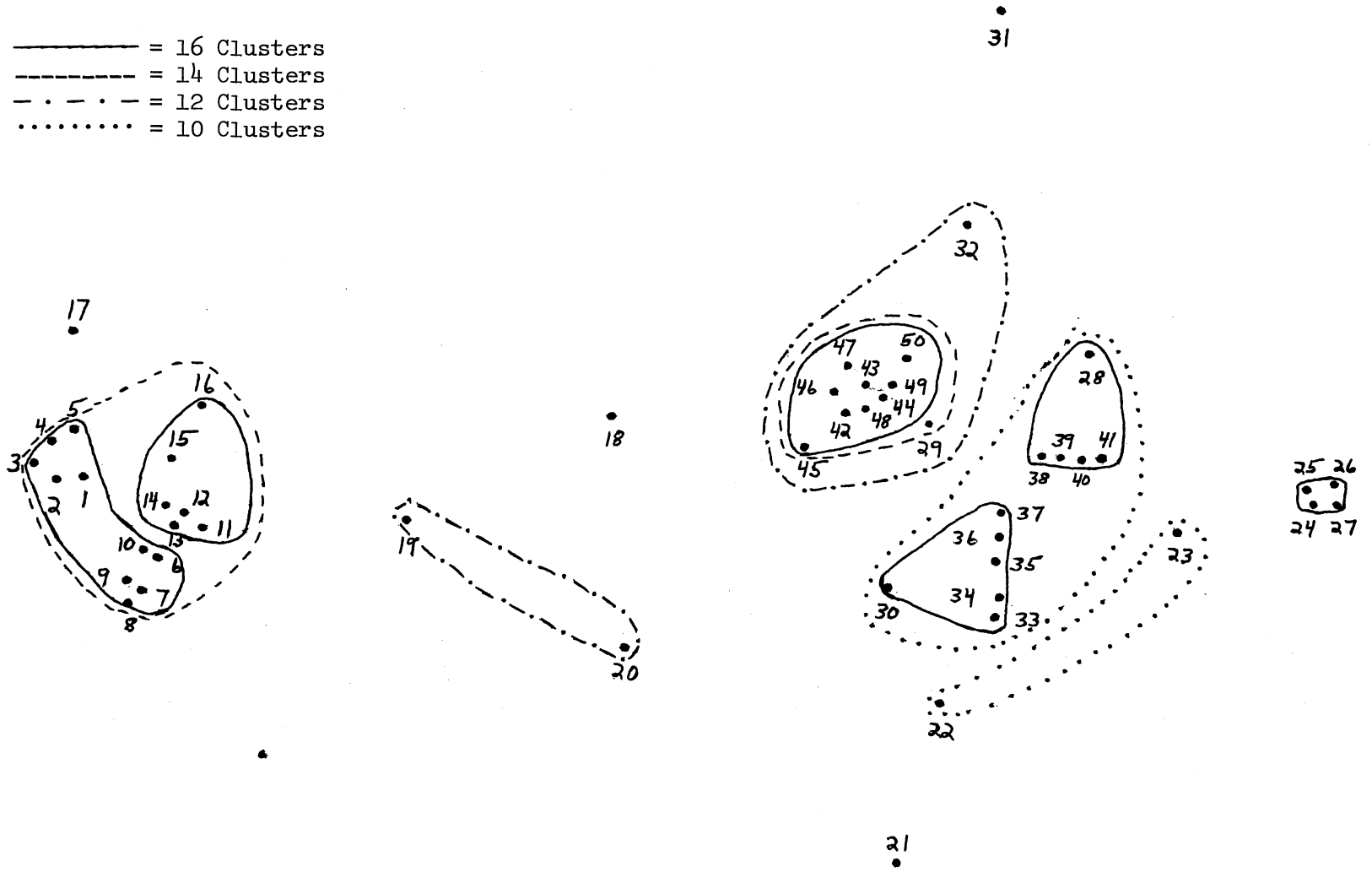


Figure 16. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Complete Linkage Algorithm Using Three Variables

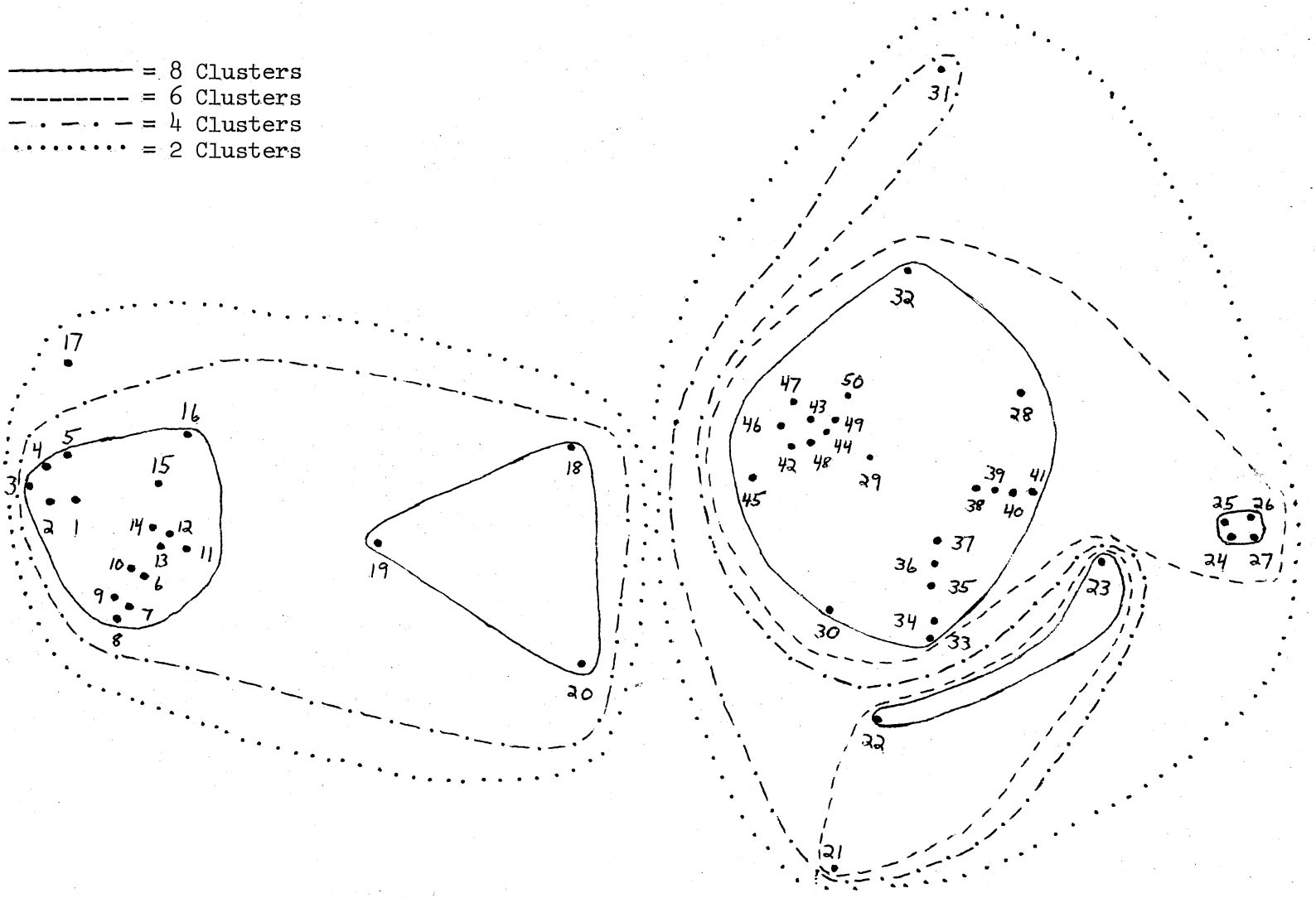


Figure 17. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Complete Linkage Algorithm Using Three Variables

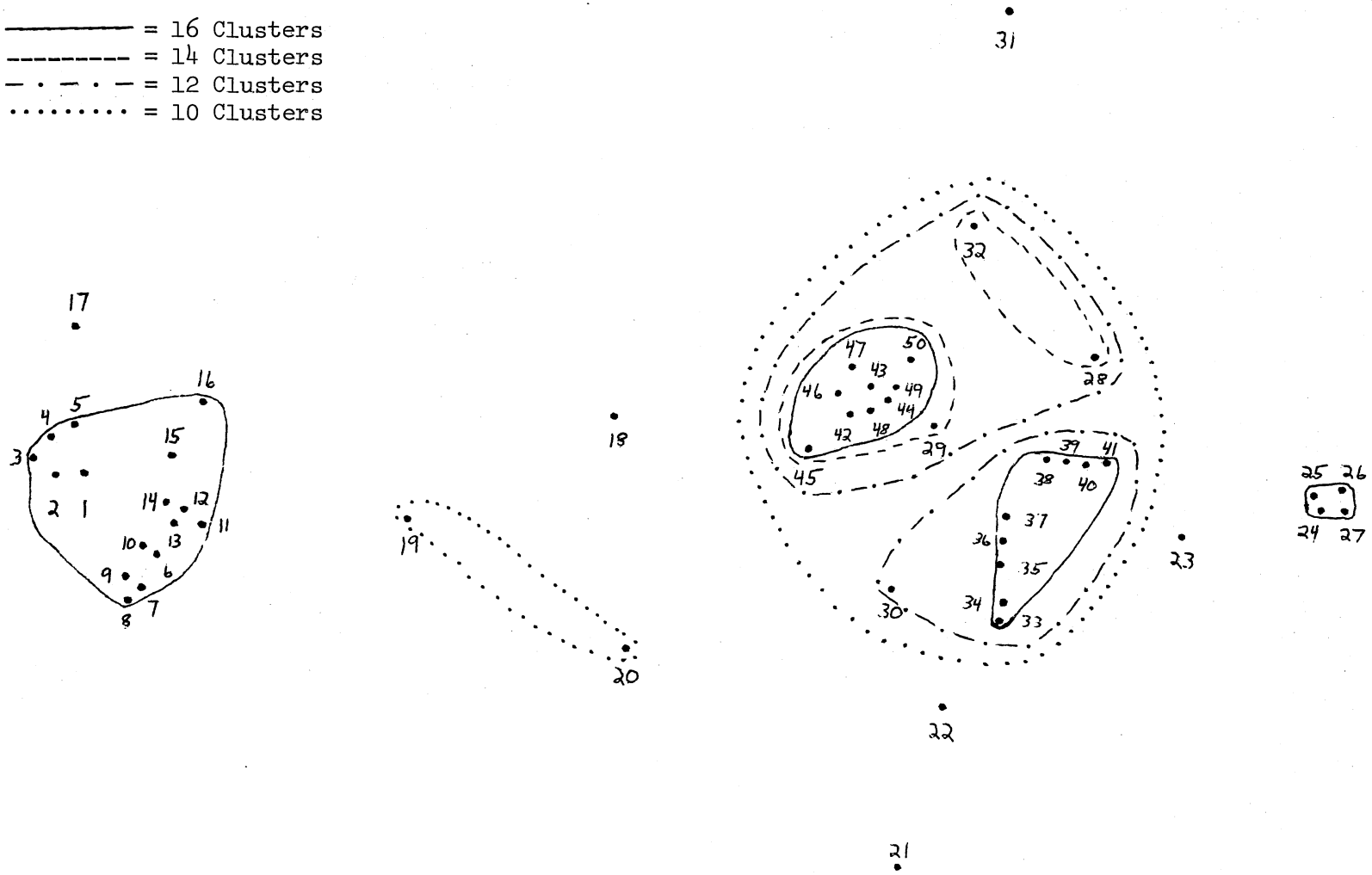


Figure 18. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Weighted Average Linkage Algorithm Using Three Variables

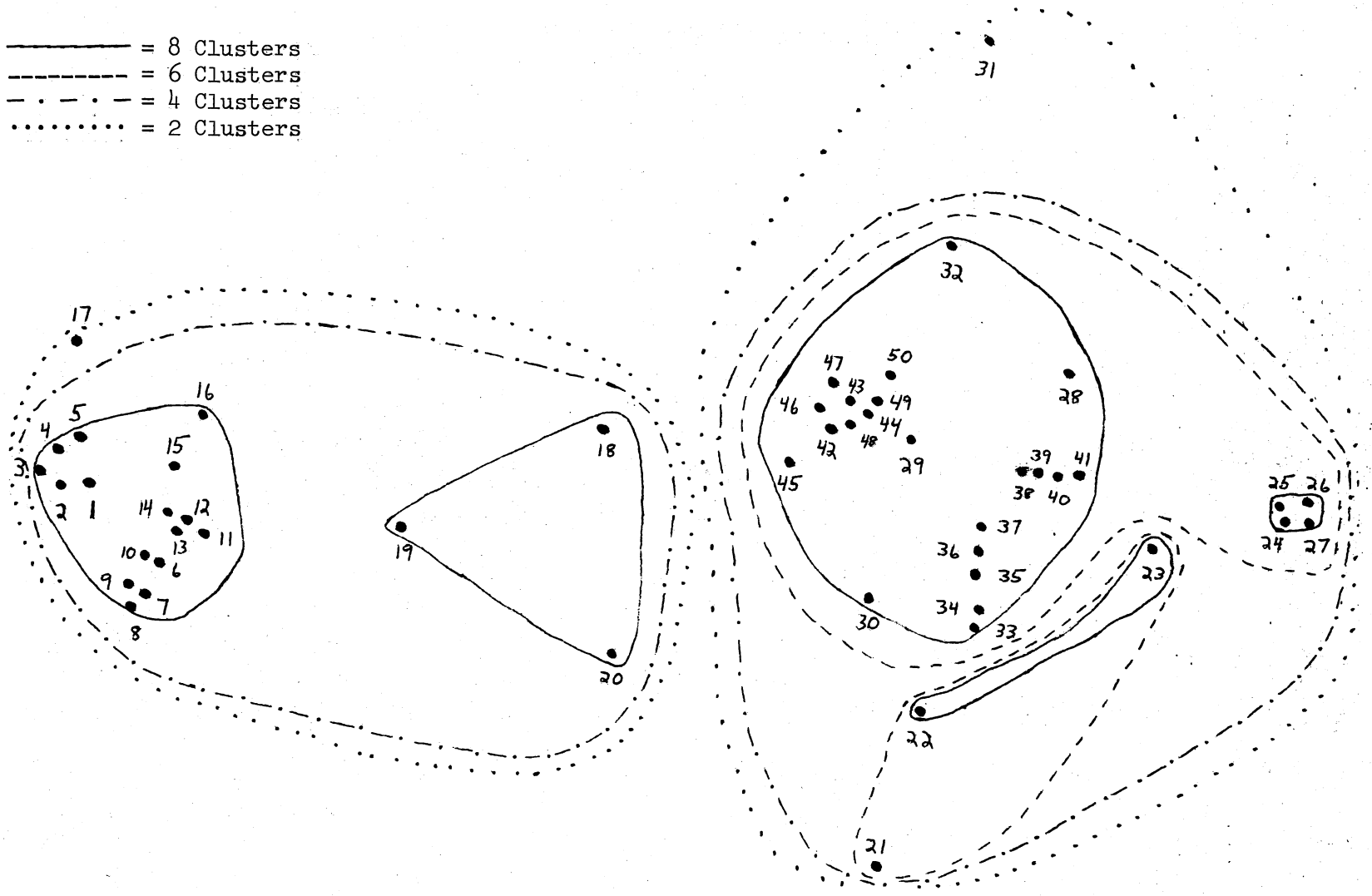


Figure 19. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Weighted Average Linkage Algorithm Using Three Variables

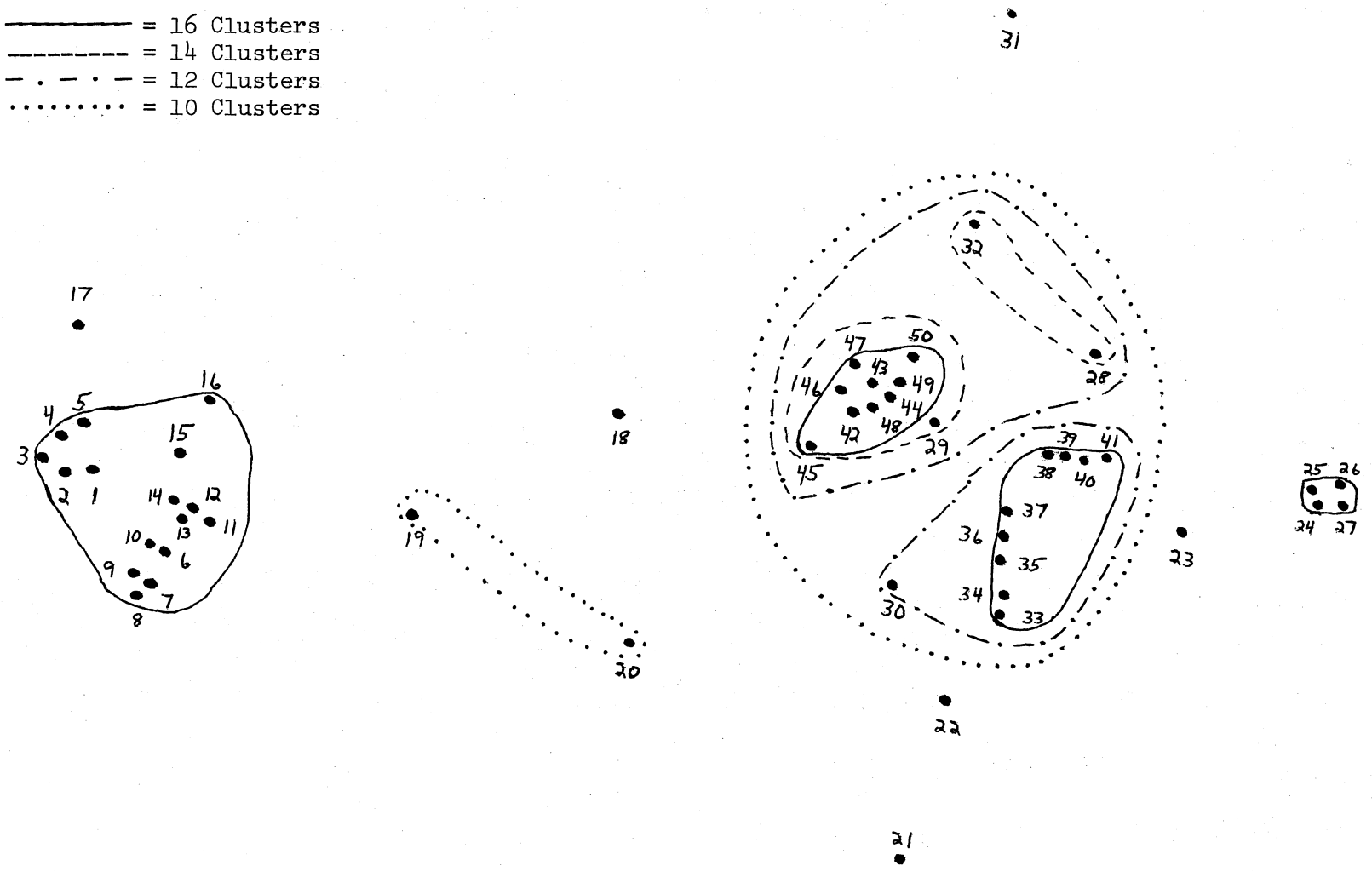


Figure 20. Assignment of Target Elements Into: 10, 12, 14, 16 Clusters by the Centroid Linkage Algorithm Using Three Variables

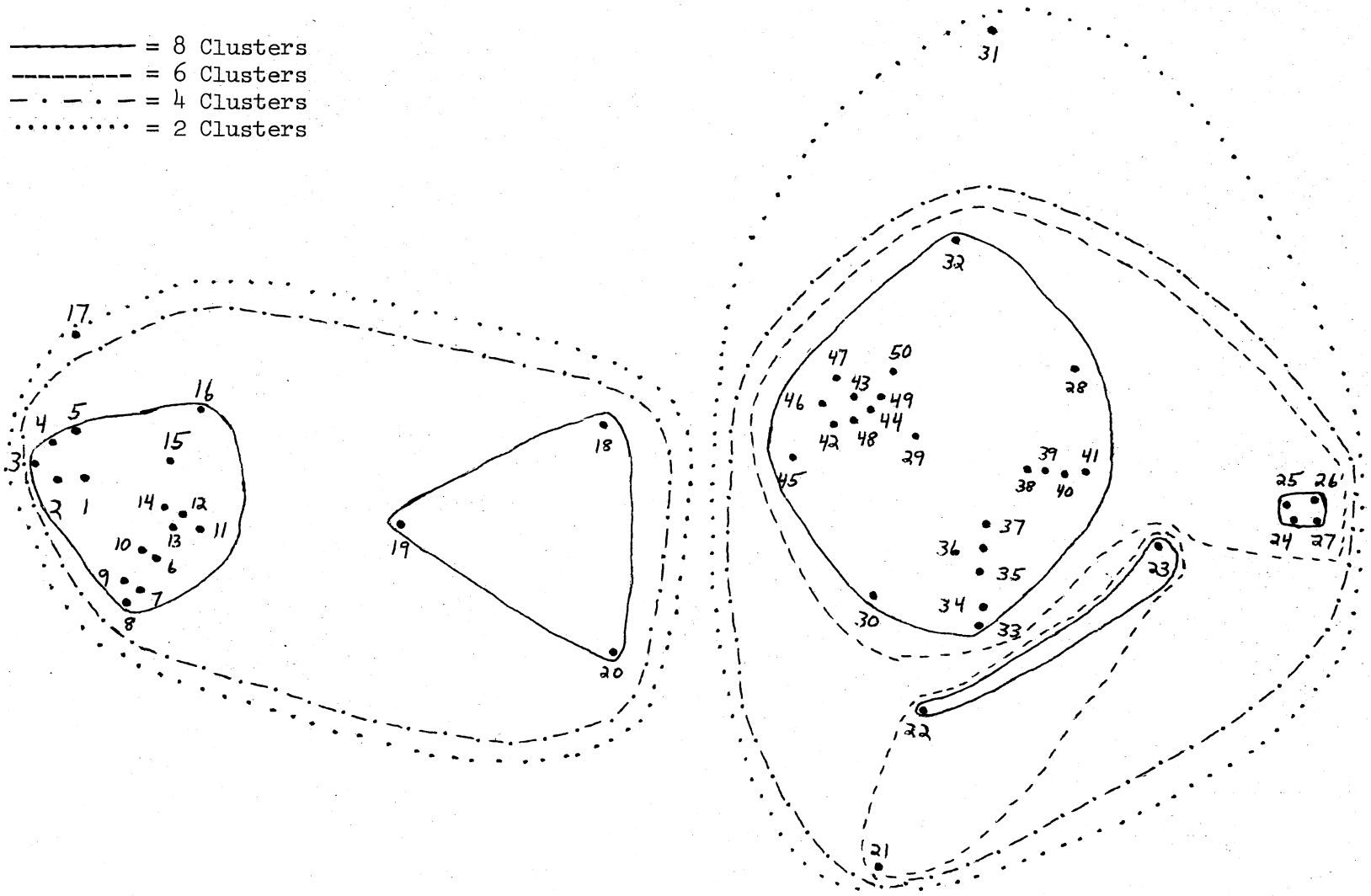


Figure 21. Assignment of Target Elements Into: 2, 4, 6, 8 Clusters by the Centroid Linkage Algorithm Using Three Variables

33-41 to the same cluster except the complete linkage algorithm. The reason for the change from the two variable clustering is that the elements 33-41 all have about the same importance and hence are relatively closer together than in the two variable case. When $k = 4$, the problem of having very important but very distant target elements grouped together is illustrated. The same general conclusions, about the types of clustering produced by the algorithms, are also true in the three variable cases.

There are some desirable features of using the importance measure as a third variable. The single linkage algorithm assigns many of the important target elements to single element clusters, where they receive maximum attention. Another advantage is that target elements of the same kind are grouped together. Thus when the differential effect of weapons on target elements is considered in more general models, elements which can be damaged by the same weapon will be grouped together.

There are some undesirable features of using the importance measure as a third variable. The $f_i$'s no longer have a physical interpretation, which seems to make the formulation of a good stopping rule impossible. As mentioned previously some clusters are formed which contain target elements of similar importance but which are so distant that a pattern could not be expected to cover them all.

The question arises: Is it necessary to use the importance as a variable to assign target elements to clusters? It could be argued that important target elements are "naturally" isolated when an airfield complex is constructed. It could also be argued that similar elements of moderate importance would tend to be grouped together because of

similar strategic functions and for convenience. It would be reasonable to expect groups of aircraft to be located near runways, etc. Consideration of these arguments has lead to the strategies suggested in a previous section.

The strategies proposed use the physical coordinates, and the pattern size and shape to cluster the target elements. Then they use the importance measure to select the most important clusters, if a pass cannot be made at each cluster. The strategies that were proposed were tested on the data for six different patterns. Three of the patterns were circles with diameters $\sqrt{37}$, 9, $\sqrt{208}$. The remaining patterns were rectangles whose dimensions were $\sqrt{634} \times \sqrt{10}$, $\sqrt{37} \times \sqrt{2}$, $\sqrt{53} \times \sqrt{17}$.

For the circular patterns the single linkage algorithm is allowed to group elements together until the maximum number of elements in any cluster multiplied by the value of $f_i$ is greater than the square of the diameter of the circular pattern. Figures 22 through 24 visually display the resulting clusters for these three patterns. The stopping rule used by this procedure is somewhat conservative. The clustering is stopped in some cases when the pattern is still large enough to cover all the elements in any cluster at several additional stages.

The complete linkage algorithm is allowed to group elements together until $f_i$ is larger than the square of the diameter of the circular pattern. Figures 25 through 27 visually display the clusters for these three patterns.

For the rectangular patterns, the single linkage algorithm is allowed to group elements together until the maximum number of elements in any cluster multiplied by the value of $f_i$ is greater than the square of the shorter side of the rectangle. The complete linkage algorithm

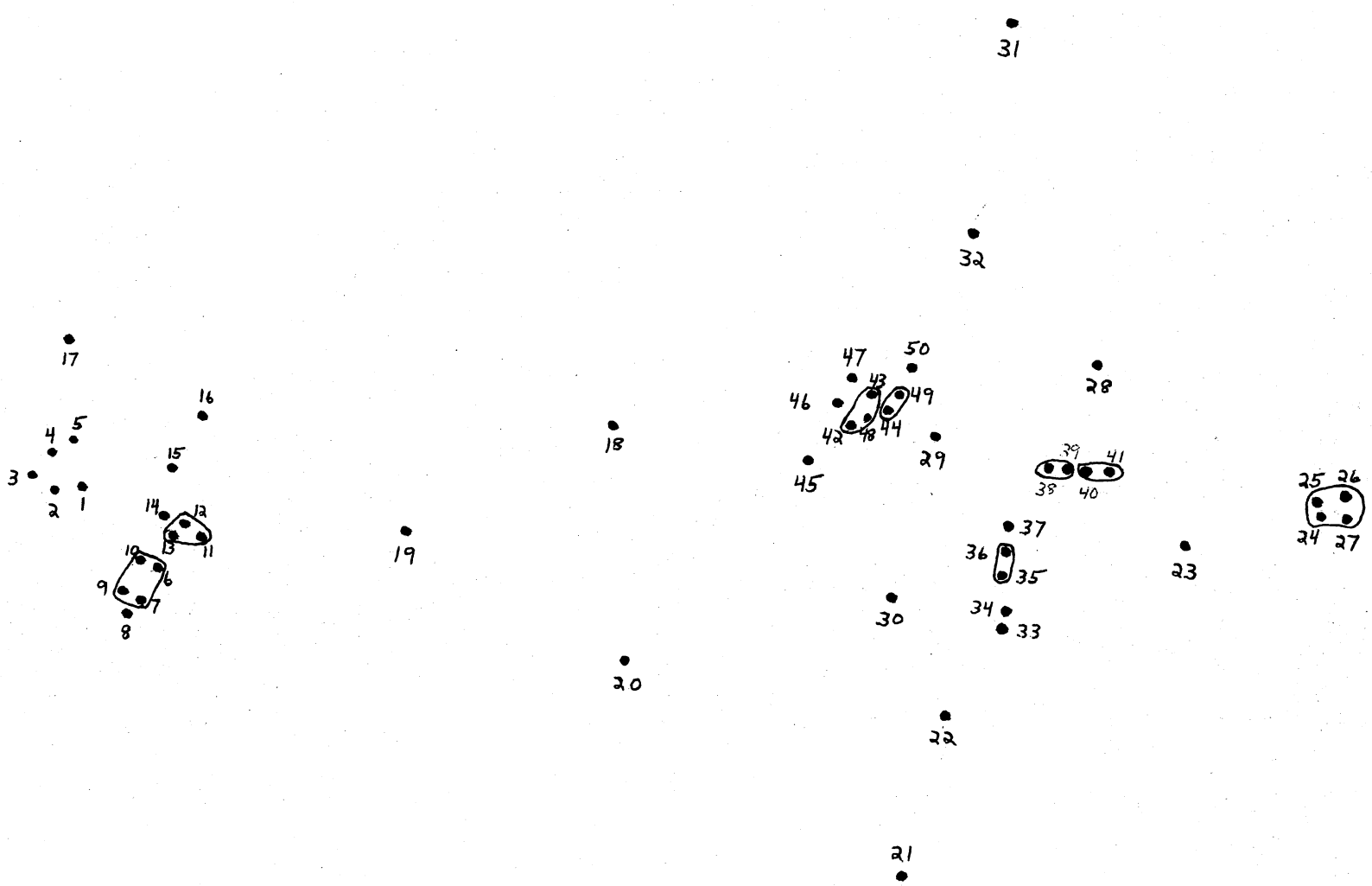


Figure 22. The Clusters Produced by Using the Stopping Rule with Single Linkage and the $\sqrt{37}$ Unit Diameter Circular Pattern

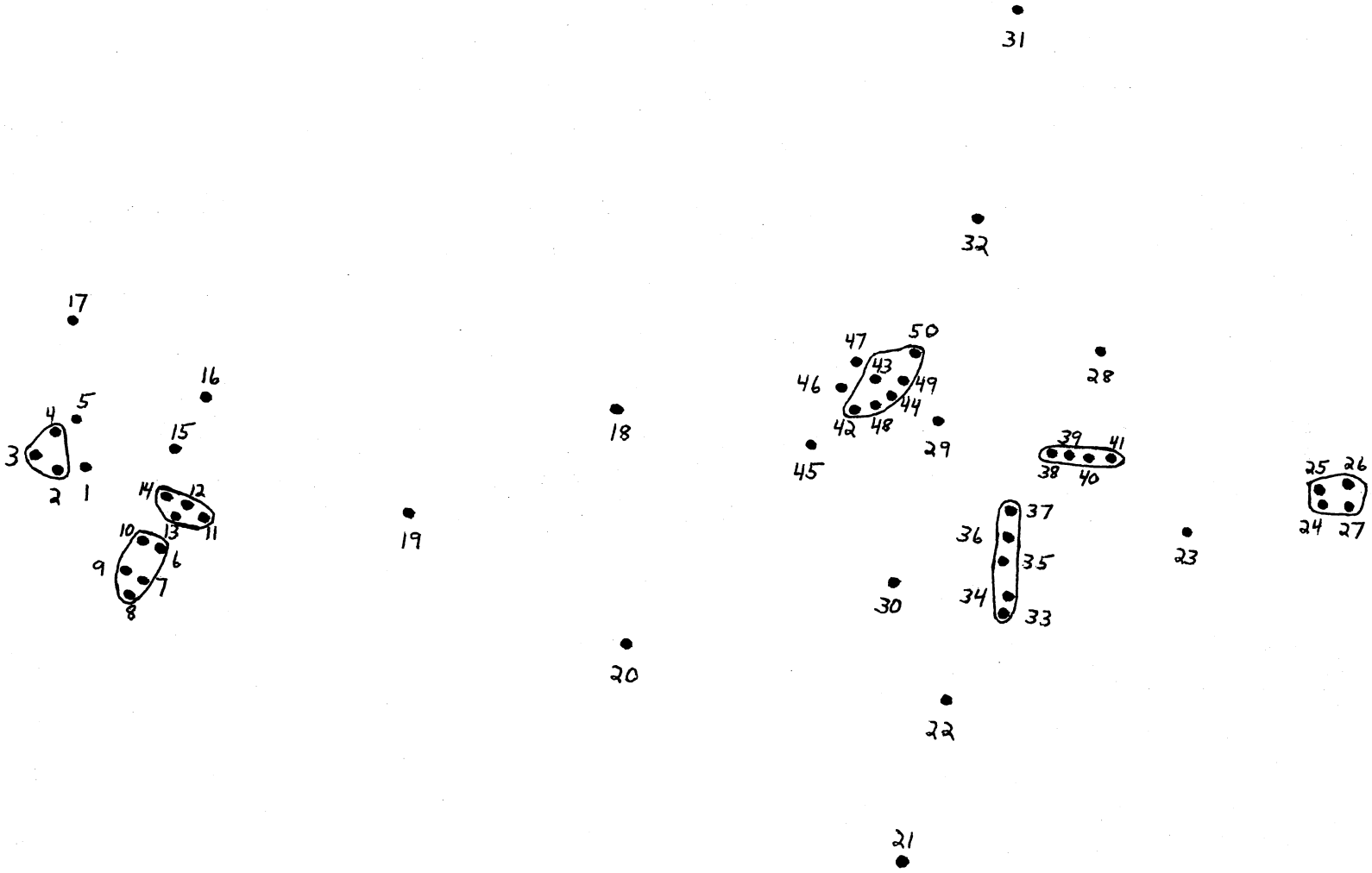


Figure 23. The Clusters Produced by Using the Stopping Rule with Single Linkage and the 9 Unit Diameter Circular Pattern

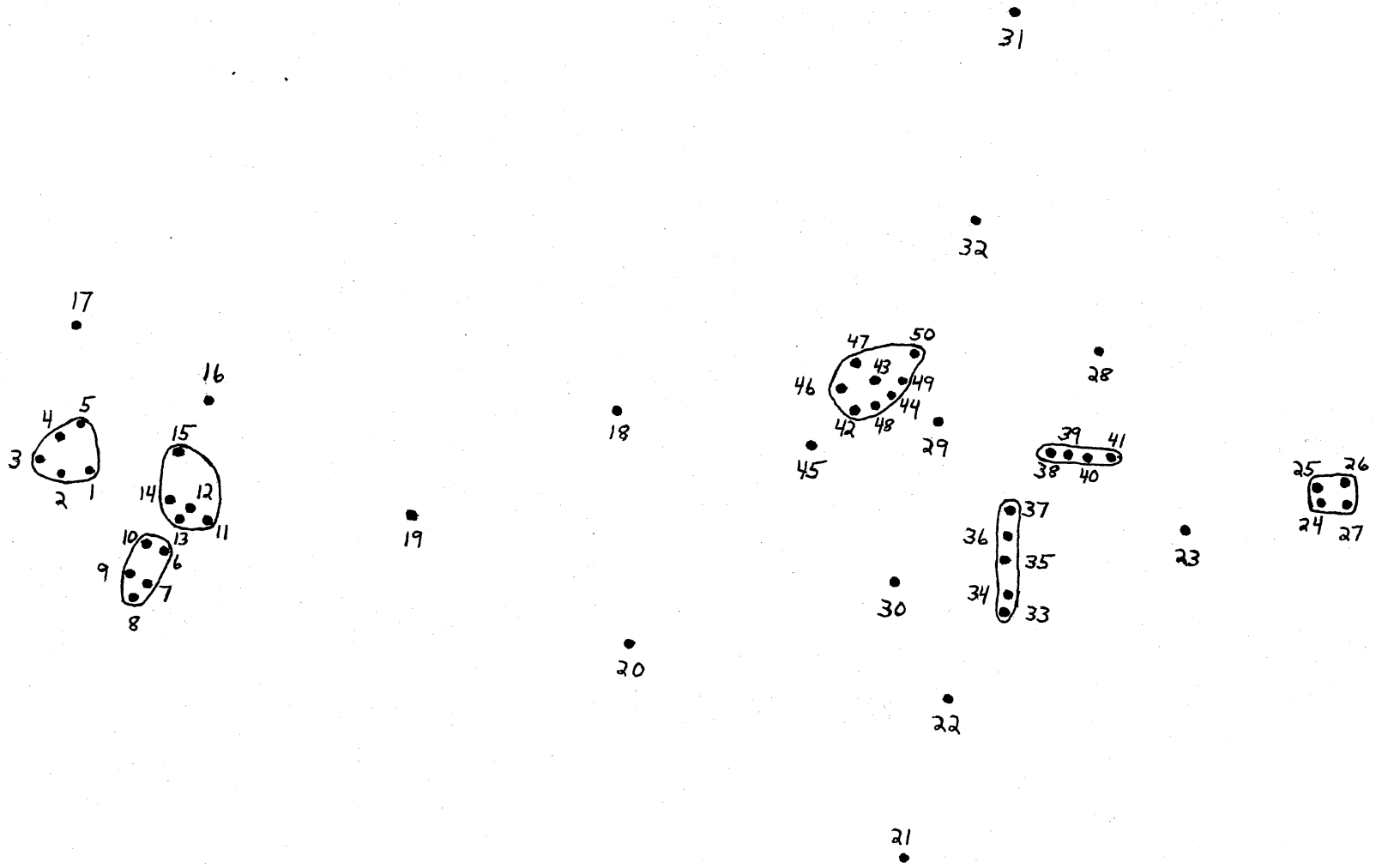


Figure 24. The Clusters Produced by Using the Stopping Rule with Single Linkage and the $\sqrt{208}$ Unit Diameter Circular Pattern

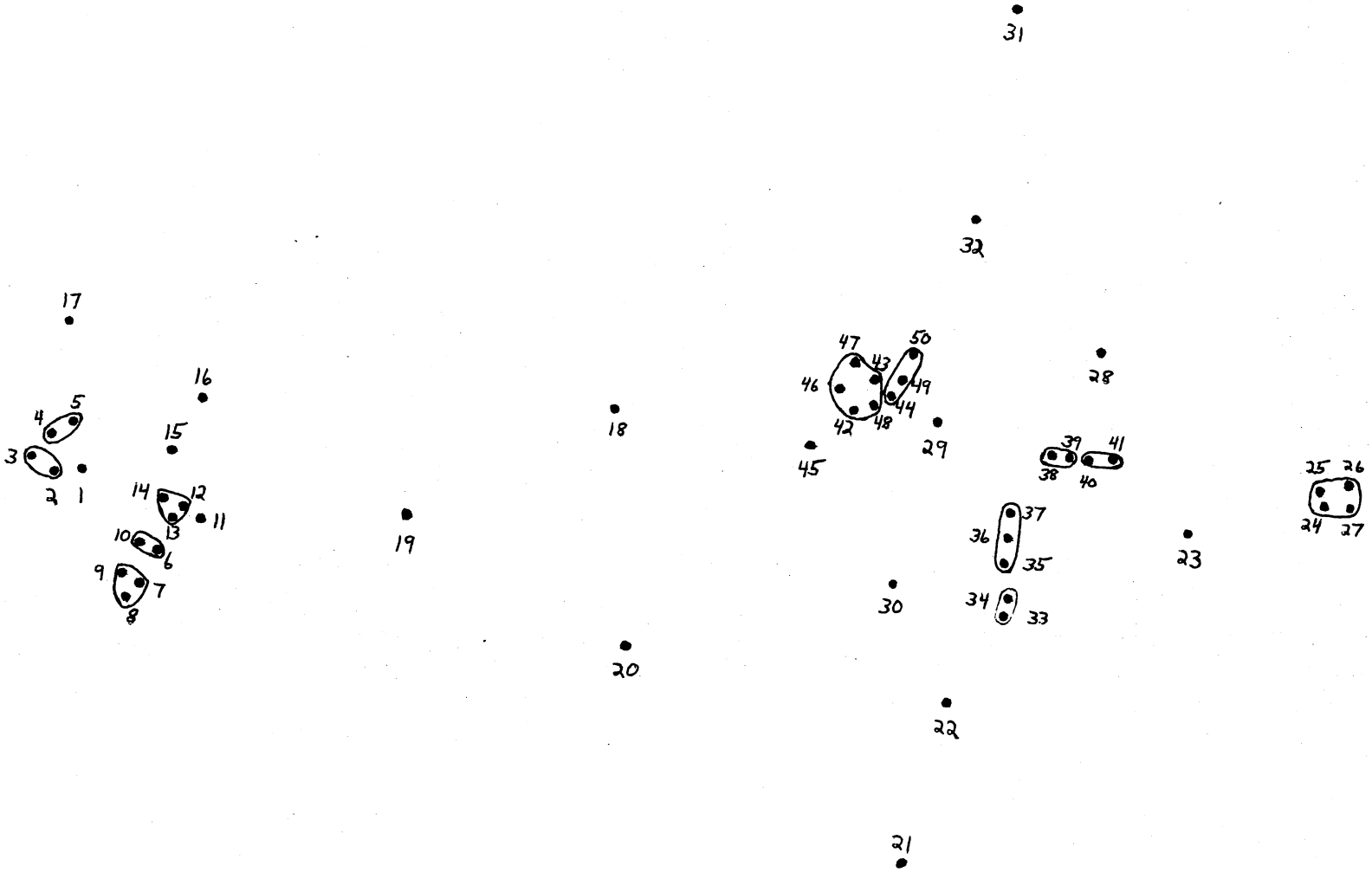


Figure 25. The Clusters Produced by Using the Stopping Rule with Complete Linkage and the $\sqrt{37}$ Unit Diameter Circular Pattern

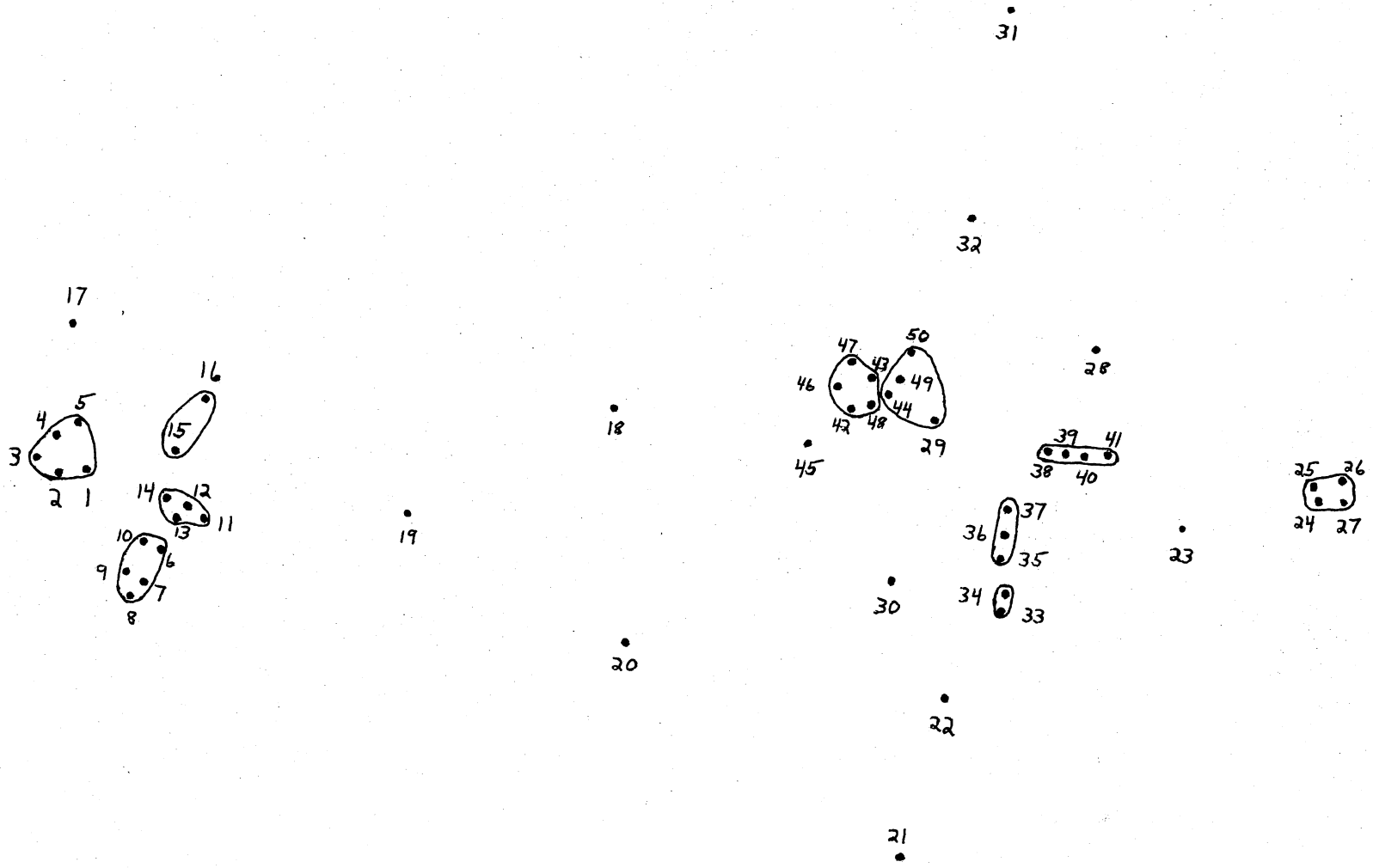


Figure 26. The Clusters Produced by Using the Stopping Rule with Complete Linkage and the 9 Unit Diameter Circular Pattern

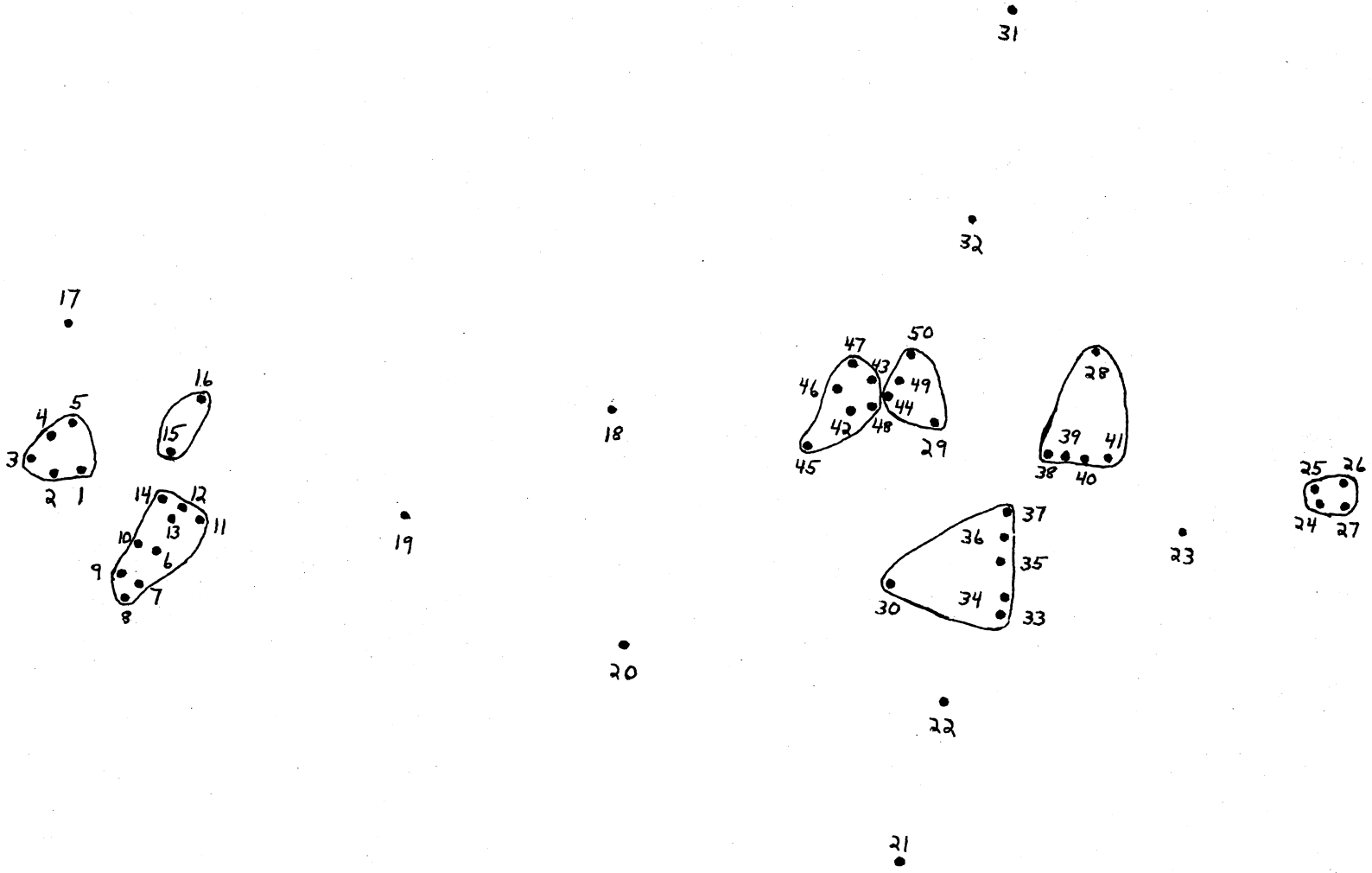


Figure 27. The Clusters Produced by Using the Stopping Rule with Complete Linkage and the $\sqrt{208}$ Unit Diameter Circular Pattern

is allowed to group elements together until $f_i$ is larger than the square of the shorter side of the rectangle.

The results were not very encouraging for the rectangular patterns, due primarily to the algorithm's failure to produce many elongated clusters. Further research is needed to determine if any other algorithms, such as Ling's (1971), would be more useful for rectangular patterns.

The conclusion for circular patterns is that the complete linkage algorithm appears to be the best algorithm and that the clusters of target elements it produces are quite reasonable. It has not yet been established whether or not the complete linkage technique will suggest "better" aim points at a more reasonable cost than would present methods. Further research has been planned to determine the "optimality" of this procedure compared to the present methods. The basis for comparison could be a criterion such as expected fractional coverage.

# CHAPTER VI

# POSSIBLE EXTENSIONS

## Additional Tables of Percentage Points

The present study has indicated that the use of tests based on the clustering suggested by various sequential agglomerative algorithms compares favorably with the Engelman-Hartigan (1969) test. In order to make these tests more useful in practice, additional tables of percentage points are needed for many different sample sizes. It would be desirable to have these additional tables based on the generation of very large numbers of data sets, so that the standard errors of the estimates would compare favorably with the Engelman-Hartigan tables, which are based on 100,000 data sets. The major problem with generating additional percentage points is the enormous amount of computer time required.

## Some Possible Sequential Test Procedures

There are at least two logical sequential test procedures which could be used in conjunction with the tests either reviewed or proposed previously, when the number of clusters in the alternative is unknown. The first procedure will be referred to as the "changing null hypothesis procedure", for reasons which will become obvious. At stage $n - 1$, the null hypothesis that all the observations are from the same normal population is tested against the two cluster alternative, using the Engelman-Hartigan test or the tests based on the agglomerative

algorithms, whichever is appropriate. If the null hypothesis is rejected, then at stage $n - 2$ each of the resulting clusters, which have at least three observations, is tested by recomputing the test statistic that was used at the previous stage. The test statistic is computed from the assignment of observations to clusters at stage $n - 2$. The sample size is smaller and the statistic is referred to a different table of critical values. This process is continued until all the "clusters" cannot be further divided. This procedure is much the same as many multiple comparison procedures, and is subject to the same type of errors. Murphy (1973) gives a detailed discussion of the different types of errors a sequential multiple comparison procedure may incur.

Another procedure will be referred to as the "changing alternative hypothesis" procedure. For this procedure, at stage $n - 1$, the null hypothesis that all the observations are from a single normal population is tested against the two cluster alternative. The ratio $B/W$ is computed from the observations assigned to clusters by one of the agglomerative procedures and then $B/W$ is referred to the proper table of critical values. If the null hypothesis is rejected then the ratio $B/W$ is computed at stage $n - 2$ from the algorithm's assignment of observations to three clusters; $B/W$ is referred to the critical point for the three cluster alternative. The process is repeated until the null hypothesis can no longer be rejected, or until all the possible alternatives have been tested. The procedure just considered began by testing the two cluster alternative first, and then conditional on a rejection of the two cluster alternative, the three cluster alternative was considered, etc. But there is no compelling reason for this to be the proper order; in fact, there are $(n - 2)!$ possible orders in which

this sequential test might be run. It is not known which order or orders are the best. A study of these sequential procedures appears to be feasible, but the investigation would probably have to be done on an empirical basis.

## Additional Investigation of the Relationship Between the Normal Mixtures Problem and Clustering

All of the test procedures that are based on the agglomerative algorithms have been investigated with respect to the Engelman and Hartigan (1969) alternative, that the number of observations from each population is a fixed but unknown constant. These same test procedures may also be adequate for the Lee (1974) alternative, that the number of observations from each population is random. The probability distribution of these random variables is determined by $k - 1$ unknown "mixing" parameters. Lee (1974) derived a test criterion for this alternative, but did not generate percentage points for its distribution. The generation of percentage points for this test would serve several purposes; first the test could be carried out in practice, although like the Engelman and Hartigan test it would require relatively large amounts of computer time to be performed. Second, the power of this procedure could be investigated for various separations of the means and for various values of the mixing parameters. After the power of the Lee (1974) test has been determined, then the power of the agglomerative algorithm based procedures could be estimated and compared with the power of the Lee (1974) test.

The Lee (1974) formulation of the likelihood suggests a new method, through clustering, of obtaining estimates for the parameters in the normal mixture problem. There appears to be a good chance to empirically estimate the distributions of the estimators. It is the author's opinion that further investigation of this approach would lead to advances in both the clustering and normal mixtures problems.

## Extension to Multivariate Cases

The generalization of the univariate results to multivariate cases is not immediate. One of the first problems is the increase in the number of unknown parameters in both the null and alternative hypotheses. No good tests were discovered in the univariate case for alternatives having more unknown parameters than the number of observations available (the partition is not included as an unknown parameter in this counting). For the bivariate normal case there are 2 unknown parameters in each mean vector and 3 unknown parameters in the covariance matrix, assuming the same covariance structure for each observation. By analogy, it may be expected that "good" tests will exist only for 2, 3, ..., $\left[\frac{n-3}{2}\right]$ population alternatives.

One reasonable test procedure (at least theoretically) for the univariate test of the two population alternative was given by maximum B/W, or maximum F, where the maximum is over all possible partitions of the observations into two clusters. By analogy it may be that a good test will be given by the maximum of one of the standard multivariate test criteria (see Everitt, 1974), over all partitions of the data into two clusters. There are at least two problems with this approach; first, which one of the multivariate test criteria should be

chosen, and secondly in multivariate cases the observations cannot be ranked as univariate observations can, so many more partitions would need to be searched. Procedures which were expensive and inconvenient to work with in the univariate case often become totally impossible in multivariate cases.

It is anticipated that the "good" tests derived by theoretical considerations will not be practical to use, so the sequential agglomerative procedures may be used as an adequate approximation to the clustering suggested by the theoretical tests. Tables of percentage points for these procedures could be generated, and the power of the procedures estimated also by empirical methods.

The amount of computer time required to generate these tables would be very large.

A SELECTED BIBLIOGRAPHY


1. Baker, F. B. and L. J. Hubert. "Measuring the Power of Hierarchal Cluster Analysis." Journal of the American Statistical Association, Vol. 70 (1975), 31-37.

2. Day, N. E. "Estimating the Components of a Mixture of Normal Distributions." Biometrika, Vol. 56 (1969), 463-474.

3. DuBien, J. L. Personal Communication, Oklahoma State University, 1975.

4. Edwards, A. W. F. and L. L. Cavalli-Sforza. "A Method for Cluster Analysis." Biometrics, Vol. 21 (1965), 372-375.

5. Engelman, L. and J. A. Hartigan. "Percentage Points of a Test for Clusters." Journal of the American Statistical Association, Vol. 64 (1969), 1647-1648.

6. Everitt, B. Cluster Analysis, New York: Wiley, 1974.

7. Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, Vol. 7, Part II (1936), 179-188.

8. Fisher, W. D. "On Grouping for Maximum Homogeneity." Journal of the American Statistical Association, Vol. 53 (1958), 789-98.

9. Friedman, H. P. and J. Rubin. "On Some Invariant Criterion for Grouping Data." Journal of the American Statistical Association, Vol. 62 (1967), 1159-1178.

10. Gay, W. L. "Aim Point Selection for Multiple Target Configurations." (Unpublished M. S. thesis, Oklahoma State University, 1974.)

11. Gibson, J. E. Personal Communication, Oklahoma State University, 1975.

12. Gower, J. C. and G. J. S. Ross. "Minimum Spanning Trees and Single Linkage Cluster Analysis." Applied Statistics, Vol. 18 (1969), 54-64.

13. Hartigan, J. A. "Representation of Similarity Matrices by Trees." Journal of the American Statistical Association, Vol. 62 (1967), 1140-58.

14. Lance, G. N. and W. T. Williams. "A General Theory of Classificatory Sorting Strategies: 1. Hierarchal Systems." Computer Journal, Vol. 9 (1967), 373-380.

15. Lee, K. L. "Some Practical Points of a Test for Clusters." (Presented at Joint Statistics Meetings, Tallahassee, Florida, 1974.)

16. Ling, R. F. "Cluster Analysis." (Unpublished Ph.D. thesis, Yale University, 1971.)

17. Mrachek, R. J. "Some Statistical Aspects of Clustering Procedures." (Unpublished M. S. thesis, Iowa State University, 1972.)

18. Murphy, J. R. "Procedures for Grouping a Set of Observed Means." (Unpublished Ph.D. thesis, Oklahoma State University, 1973.)

19. Norton, J. M. and W. D. Warde. "A Comment on F. J. Rohlf's Paper." (Submitted to Biometrics, 1975.)

20. Orloci, L. "An Agglomerative Method for Classification of Plant Communities." Journal of Ecology, Vol. 55 (1967), 193-206.

21. Pearson, K. "Contributions to the Mathematical Theory of Evolution." Phil. Trans. R. Soc. A., 185, (1894), 71-110. (Quoted in paper by N. E. Day.)

22. Rand, W. M. "The Development of Objective Criteria for Evaluating Clustering Methods." (Unpublished Ph.D. thesis, University of California at Los Angeles, 1969.)

23. Rohlf, F. J. "Generalization of the Gap Test for the Detection of Multivariate Outliers." Biometrics, Vol. 27 (1971), 387-398.

24. Rubin, J. "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem." Journal of Theoretical Biology, Vol. 15 (1967), 103-144.

25. Scott, A. J. and M. J. Symons. "Clustering Methods Based on Likelihood Ratio Criteria." Biometrics, Vol. 27 (1971), 387-398.

26. Searle, S. R. Linear Models. New York: Wiley, 1971.

27. Sneath, P. H. A. "The Application of Computers to Taxonomy." Journal of General Microbiology, Vol. 17 (1957), 201-226.

28. Sneath, P. H. A. and R. R. Sokal. Numerical Taxonomy. San Francisco: Freeman, 1973.

29. Snow, R. and M. Ryan. "A Simplified Weapons Evaluations Model." Memorandum RM-5677-1-PR. Santa Monica, California: Rand Corp., 1970.

30. Sokal, R. R. and C. D. Michener. "A Statistical Method for Evaluating Systematic Relationships." University of Kansas Science Bulletin, Vol. 38 (1958), 1409-38.

31. Sokal, R. R. and P. H. A. Sneath. Principles of Numerical Taxonomy. San Francisco: Freeman, 1963.

32. Tukey, J. W. "Data Analysis and Behavioral Science." (Unpublished manuscript, 1962.)

33. Ward, J. H. "Hierarchal Grouping to Optimize an Objective Function." Journal of the American Statistics Association, Vol. 58 (1963), 236-44.

34. Warde, W. D. (Unpublished manuscript, Oklahoma State University, 1975.)

35. Williams, W. T. and W. B. Dale. "Fundamental Problems in Numerical Taxonomy." Advances in Botanical Research, Vol. 2 (1965), 35-68.

# APPENDIX

Generation of percentage points for null distributions or estimates of power will require modifications of the parameters in these programs. The first program generates percentage points or estimates of power for the four agglomerative procedures. The parameters in this program will be briefly described.

N is the sample size.

M is the number of replications.

IX and JX are random, odd, six-digit integers to start the normal generator.

POWER is the separation of the two normal populations which have unit variances.

If N is larger than 10, modifications must be made in all the array sizes. The loop on line 24 must be modified to generate the desired partition into 2 groups.

There are four subroutines in this program and their functions will be briefly described.

BUTLER generates random normal (0, 1) variables.

EUCLID computes a vector of pairwise squared Euclidean distances from the generated data.

MESA carries out the clustering using the single, complete, weighted average and centroid linkage methods.

F computes an F value for each stage of the clustering as long as that value is less than or equal to 999,999. If the F value is greater than 999,999 then the subroutine will report the value as 999,999.

The second program generates data sets to be used to estimate the power of the Engelman and Hartigan test. The parameters are the same

as those previously used except the loop on line 14 must be modified to generate the desired partition into two groups.

```
CARD
   1        DIMENSION V(10,10),W(10,10),D(100),V1(20)
   2        DIMENSION V2(20)
   3        DIMENSION DDD(32)
   4        COMMON CF,STOT
   5        REAL*8 W
   6        KKK=1
   7        NEND=KKK+1
   8        JX=689633
   9        N=10
  10        NS=N*(N-1)/2
  11        L=4
  12        NN=4
  13        NN1=NN-1
  14        K=1
  15        M=500
  16        POWER=4.
  17        IX=515167
  18        DO 90 IJK=1,M
  19        DO 10 I=1,N
  20        DO 10 J=1,K
  21        CALL BUTLER (KKK,W(I,J),IX,JX,NEND)
  22   10   CONTINUE
  23        DO 11 I=1,K
  24        DO 11 J=1,K
  25        W(I,J)=W(I,J)+POWER
  26   11   CONTINUE
  27        IKJ=1
  28        DELTA=0.
  29        DO 12 I=1,N
  30        DO 12 J=1,K
  31      12 V(I,J)=W(I,J)
  32   13   CONTINUE
  33        DO 25 I=1,N
  34        V1(I)=W(I,1)
  35   25   CONTINUE
  36        CF=0.
  37        DO 26 I=1,N
  38        CF=CF+V1(I)
  39   26   CONTINUE
  40        CF=CF*CF/N
  41        STOT=0.
  42        DO 27 I=1,N
  43        STOT=STOT+V1(I)*V1(I)
  44   27   CONTINUE
  45        STOT=STOT-CF
  46        CALL EUCLID(N,K,V,D)
  47        CALL MESA(N,D,V1)
  48      90 CONTINUE
  49   100  CONTINUE
  50        STOP
  51        END
  52        SUBROUTINE EUCLID(N,K,V,D)
  53        DIMENSION V(10,10),D(100)
  54        NR=N*(N-1)/2
```

```
CARD
   55        DO 5 I=1,NR
   56      5 D(I)=0.
   57        II=0
   58        DO 300 I=2,N
   59        IJ=I-1
   60        DO 300 J=1,IJ
   61        II=II+1
   62        DO 300 JJ=1,K
   63    300 D(II)=D(II)+(V(I,JJ)-V(J,JJ))*(V(I,JJ)-V(J,JJ))
   64        RETURN
   65        END
   66        SUBROUTINE MESA(N,DR,V1)
   67        DIMENSION IA(20,4),IB(20,4),DS(20,4),DR(100)
   68        DIMENSION K(20),KC(20),D(100),K1(20),K2(20)
   69        DIMENSION V1(20),V2(20)
   70        DIMENSION DDD(32)
   71        IIIII=1
   72        JJJJJ=2
   73        KKKK=3
   74        LLLL=4
   75        NN=N*(N-1)/2
   76        N2=N-1
   77        N1=N-2
   78        KCT=1
   79        DO 70 KODE=1,4
   80        DO 15 I=1,N
   81        V2(I)=V1(I)
   82   15   CONTINUE
   83        DO 5 I=1,NN
   84      5 D(I)=DR(I)
   85        GO TO (1,1,2,3),KODE
   86      1 AP=.5
   87        AQ=.5
   88        B=0.
   89        G=-.5
   90        IF(KODE.EQ.2) G=-G
   91        GO TO 4
   92      2 B=0.
   93      3 G=0.
   94      4 DO 10 I=1,N
   95     10 K(I)=1
   96        KG=1
   97     20 II=0
   98        III=1
   99        J1=0
  100        I1=0
  101        DD=1000000.
  102        DO 30 J=2,N
  103        IJ=J-1
  104        DO 30 I=1,IJ
  105        II=II+1
  106        IF(DD.LE.D(II)) GO TO 30
  107        IF(K(J ).EQ.0) GO TO 30
  108        IF(K(I).EQ.0) GO TO 30
```

```
CARD
 109          DD=D(II)
 110          III=II
 111          I1=I
 112          J1=J
 113       30 CONTINUE
 114          IA(KG,KODE)=I1
 115          IB(KG,KODE)=J1
 116          DS(KG,KODE)=DD
 117          II=0
 118          JJ=0
 119          LL=0
 120          DO 60 J=2,N
 121          IJ=J-1
 122          DO 60 I=1,IJ
 123          II=II+1
 124          IF(II.EQ.III) GO TO 60
 125          IF( I.EQ.I1) GO TO 40
 126          IF(J.EQ.I1) GO TO 40
 127          IF(J.EQ.J1) GO TO 50
 128          IF( I.EQ.J1) GO TO 50
 129          GO TO 60
 130       40 JJ=JJ+1
 131          K1(JJ)=II
 132          GO TO 60
 133       50 LL=LL+1
 134          K2(LL)=II
 135       60 CONTINUE
 136          DO 65 I=1,N1
 137          II=K1(I)
 138          LL=K2(I)
 139          IF(KODE.LE.2) GO TO 65
 140          GG=K(I1)+K(J1)
 141          AP=K(I1)/GG
 142          AQ=K(J1)/GG
 143          IF(KODE.EQ.3) GO TO 65
 144          B=-AP*AQ
 145       65 D(II)=AP*D(II)+AQ*D(LL)+B*D(III)+G*ABS(D(II)-D(LL))
 146          K(I1)=K(I1)+K(J1)
 147          K(J1)=0
 148          V2(I1)=V2(I1)+V2(J1)
 149          V2(J1)=0.
 150          IF(KG.EQ.N-1) GO TO 66
 151          CALL F(N,V2,K,KG,FSTAT)
 152          DDD(KCT)=FSTAT
 153   66     CONTINUE
 154          KG=KG+1
 155          IF(KG.EQ.N-1) GO TO 67
 156          KCT=KCT+1
 157   67     CONTINUE
 158          IF(KG.NE.N )    GO TO 20
 159       70 CONTINUE
 160          WRITE(7,550) (DDD(I),I=1,8),IIIII
 161          WRITE(7,550) (DDD(I),I=9,16),JJJJJ
 162          WRITE(7,550) (DDD(I),I=17,24),KKKK
```

```
CARD
 163          WRITE(7,550) (DDD(I),I=25,32),LLLL
 164          RETURN
 165   550    FORMAT(7F10.3,F9.3,I1)
 166          END
 167          SUBROUTINE F(N,V1,K,KG,FSTAT)
 168          COMMON CF,STOT
 169          DIMENSION V1(20),K(20)
 170          SYY=0.
 171          DO 10 I=1,N
 172          IF(K(I).EQ.0) GO TO 10
 173          SYY=SYY+V1(I)*V1(I)/K(I)
 174   10     CONTINUE
 175          SYY=SYY-CF
 176          SWITH=STOT-SYY
 177          IF(SWITH.LE..000001) GO TO 15
 178          FSTAT=(SYY*KG)/(SWITH*(N-KG-1))
 179          IF(FSTAT.LE.999999.) GO TO 16
 180   15     FSTAT=999999.
 181   16     CONTINUE
 182          RETURN
 183          END
 184          SUBROUTINE BUTLER (L, RAND, IX, JX, NEND)
 185   C      RAND IS THE RANDOM DEVIATE GENERATED
 186   C      IX AND JX ARE INITAL VALUES
 187   C      NEND IS L + 1
 188   C      L IS AN INITIAL INTEGER
 189   C
 190   C      RANDOM NORMAL DEVIATES GENERATING PROGRAM
 191   C
 192   C      L IS THE INDEX FOR THE L TH RANDOM VARIABLE GENERATED
 193   C      RAND IS THE RANDOM VARIABLE GENERATED, (DISTRIBUTED NORMAL(0,1)
 194   C
 195   C      COMPUTER PROGRAM WRITTEN BY C. E. GATES, ESQ. 2/6/73
 196   C      FOR GENERATING RANDOM VARIABLES FROM THE NORMAL DISTRIBUTION
 197   C
 198          IMPLICIT REAL*8 (A-H,O-Z)
 199          REAL*4 C
 200          DIMENSION C(6),X(257),U(3),R(256)
 201          DATA C/2.515517,.802853,.010328,1.43279,.189269,.001308/
 202          IF (L.GT.NEND) GO TO 70
 203          CONST = DSQRT (1.000/(2.0D0 * 3.14159D0))
 204          X(1) = -3.6
 205          X(257) = 3.6
 206          FCLC = 0.0
 207          RAT = 1./256.
 208          RAND = 0.0
 209          DO 10 I = 1,255
 210          RAND = RAND + RAT
 211   C
 212   C      C.D.F. VALUE IS I/256
 213   C
 214          IF(I.GT.128) GO TO 12
 215          T = DSQRT(-2.0D0 *DLOG(RAND))
 216          GO TO 14
```

```
CARD
 217      12 T = DSQRT(-2.0D0 * DLOG(1.0D0 - RAND))
 218   C
 219   C     Z VALUE IS THE VALUE ALONG THE X-AXIS , I.E. X
 220   C
 221      14  Z = T - (C(1) + C(2)*T + C(3)*T**2)/(1. + C(4)*T + C(5)*T**2 +
 222     $ C(6)*T**3)
 223          IF (I.LT.129) Z = -Z
 224          X(I +1) = Z
 225   C
 226   C     FNEW IS THE CURRENT VALUE OF F(X); R(I) IS BUTLER,S R(I)
 227   C
 228          FNEW = CONST *DEXP(-Z**2 /2.0D0)
 229      20 R(I) = (FNEW-FOLD)/(FNEW + FOLD)
 230      10 FOLD = FNEW
 231          FNEW = 0.0
 232          R(256) = (FNEW-FOLD)/(FNEW + FOLD)
 233   C
 234   C     HERE WE START TO DO THE SAMPLING PHASE
 235   C
 236      70 CONTINUE
 237   C
 238   C     SELECT THE I TH INTERVAL WITH PROBABILITY 1/256
 239   C
 240          IX = IX *65539
 241          JX = JX * 262147
 242          RAND = .4656613D-9 * DFLOAT(IABS( IX + JX))
 243          I = 256.*RAND + 1.0
 244   C
 245   C     WE GENERATE THE THREE RANDOM UNIFORMS NEEDED
 246   C
 247          DO 32 K = 1,3
 248          JX = JX *262147
 249          IX = IX *65539
 250      32 U(K) = .4656613D-9* DFLOAT(IABS(IX + JX))
 251          Z = X(I + 1 ) - X(I)
 252   C
 253   C     U(3) IS USED TO DETERMINE WHETHER WE SAMPLE WITH PROBABILITY
 254   C     ABS(R(I)) OR 1 - ABS(R(I))
 255   C
 256          IF (U(3).LT. DABS(R(I))) GO TO 34
 257          RAND = X(I) + Z*U(1)
 258          GO TO 36
 259   C
 260   C     WE DETERMINE THE MAX. OR MIN. OF R(I) DEPENDING OF WHETHER
 261   C     R(I) .LT. OR.GT. 0
 262   C
 263      34 IF (R(I).LT.0.0) GO TO 50
 264          RAND = DMAX1(U(1),U(2))
 265          GO TO 52
 266      50 RAND = DMIN1(U(1),U(2))
 267      52 RAND = X(I) + Z*RAND
 268      36 CONTINUE
 269          RETURN
 270          END
```

```
CARD
   1          DIMENSION D(10),DD(10)
   2          MMM=1
   3          KKK=1
   4          NEND=KKK+1
   5          M=500
   6          PCWER=4.
   7          JX=476963
   8          IX=445123
   9          A=100000.
  10          DO 100 III=1,M
  11          DO 10 I=1,10
  12          CALL BUTLER (KKK,D(I),IX,JX,NEND)
  13    10    CONTINUE
  14          DC 20 I=1,MMM
  15          D(I)=D(I)+PCWER
  16    20    CONTINUE
  17          DO 50 K=1,10
  18          DC 40 I=1,10
  19          IF(A.LE.D(I)) GO TO 30
  20          A=D(I)
  21          J=I
  22    30    CONTINUE
  23    40    CONTINUE
  24          DD(K)=A
  25          D(J)=100000.
  26          A=100000.
  27    50    CONTINUE
  28          WRITE(7,500) (DD(JJ),JJ=1,10)
  29   100    CCNTINUE
  30   500    FORMAT(10F8.4)
  31          STOP
  32          END
  33          SUBROUTINE BUTLER (L, RAND, IX, JX, NEND)
  34    C     RAND IS THE RANDOM DEVIATE GENERATED
  35    C     IX AND JX ARE INITAL VALUES
  36    C     NENC IS L + 1
  37    C     L IS AN INITIAL INTEGER
  38    C
  39    C     RANDOM NORMAL DEVIATES GENERATING PROGRAM
  40    C
  41    C     L IS THE INDEX FOR THE L TH RANDOM VARIABLE GENERATED
  42    C     RAND IS THE RANDOM VARIABLE GENERATED, (DISTRIBUTED NORMAL(0,1)
  43    C
  44    C     COMPUTOR PROGRAM WRITTEN BY C. E. GATES, ESQ. 2/6/73
  45    C     FOR GENERATING RANDOM VARIABLES FROM THE NORMAL DISTRIBUTION
  46    C
  47          IMPLICIT REAL*8 (A-H,O-Z)
  48          REAL*4 C
  49          DIMENSION C(6),X(257),U(3),R(256)
  50          DATA C/2.515517,.802853,.010328,1.43279,.189269,.001308/
  51          IF (L.GT.NEND) GO TO 70
  52          CONST = DSQRT (1.0D0/(2.0D0 * 3.14159D0))
  53          X(1) = -3.6
  54          X(257) = 3.6
```

```
CARD
   55         FOLD = 0.0
   56         RAT = 1./256.
   57         RAND = 0.0
   58         DO 10 I = 1,255
   59         RAND = RAND + RAT
   60   C
   61   C     C.D.F. VALUE IS I/256
   62   C
   63         IF(I.GT.128) GO TO 12
   64         T = DSQRT(-2.0D0 *DLOG(RAND))
   65         GO TO 14
   66      12 T = DSQRT(-2.0D0 * DLOG(1.0D0 - RAND))
   67   C
   68   C     Z VALUE IS THE VALUE ALONG THE X-AXIS , I.E. X
   69   C
   70      14  Z = T - (C(1) + C(2)*T + C(3)*T**2)/(1. + C(4)*T + C(5)*T**2 +
   71        $ C(6)*T**3)
   72         IF (I.LT.129) Z = -Z
   73         X(I +1) = Z
   74   C
   75   C     FNEW IS THE CURRENT VALUE OF F(X); R(I) IS BUTLER,S R(I)
   76   C
   77         FNEW = CONST *DEXP(-Z**2 /2.0D0)
   78      20 R(I) = (FNEW-FOLD)/(FNEW + FOLD)
   79      10 FOLD = FNEW
   80         FNEW = 0.0
   81         R(256) = (FNEW-FOLD)/(FNEW + FOLD)
   82   C
   83   C     HERE WE START TO DO THE SAMPLING PHASE
   84   C
   85      70 CONTINUE
   86   C
   87   C     SELECT THE I TH INTERVAL WITH PROBABILITY I/256
   88   C
   89         IX = IX *65539
   90         JX = JX * 262147
   91         RAND = .4656613D-9 * DFLOAT(IABS( IX + JX))
   92         I = 256.*RAND + 1.0
   93   C
   94   C     WE GENERATE THE THREE RANDOM UNIFORMS NEEDED
   95   C
   96         DO 32 K = 1,3
   97         JX = JX *262147
   98         IX = IX *65539
   99      32 U(K) = .4656613D-9* DFLOAT(IABS(IX + JX))
  100         Z = X(I + 1 ) - X(I)
  101   C
  102   C     U(3) IS USED TO DETERMINE WHETHER WE SAMPLE WITH PROBABILITY
  103   C     ABS(R(I)) OR 1 - ABS(R(I))
  104   C
  105         IF (U(3).LT. DABS(R(I))) GO TO 34
  106         RAND = X(I) + Z*U(1)
  107         GO TO 36
  108   C
```

```
CARD
 108  C      WE DETERMINE THE MAX. OR MIN. OF R(I) DEPENDING OF WHETHER
 110  C      R(I) .LT. OR.GT. 0
 111  C
 112      34 IF (R(I).LT.0.0) GO TO 50
 113         RAND = DMAX1(U(1),U(2))
 114         GO TO 52
 115      50 RAND = DMIN1(U(1),U(2))
 116      52 RAND = X(I) + Z*RAND
 117      36 CONTINUE
 118         RETURN
 119         END
```

VITA

James Michael Norton

Candidate for the Degree of

Doctor of Philosophy

Thesis: SOME STATISTICAL PROCEDURES TO AID IN THE EVALUATION OF A CLUSTER ANALYSIS

Major Field: Statistics

Biographical:

Personal Data: Born in San Diego, California, April 5, 1948, the son of James and Norma Norton.

Education: Attended elementary school in Brandon, Vermont; attended high school in Brandon, Vermont; graduated from Otter Valley Union High School in 1966; received the Bachelor of Science degree from the University of Vermont with a major in Mathematics in August, 1970; received the Master of Arts degree from the University of Vermont with a major in Mathematics in May, 1972; completed requirements for the Doctor of Philosophy degree in December, 1975, at Oklahoma State University.

Professional Experience: Research and teaching assistant at the University of Vermont from September, 1970 to May, 1972; research and teaching assistant at Oklahoma State University from August, 1972 to August, 1975.

Professional Organizations: American Statistical Association, Institute of Mathematical Statistics, Oklahoma Chapter of the American Statistical Association.