An Investigation of Cluster Analysis

John C. Klingel

Utah State University
AN INVESTIGATION OF CLUSTER ANALYSIS

by

John C. Klingel

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Applied Statistics

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1973
ACKNOWLEDGMENTS

Cordial thanks in the construction of this thesis is hereby extended to the following persons: To all members of my committee; and to Dr. A. R. Klingel, my father, for several lengthy discussions on various aspects of this thesis.

I would also like to thank the following persons for help in procuring computer programs: Mr. Kim Marshall and Mr. Bob Johnson.

John C. Klingel
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Definitions</td>
<td>6</td>
</tr>
<tr>
<td>Clustering Phases</td>
<td>7</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>10</td>
</tr>
<tr>
<td>Single Linkage</td>
<td>10</td>
</tr>
<tr>
<td>Average and Complete Linkage</td>
<td>15</td>
</tr>
<tr>
<td>Goodall's Probability Model</td>
<td>15</td>
</tr>
<tr>
<td>Column Discrepancy</td>
<td>17</td>
</tr>
<tr>
<td>Orloci's Within-Group Dispersion</td>
<td>19</td>
</tr>
<tr>
<td>K-Means Analysis</td>
<td>21</td>
</tr>
<tr>
<td>Incremental Sum of Squares</td>
<td>23</td>
</tr>
<tr>
<td>Information</td>
<td>25</td>
</tr>
<tr>
<td>METHODS</td>
<td>28</td>
</tr>
<tr>
<td>RESULTS AND DISCUSSION</td>
<td>33</td>
</tr>
<tr>
<td>SUMMARY AND CONCLUSIONS</td>
<td>45</td>
</tr>
<tr>
<td>LITERATURE CITED</td>
<td>48</td>
</tr>
<tr>
<td>VITA</td>
<td>50</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gross comparisons of 8 cluster methods</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Analysis of Sneath's bacteria strains</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Common attribute correlation matrix used to generate the data</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Attribute means and standard deviations used to generate the data used in all 3 cluster programs</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Group correlations from the program GRPSIM</td>
<td>35</td>
</tr>
<tr>
<td>6</td>
<td>Population distances, using attribute means as the scores</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>Mutual information between the population means</td>
<td>44</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>The 4 attributes for one individual, a snail, as a vector</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>A cluster space for fictitious plants, measured on 3 attributes</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Clusters forming as the sorting level increases</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Individuals columnized to show relationships</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>Author-prescribed hierarchy</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>MINT Dendrogram for single linkage</td>
<td>34</td>
</tr>
<tr>
<td>7</td>
<td>The correlation between 2 individuals</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>The MDISP dendrogram</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>Relative population distances</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>Dendrogram for MINFO</td>
<td>44</td>
</tr>
</tbody>
</table>
ABSTRACT

An Investigation of Cluster Analysis

by

John C. Klingel, Master of Science

Utah State University, 1973

Major Professor: Dr. Rex L. Hurst
Department: Applied Statistics

Three cluster analysis programs were used to group the same 64 individuals, generated so as to represent eight populations of eight individuals each. Each individual had quantitative values for seven attributes. All eight populations shared a common attribute variance-covariance matrix.

The first program, from F. J. Rohlf's MINT package, implemented single linkage. Correlation was used as the basis for similarity. The results were not satisfactory, and the further use of correlation is in question.

The second program, MDISP, bases similarity on Euclidean distance. It was found to give excellent results, in that it clustered individuals into the exact populations from which they were generated. It is the recommended program of the three used here.

The last program, MINFO, uses similarity based on mutual information. It also gave very satisfactory results, but, due to visualization reasons, it was found to be less favorable than the MDISP program.

(56 pages)
INTRODUCTION

Cluster analysis, one aspect of the more general field of numerical taxonomy, is the main topic of this paper. Discussion will begin with an introduction to numerical taxonomy and its relationship to cluster analysis. Focus will then be on discussing cluster analysis, in particular several clustering methods.

The study of relationships among M individuals has evolved to utilizing numerical methods, whereby some N features, or attributes, of the M individuals are measured and quantified. An individual therefore becomes an array of numerical values, one value per attribute. Once quantified, the attribute values can be manipulated so as to identify subsets of the M individuals as groups of similarity, or "alikes." The measured features might, for instance, include combinations of heights, colors, numbers of legs, ecological factors, and weights.

The process by which these numerical valued attributes are manipulated is called numerical taxonomy. This process not only seeks groups of similar individuals, but also the relationships between these groups. It is important for a taxonomist to know both that he has distinct groups and also the extent to which the groups are similar. One thus tries to establish a pyramid of similarity, a good example of which is the taxonomic pyramid whereby living organisms are classified according to their kingdom, then phylum, then class, order, family, genus, and finally species. Obviously, within-group similarity is lowest at the apex of this pyramid (kingdom), increasing to its highest at the base (species).
Knowing the similarity of groups of individuals gives a taxonomist an idea of how distantly related they are. Heywood (1967, p. 44), defines numerical taxonomy as "the numerical evaluation of the similarity between groups of organisms, and the ordering of these groups into higher ranking taxa (i.e. more general groups) on the basis of these similarities."

It will now be instructive to investigate the mechanics of numerical taxonomy. Imagine several men, mice, crabs, and spiders being mixed together in a barrel, labeled A. There is then a mixture of the two orders Primata (man) and Rodentia (mice), and the two classes Crustacea (crabs) and Arachnida (spiders). If our representatives of these 4 taxonomic groups are then subjected to an ideal classification scheme which enters barrel A and puts similar objects into smaller barrels, 4 barrels, 1, 2, 3, 4, would be needed. Each barrel would contain 1 of the 4 animal types. This segregates the 4 animal types, but says nothing as of yet about inter-animal-type similarities. The actual distance between the smaller barrels is indicative of this, though, and is of interest to taxonomists.

Within barrel A, then, 2 pairs of barrels, (1-2) and (3-4) would be found. These would contain (men-mice) and (crabs-spiders), respectively. The 2 pairs (1-2) and (3-4) would be set far apart to indicate that they are only distantly related subsets of some larger population and do not unite into a common set, (1, 2, 3, 4), until much higher on the taxonomic pyramid. Union, in fact, occurs at the top of the
pyramid, at kingdom. Within each pair, the members (1 and 2) and
(3 and 4) would be close together, indicating more similarity within
each pair than between the pairs. As men and mice are closer together
than crabs and spiders, the (1-2) pair will fuse slightly before the
(3-4) pair.

Concern is next directed toward visualizing an individual which
is quantified on N attributes and then classified in relation to other
individuals. This author finds convenience in thinking of an individual
as a vector, the N elements of which are its numerical scores for the
N attributes measured. Refer to Figure 1, the 4 attributes of a snail.
A population of M individuals can then be represented as M points in an
N-dimensional space, where the dimensions are the attributes. In
Figure 2 is found the 3-dimensional space for some ficticious plants
measured on 3 attributes. There, $X_1$ is a leaf feature, $X_2$ is a root
type, and $X_3$ is a flower type.

The individuals in Figure 3 have been arbitrarily enclosed in
3 volumes, or clusters, or illustration. Because these groups are rela-
tively dense and each far apart, one might assume that the population is
composed of only members of 3 orders, for instance. If the centroids
were a good deal closer, though, question might arise about the validity
of assuming 3 distinct orders and not just one diffuse one. In addition,
the dimension of a real data set may be 15 or more. Both of these con-
ditions defy interpretation by inspection alone. For resolution of
these interpretational problems, the data should be subjected to another
tool. Cluster analysis is one such tool. It is used in many fields,
including archeology, biology, sociology, and linguistics.
Figure 1. The 4 attributes for one individual, a snail, as a vector.

Figure 2. A cluster space for ficticious plants, measured on 3 attributes.
Figure 3. Clusters forming as the sorting level increases. The example is a simple linkage algorithm. Notice the increase in group (cluster) generality as the sorting level increases from 10 to 40.

The primary function of cluster analysis is to suggest groups of similar individuals, and also some information on the inter- and intragroup relationships. It proceeds by first finding either the similarity between all individuals studied or that between individuals and user-specified points. Based on these similarities (via relationships), individuals are then pooled into C clusters of alike objects.

The bulk of cluster procedures have been presented in the last 15 years, and frequently appear to lack mathematical rigor. This may
be true, but the majority of these methods make some sense out of chaos. Methods presented here are only representative of the more popular approaches.

Before looking at specific methods, time will be taken to present a little more foundation. A resume of the geometry of a simple cluster method will be given. The example is found in Sneath (1966), where the criterion used in deciding whether or not two groups should fuse is a Euclidean distance. The level of this criterion is called the sorting level, which changes as the clustering algorithm advances.

Figure 3 shows how clusters, indicated by a solid line, are formed as the sorting level increases from cycle to cycle. This is to say that, at each subsequent cycle, points farther and farther from the group (less similar ones) are allowed to enter it. When the sorting level is "so large" all individuals are in one group and the algorithm terminates. Notice, in Figure 3, how new groups form and new points enter existing groups as the sorting level increases, until all individuals are in one group (level 40).

Definitions

The following definitions should prove beneficial to anyone reviewing the cluster analysis literature for the first time. The word "group" can mean from one to all of the individuals in a study.

1. Agglomerative procedure. A subset of the data is chosen as a starting group. Subsequent clusters are formed by adding to this subset or creating more subsets like it.
2. Divisive procedure. Here, the entire population is split
into smaller and smaller subsets to form groups. The splitting con-
tinues until the number of subsets equals the number of individuals in
the population.

3. Sorting level. This is a magnitude of the similarity between
groups. It is used as a decision function in determining whether or
not two groups should fuse.

4. Hierarchic procedure. Clusters at any one sorting level may
not overlap, i.e. an individual is in only one cluster.

5. Monothetic. The cluster-forming criterion is based on one
attribute.

6. Space dilation. As groups grow, they appear to recede from
each other. Tendency for the algorithm to allocate new individuals
to the smallest groups.

7. Dendrogram. A "tree diagram" which shows the group structures
(relationships) at any one sorting level. Usually constructed by the
user after all the data has been clustered.

8. Cophenetic values. Measures of the agreement between 2 dendro-
grams, or between a dendrogram and its original similarity matrix.

Clustering Phases

Lance and Williams (1968) report that any cluster method may entail
4 distinct phases, as listed below. One section of this paper will be
a comparison of several cluster procedures and will use these 4 phases
as a guideline. The first two are essential, while the last two are
often lacking or in need of more emphasis.
Initiation phase

Clustering may be initiated in a multitude of ways, generally in an hierarchical manner. The underlying statistic is usually an individual-individual similarity measure, referred to as an $(i,j)$ measure by Jardine and Sibson (1971). Some methods have arbitrary inceptions, while others are predetermined. MacQueen (1966), for example, begins his algorithm by defining the first $K$ data points as initial means, with which the remainder of the population will then be compared. Sneath (1957), Sokal and Michener (1958), and Peters (1971), for example, first establish a matrix of pairwise similarities which are then used to determine future clusters. This approach is most common.

Allocation phase

Most methods that this author came across rely on a "distance" in Euclidean space as a criterion for allocating new elements to clusters or for fusing existing ones. MacQueen's, for instance, is quite simple, utilizing the actual distance from points to a centroid. Others, such as Burr's (1970), use a transformation of "distance," a group sum of squares in this case. Peters (1971) uses the displacements between correlation coefficients ranked in columns.

Reallocation phase

Sokal and Michener (1958) describe a method by which they cluster in 2 phases, the second being dependent on the results of the first. Peters (1971) also reallocates, in a sense, when he clusters elements according to secondary correlations, which are based on preliminary ones.
MacQueen (1966) obtains $J$ initial clusters, then, using the centroids of these $J$ clusters as fixed points, the whole population is re-clustered about them.

**Single element phase**

The last phase is where a decision is made about leaving single elements as clusters of size one. Generally, a reallocation phase will minimize these, and inspection of defined populations may be sufficient to locate them within any group.

Enough groundwork should now be digested, that we can go to specific methods. These are only representative of the various procedures. Sokal and Sneath (1963), Jardine and Sibson (1971), and Williams (1971) provide excellent references for the study of numerical classification.
REVIEW OF LITERATURE

There are two parts to this chapter, the first being a summary table (Table 1) whereby gross comparisons of 8 cluster analysis methods can be made. The table headings are the 4 phases which Lance and Williams (1968) feel a cluster analysis procedure may entail. Any reference to a "similarity matrix" in that table is to mean an individual-individual similarity matrix. The second part of this chapter is a method by method description of these 8 cluster methods.

**Single Linkage**

Introduced by Florek in 1951 (Jardine and Sibson, 1971), this is the simplest of clustering methods. The easiest way to think of this process is to first set up a matrix of similarity coefficients for all data pairs in the population. Several coefficients are in use, including Yule's (Clement, 1954), Jaccard's (Jardine and Sibson, 1971), product moment correlation, and one used by Sneath (1957), to be illustrated shortly.

The next step is to unite the most similar pair, say C and D, and replace them with the value, X, that minimizes the dissimilarity between both (C and D) and X. Repeat this process, changing the sorting level accordingly, until only one element is left. Print groups and sorting levels as the algorithm proceeds.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Initiation</th>
<th>Allocation</th>
<th>Termination</th>
<th>Reallocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single link (Sneath)</td>
<td>Calculate a similarity matrix. Several coefficients are available</td>
<td>Fuse groups if any links exceed sorting level. Change sorting level, repeat</td>
<td>When all individuals are in a group</td>
<td>None suggested</td>
</tr>
<tr>
<td>Average link (Sokal, Michener)</td>
<td>As above. Use product moment correlation as similarity index</td>
<td>Fuse if the average link exceeds the sorting level</td>
<td>As above</td>
<td>Group centroids from phase 1 become points to re-cluster around</td>
</tr>
<tr>
<td>Probability (Goodall)</td>
<td>Similarity is based on a probability.</td>
<td>Replace most similar pair with new element. Calculate new matrix, repeat</td>
<td>As above</td>
<td>Not conducive to</td>
</tr>
<tr>
<td>Column discrepancy (Peters)</td>
<td>Arrange similarity (based on column discrepancy) in columns</td>
<td>Groups are fused so as to minimize the within-column discrepancy</td>
<td>When all individuals are classified</td>
<td>Clusters points according to secondary similarities, which are based on initial ones.</td>
</tr>
<tr>
<td>Within-group dispersion (Orloci)</td>
<td>Similarity matrix is based on distance</td>
<td>Unite groups so as to minimize the within-group dispersion</td>
<td>When mean vectors become significantly different</td>
<td>Some information on group relationships is regained by centroid ordination.</td>
</tr>
<tr>
<td>K means (MacQueen)</td>
<td>First k data points are k means about which clustering begins</td>
<td>Groups fuse if the distance between them is less than C</td>
<td>When all data is group</td>
<td>Recluster about the means of the groups found in phase 1.</td>
</tr>
<tr>
<td>Methods</td>
<td>Initiation</td>
<td>Allocation</td>
<td>Termination</td>
<td>Redlocation</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------------</td>
<td>------------------------------------------------</td>
<td>------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Incremental sum of sqrs. (Burr)</td>
<td>Similarity based on distance between groups</td>
<td>Unite closest groups. Recalculate between-group distances. Repeat.</td>
<td>As above.</td>
<td>None suggested.</td>
</tr>
<tr>
<td>Information (Orloci)</td>
<td>Calculate mutual information between subsets of the population</td>
<td>Minimums are fused, replaced by an average value. A new matrix is made. Repeat.</td>
<td>When information matrix is of order 1</td>
<td>None suggested.</td>
</tr>
</tbody>
</table>
For Sneath's (1957) variation of this single linkage, calculations are initiated by establishing a matrix, S, of similarity coefficients for all possible data pairs. Consider binary attributes (presence, absence) of an A,B pair as occurring in one of three ways: (a) Only A possesses a certain attribute; (b) Both A and B have it; and (3) Only B has it. Letting "a" represent the number of attributes that only A has, "b" those that A and B share, and "c" the number that only B has, similarity can be expressed as the ratio of 

\[
\frac{b}{a+b+c} = \text{similarity.}
\]

Assuming (i) individuals in a population, it is this (i x i) matrix of similarity coefficients, S, that is used in determining groups. The similarity coefficient for the arbitrary pair X, Y is denoted by \(S_{X,Y}\).

The next step is to scan the X matrix and group individuals with an S value greater than a sorting level of .99, for instance. Then, print groups as sets of individuals, the appropriate sorting level, and return to scan the X matrix again. Lower the sorting level and find more groups, repeating this until all individuals are in a group. As the sorting level decreases, more diffuse groups are formed, as less similarity is required for fusion. This stage of the analysis only gives rough group relationships.

For more insight into group relationships, Sneath next looks at the intergroup and intragroup Mean Similarity Values, \(\bar{S}_{ij}\) and \(\bar{S}_k\), respectively. These values give insight into the density of a group and the similarity between pairs of groups. Following is Sneath's example of how to use these values.
Table 2 contains the groups-within-sorting-level analysis for several strains of bacteria. At the high level of .99, only strains 1 and 2 form a group. Just below this, at .98, strains 4 and 5 unite, forming group 2. These two groups remain distinct until a sorting level of .80.

Table 2. Analysis for Sneath's bacteria strains

<table>
<thead>
<tr>
<th>SL</th>
<th>Groups</th>
<th>Individuals</th>
<th>$\bar{S}$ values</th>
</tr>
</thead>
<tbody>
<tr>
<td>.99</td>
<td>A</td>
<td>1,2</td>
<td>$a = .988$</td>
</tr>
<tr>
<td>.98</td>
<td>A</td>
<td>1,2,3</td>
<td>$b = .981$</td>
</tr>
<tr>
<td>.</td>
<td>B</td>
<td>4,5</td>
<td>$ab = .781$</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.80</td>
<td>A,B</td>
<td>1,2,3,4,5</td>
<td></td>
</tr>
</tbody>
</table>

The high $S_a$ and $S_b$ values in Table 2 show that both groups are compact, indicating cause for individual recognition. Simultaneously, the $S_{ab}$ value indicates a fair amount of similarity between the groups, and that both are probably highly related on the next highest taxonomic level.

The Utah State University Desert Biome computer library contains a statistical package, MINT, which has single linkage as a cluster analysis option. The package was constructed by F. J. Rohlf, at the State University of New York, Stony Brook. The MINT package is used in this study.
Average and Complete Linkage

The principles of both of these methods are very similar to single linkage and will be presented here mainly by noting differences. Sokal and Michener (1958) present a widely used average linkage method, and apply their method to a bee population in that paper. Sorensen (1948) is responsible for the introduction of the complete linkage system, according to Jardine and Sibson (1971).

The major difference between single and complete linkage lies in the group acceptability criterion. Assume that the groups EDAB and BC are in question of fusion. Single linkage will fuse them immediately, as one element of the BC pair is in common to both groups and thus one "distance" between the groups is spanned. Fusion by the average linkage method, however, necessitates the average similarity between all pairs (BD, BE, etc.) to be at least as large as the sorting level. Average linkage is thus more conservative.

Complete linkage operates quite like the other two linkages. Here, however, all between group similarities must satisfy the sorting criterion before groups are united. Complete linkage thus makes certain of strict peer identity before uniting groups and is thus the most restrictive.

Goodall's Probability Model

The underlying similarity criterion used in this method is the complement (1-P) of the probability (p) that a random pair will have a similarity greater than the pair in question. Goodall (1966) examines 3 classifications of data to which the probability model is applicable. These are considered below.
First mentioned is qualitative data. By defining all differing pairwise values as equally dissimilar and ordering non-differing pairwise values as prescribed above, (probability complement), closer similarity is demonstrated between a pair which shares an uncommon attribute than between a pair sharing a common one.

Secondly, Goodall refers to "ordered attributes," which are discontinuous ranks. Representing 5 individuals by A, B, C, D, E, a similarity progression might be

\[
\text{the similarity of } AB > AC > BD > BE > AE,
\]
because the distance on the ranked scale is greatest for AE and least for AB. However, the relationship of AB to CD, or AC to CE, is not yet clear. These are determined by utilizing the sum of the probabilities of all possible pairs of values included in each group pair. For the AC-BD group pair, this would mean summing the probabilities of AC, AD, BC, and CD. In other words, ordering is carried out by implementing the population proportion contained in each group pair. The smallest total probability is indicative of greatest similarity.

As an example, assume that 4 ranked categories contain proportions of individuals in the population as follows: .1, .4, .4, and .1. Then, the pairs (1,3) and (2,4) are equally similar since \((.1+.4+.4) = (.4 + .4+.1)\). However, (1,2) is more similar than (2,3) since \((.1 + .4) < (.4+.4)\).

When the ordered attributes are discrete, all pairs of values are again ordered according to the population proportion encompassed by the two values. Upon ordering of all pairs of values, in whatever way, similarity values are determined by the pattern established for the qualitative attributes above.
The metric (quantitative) attributes are considered next. Ordering is based on differences of values, the similarity increasing as the difference of the values decreases. As for pairs whose value differences are equal, one again utilizes proportions (as above) to "weight" the differences before attempting to order the pairs. Again, similarity is expressed as the complement of the cumulative probability of at least the pair in question.

The final step is to examine the data not merely attribute by attribute, but by a number of attributes simultaneously. (An assumption of independence of attributes within any individual is necessary here, causing some criticism of the approach.) Thus, we will now consider pairs or sets of values instead of only pairs of values. To determine similarity between sets, assuming ordering has been previously established within each attribute, the product of cumulative probabilities for all pairs on all attributes is examined. Set-pairs with smaller probability products are considered more similar than set-pairs with larger products. Analogous to single attributes, individual probabilities are summed over all set-pairs at least as similar as that pair under consideration. This results in a similarity measure, for which the similarity index is the complement.

Column Discrepancy

This method, due to J. A. Peters, was developed because its author was dissatisfied with the believed information loss when other methods are applied to biogeographic data. Peters feels that the relationship of any individual with all other individuals, not just one, should be
retained throughout an analysis and therefore invented a statistic to do so. He also feels that other measurements can be highly biased if the failure to discover a particular attribute is due to the search mechanics and not really its absence.

Simple similarity is based on a coefficient similar to Sneath's, i.e. "common taxa divided by total taxa present." The method operates by first establishing ranked columns, one for each individual i, of the similarity between individual i and all others studied. Along with each coefficient is its associated individual. The column starts with the highest coefficient and its corresponding individual, and descends. The similarity measure used thus far is only initial and will be replaced soon.

Examining any two of these ranked lists should indicate the degree of similarity between the two individuals for which the columns exist. Agreement is determined by comparing the $j^{th}$ ranked individual down each column and noting they are the same. If both $j^{th}$ rankings frequently refer to the same individual, then these two columnized individuals are similar. If many discrepancies exist, there is probably a natural barrier of some sort between these two individuals. An illustration of this is found in Figure 4, taken from Peters (1971). Notice how the first 3 form group 1, and the last 3 form group 2.

Realize that the number of column discrepancies in the whole study is a function of the relative column positions with each other. Peters therefore permutes the columns to maximize the number of discrepancies and calculates similarity of $k,n = 1-(2D_{kn}/N^2)$ where $D_{kn}$ is the number of discrepancies between column k and n, and N is the number of individuals in the study.
This similarity value ranges from 0 (no similarity) to 1 and is calculated for each data pair. These more complicated similarity coefficients are then put into a matrix and examined.

For large amounts of data, Peters devised the program JPFRF to do the inspection. Here, however, he counts the displacements instead of the discrepancies. Displacements are defined as the number of moves of ranked coefficients needed to get all individuals ranked adjacently in the two columns. Total displacements are then printed out, and individuals with displacements less than some specified value are clustered together.

**Orloci's Within-Group Dispersion**

This is an agglomerative method similar to that of Jancey (1966) and utilizes an increase in the within-group dispersion as a cluster-or-not criterion. The method is dependent on the works of Edwards and Cavalli-Sforza (1966). It handles quantitative and qualitative data equally well (Orloci, 1967).
Two groups are fused only if the fused, within-group dispersion is less than that for the fusion of either of the groups with any of the others in the study. Consider having $N$ individuals and $P$ different attributes, as with $N$ stands and $P$ species. The challenge here is to locate clusters in $P$-space and maximize their density. To do this, one first needs to define the distance measurement used in calculating the within-group dispersion. Though absolute distance seems an obvious solution, Orloci points out that stand differences $N_i$ and $N_j$, which are measured in this way are dependent on the magnitude of the species composition as well as the species present themselves.

As this is not desirable, Orloci relies on a standard distance to avoid this problem. Below are the definitions necessary for understanding the mechanics of this method.

1. $X$ is a data matrix. Columns are attributes, rows are individuals.

2. $X_{ej}$ is the value of the $e$th attribute, $j$th individual.

3. $V_j$ is the $\sqrt{\sum_{e=1}^{p} X_{ej}^2}$, summing $e = 1$ to $p$. This is the length of the $j$th position vector.

4. $D_{jk} = \sqrt{\left(\sum_{e=1}^{p} X_{ej} - X_{ek}\right)^2}$, summing $e = 1$ to $p$, is the distance between stands $j$ and $k$.

5. $Q_m = \sum_{j=1}^{k_m} V_j^2$, summing $j = 1$ to $k_m$, is the average within-group dispersion. Here, $k_m$ is the number of individuals in group $m$.

This idea was introduced by Edwards and Cavalli-Sforza.

6. $I = \sum_{m=1}^{G} Q_m / Q_s$, summing $m = 1$ to $G$, is the overall heterogeneity in a sample of $G$ groups. $Q_s$ is the sample sum of squares.

7. $E = (Q_s - \sum_{m} Q_m) = (I-1)$ is the measure of overall group differences, subject to the restriction that $0 < E < 1$. 
The values of the $Q_i$ are then used as the cluster-or-not decision function. Clusters are then formed in successive cycles, fusing at most 2 individuals per cycle, so as to repeatedly maximize $E$. This simultaneously minimizes the within-group sum of squares.

In short, groups $A$ and $B$ are fused if

$$(Q_{AB} - Q_a - Q_B) < (Q_{RT} - Q_R - Q_T),$$

where $R = A$ or $B$, and $T =$ any other individual.

If the above two quantities happen to be equal, the first encountered is the smaller, by definition.

The output of this procedure allows for the construction of a dendrogram, stems joining at appropriate $Q/k$ values. The value of $I$ serves as a classification efficiency measure, since $I = 0$ implies intensive classification, or the presence of $N$ different individuals. Classification can be terminated when stem mean vectors become significantly different. The FORTRAN program for this method, MDISP, is clearly written and easy to manipulate if desired.

**K-Means Analysis**

This method is the most robust yet mentioned, as the use of it as a clustering method is only one of its uses. Among the others mentioned by MacQueen (1966) are relevant classifications (predictions), distribution approximation, test for independence of several variables, and a scheme for lexicographic classification systems.

Three values, $C$, $R$, and $K$ must be established by the user to initiate calculations. The value of $K$ prescribes that the first $K$ sample points will be $K$ means about which the remaining $N-K$ points
will initially be clustered, as follows. First, comparisons between C and all possible pair wise distances among the K means are made. Pairs of these means with a distance less than C are fused together. Once this is done, sample point K + 1 is read and sent to cohabit with the mean to which it is nearest, say the I\(^{th}\). A new, weighted mean is then calculated for the (K + 1)\(^{th}\) point and the I\(^{th}\) mean, this new mean then replacing both values. The process continues to read a point, cluster and calculate a new average, compare all distances with C, and regroup, if necessary. This goes on until all the data is read. This generally reduces K and therefore "coarsens" the partition.

The final parameter in the grouping process, R, is used in conjunction with C to "refine" the partition, i.e. increase K. If the distance from the new point to the nearest mean is existence is greater than R, the point temporarily becomes its own centroid and K increases. Thus, clustering is restricted by C and R, and K becomes a variable dependent on them.

Once all points are clustered and some number, M, of means are located, a reallocation phase begins. Here, all original points are reclustered around the final M means of the first phase. These means stay constant at M and final groupings are considered to be the groups formed around them.

The output of the process consists of points within a group, the distance to their mean, and up to 18 characters of information on any point. Also, group averages, between mean distances, sum of squares of deviations of points about respective means, and a grand average of these quantities (over groups) are printed. The grand average is
called a within-class variation, but is not quite a within-class variance. For the mathematics of the process, see MacQueen (1966), where a great deal more detail is presented than is desirable here.

**Incremental Sum of Squares**

In a recent publication, Burr (1970) describes a cluster analysis package that entails six cluster strategies. Only two of these, the incremental sum of squares and the variance procedures, were new. Here, discussion is limited to the incremental sum of squares method, as the variance procedure is very similar. The other 4 strategies are the 3 linkage methods discussed earlier and a centroid method.

The incremental sum of squares procedure seems quite commendable and is easy to comprehend. It implements squared Euclidean distance (SED) as its similarity criterion, measured from a data point to a centroid. Four general steps, listed below, are carried out.

1. With all N data points forming N clusters, a pairwise distance matrix, D, is calculated. \( D_{ij} \) is the distance between i and j.
2. D is scanned to locate the closest pair, say i and j.
3. Points i and j are fused and information on them is printed. If only one cluster remains, processing terminates.
4. D is updated to \( D' \), where distances are now from cluster \( J \) (groups i+j) to the remaining K groups. Repeat step 2 to 4.

The symbolism in the procedure is as follows. Letting

- \( N_i \) = the size of cluster i
- \( \ell_{ij} \) = the j points in cluster i, and
- \( C_i \) = the centroid of cluster i,
we define $S_i$, the sum of squares within cluster $i$, as

$$S_i = \sum (t_{ij} - C_i)^2$$

summing $j = 1$ to $N_i$.

Also, the sum of squares within cluster $J$, if $i$ and $j$ fuse to $J$, is defined as

$$S_J = \sum (t_{ik} - C_i)^2 + \sum (t_{kj} - C_j)^2$$

summing the first term from $1$ to $N_i$ and the second term from $1$ to $N_j$. Here, $C_{ij}$ is the centroid common to cluster $J(=i+j)$.

Next, the distance between two clusters, $J(=i+j)$ and $K$, is defined as

$$D_{JK} = \left[ \left( (N_i - N_k) / D_{ik} \right) + \left( (N_j + N_k) / D_{jk} \right) - \left( N_k / D_{ij} \right) \right] / (N_i + N_j + N_k).

Finally, the increment in the within-class sum of squares, $S_{ij}$, where $i$ and $j$ fuse, is defined as

$$S_{ij} = (N_i * N_j * D_{ij}) / (N_i + N_j).$$

The algorithm seeks to minimize the $S_{ij}$ values during any fusion. Question might arise as to whether one should minimize the within-group sum of squares or to maximize the between-group sum of squares. The problem is resolved by observing the following, assuming a total of $T$ points. Let $SSWC$ be the total sum of squares within clusters, and $GTSS$ be the grand total sum of squares obtained by summing the SED's of all $T$ points from the overall data centroid. If $D_i$ is the SED between the centroid of cluster $i$ and the grand centroid, and $N_i$ is the number of elements in this cluster, then the between cluster sum of squares ($SSBC$) is defined as

$$SSBC = \sum N_i * D_i.$$

the identity

$$GTSS = SSWC + SSBC.$$
then follows. During any particular fusion cycle, GTSS is a constant. However, SSWC ranges, in $t-1$ steps, from zero to GTSS, while the SSBC decreases, $t-1$ times, from GTSS to zero. It is therefore seen that minimizing the SSWC is identical, in any fusion, to maximizing the SSBC.

**Information**

Information is best applicable when data is either ordinal or binary, such as in ecological surveys where the recognition of outsiders is less important than the recognition of patterns. It is weakest if the data is mainly continuous, as it is group-size dependent and doubtfully conducive to reallocation (Silliams, Clifford, and Lance, 1971).

In 1968, Lance and Williams introduced an information-based cluster program, named DIVINF, as a counterpart to their agglomerative approach, MULTBET. Both of these programs worked well, but neither was complete. Thus, in 1971 they extended the capacities of their previous programs to include mixed data, i.e. continuous and binary. This extension is called MULTDIV, and is a divisive procedure.

The system operates by first dividing all individuals into pairwise sets. Similarity is then based on the gain in information when two groups are re-fused. Groups with the maximum change in information, at each stage in the analysis, are fused together. The method then repeats itself, establishing new sets, dividing them and re-fusing, until all individuals are in only one set. If data is continuous, it is categorized into 8 states and handled as if it were 8-state ordinal data.
Orloci's information procedure will be discussed in a little more detail, as it is used further on in this study. Similarity here is based on the mutual information, as defined by Shannon in 1948, between two subsets of the population. A data matrix, X, is established, where the R rows represent individuals and the C columns are the attributes. Three agglomerative models, each with a different definition of dissimilarity, are presented by Orloci. The second definition is used in the MINFO program, which this author will use for comparative purposes later in the present study. Dissimilarity = $d_{xy}$ below.

\[
\text{INF1: } \Sigma (I(F_j;F_k) - I(F_j) - I(F_k)) = d_{jk}.
\]

\[
\text{INF2: } I(\text{rows}; \text{columns})_{jk} - I(r;c)_{j} - I(r;c)_{k} = d_{jk}.
\]

\[
\text{INF3: } I(F_j,F_k) - I(F_j;F_k) = d_{jk}.
\]

Here, $F_i$ is an $s_i$ valued vector which summarizes the values in any row or column, as frequencies, such that 3 properties hold.

1. $\Sigma f_{ij} = n$, summing $j$ from 1 to $s_i$.
2. $p_{ij} = f_{ij} / n$.
3. $\Sigma p_{ij} = 1$, summing $j = 1$ to $s_i$, where $f_{ij}$ is a frequency in row $i$, column $j$ of the $X$ matrix.

A pairwise heterogeneity matrix, D, is then calculated as prescribed by the preceding definitions. If INF1 is used, the emphasis is on the total information in the subsets $X_j$, $X_k$, and $X_{jk}$. INF2 can be used if the entries in $X$ are binomial counts, as the mutual row-column information is used. Model INF3 uses the equivocation information separating pairs of individuals, based on the joint information $I(F_j,F_k)$ and the mutual information $I(F_j;F_k)$.
Part 2 of the algorithm proceeds by scanning the D matrix for local minimum values and uniting the corresponding subsets. If INF1 or INF2 is used, D is recomputed using the basic data for the revised subsets. For INF3, an average \( d_{j,k} \) value replaces any fused subsets. In all three models, these steps are repeated until D is of order 1. A dendrogram is easily constructed from the output. The program for this method, MINFO, is a clearly written FORTRAN program. The method is discussed by Orloci (1969).
METHODS

After reviewing the literature, three cluster analysis methods were chosen for detailed investigation. These methods, their authors, and program names follow, with references given previously.

2. Group dispersions. Orloci. MDISP.
3. Information. Orloci. MINFO.

For comparison of these methods, data was generated from known origin and the same data used in each program. A hierarchy of 8 populations was established, as in Figure 5, and data generated to fit it as best as could be done. Attributes on which populations differ appear outside the fork at which those populations separate, common attributes appearing within the fork. For example, since populations A and B differ on only attribute 6, a 6 appears on their dividing fork. All other attributes are common for these two populations and are within the fork.

Multivariate normal data was used, and was generated using Monte Carlo techniques (Hurst, Knop, 1971). The algorithm used requires 3 statistics. First, one needs means for all N attributes of each population. These means locate the centroid of the population in N-space. Secondly, attribute variances are required. These were selected arbitrarily and were kept identical for corresponding attributes of each population. Lastly, an N x N attribute correlation matrix is needed.
This was developed by trial and error, because it was too difficult to preselect values of \( r \) that would make the matrix positive definite. This matrix is shown in Table 3. All populations share this correlation matrix.

As implied above, populations were differentiated only by varying the necessary attribute means. The final magnitude of these mean differences (relative to standard deviations) was established after 3 trial runs, generating 3 populations each trial and seeing how well single linkage in the MINT package could recognize them. Populations for
Table 3. Common attribute correlation matrix used to generate the data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>.601</td>
<td>.534</td>
<td>.294</td>
<td>.623</td>
<td>.652</td>
<td>.557</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>.482</td>
<td>.316</td>
<td>.542</td>
<td>.523</td>
<td>.408</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>.607</td>
<td>.486</td>
<td>.368</td>
<td>.385</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>.585</td>
<td>.424</td>
<td>.483</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>.245</td>
<td>.482</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>.426</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

trial 1, where differing attribute means were only about 1-1/4 standard deviations apart, were not separable. Differing means in trial 2 were therefore set from 3 to 5 standard deviations apart. This resulted in excellent separations, but mean differences of this magnitude are separable by inspection. Therefore, separations of 2 standard deviations were tried in trial 3. Good recognition of all populations, save A, resulted. Population A remained mixed with the others throughout the investigation with the MINT package, reasons being discussed in the results section.

Based on the results of trial 3, it was decided that if populations are to differ on an attribute their respective means for that attribute should be about 2 standard deviations apart. The means finally adopted for program comparison appear in Table 4. The 8 populations used have 8 individuals each, 7 attributes then being measured on each
of the 64 individuals. To speed the mechanics of interpretation, the population individuals were established as follows: Population A is individuals 1 to 8, B is individuals 9 to 16, C is 17 to 24, and so forth. A shuffling of the trial 2 data confirmed hope that the results of the MINT package were not dependent on the order in which the data is read.

Next, the 64 individuals were fed into the single linkage program, where similarity was based on correlation. This method was chosen because it is simple, widely used, and some authors feel it is of much use, to biologists in particular (Jardine and Sibson, 1971).

The second program used to analyze the data was Orloci's MDISP program, which bases the cluster-or-not decision on a within-group dispersion. This technique is similar to that of Burr (1970), which implements

<table>
<thead>
<tr>
<th>Attributes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.0</td>
<td>21.6</td>
<td>18.8</td>
<td>18.1</td>
<td>19.5</td>
<td>20.4</td>
<td>21.7</td>
<td>19.2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>20.0</td>
<td>18.8</td>
<td>21.4</td>
<td>19.2</td>
<td>18.4</td>
<td>13.3</td>
<td>30.1</td>
<td>25.9</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>20.0</td>
<td>21.3</td>
<td>19.3</td>
<td>22.0</td>
<td>21.3</td>
<td>14.6</td>
<td>26.9</td>
<td>25.3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>20.0</td>
<td>19.6</td>
<td>11.6</td>
<td>29.7</td>
<td>33.8</td>
<td>41.2</td>
<td>49.5</td>
<td>56.8</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>20.0</td>
<td>20.2</td>
<td>23.8</td>
<td>30.1</td>
<td>10.5</td>
<td>38.2</td>
<td>37.7</td>
<td>38.4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>20.0</td>
<td>10.6</td>
<td>29.6</td>
<td>38.4</td>
<td>33.2</td>
<td>43.2</td>
<td>40.2</td>
<td>41.8</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>20.0</td>
<td>19.1</td>
<td>22.7</td>
<td>31.5</td>
<td>10.8</td>
<td>40.7</td>
<td>39.9</td>
<td>42.1</td>
<td></td>
</tr>
</tbody>
</table>
an incremental sum of squares method. MDISP is a clearly written FORTRAN program and is easily revised or extended if desired.

The third program used, MINFO, was also written in FORTRAN by Orloci. It utilizes the most sophisticated technique for clustering, that of information. Other authors of information programs are Lance and Williams (1971), who also do a great deal of publication on the cluster analysis field in general.

To aid in the interpretation of the single linkage dendrogram, this author wrote a program (GRPSIM) to calculate the correlation between the groups which the dendrogram prescribes. The basic statistic is Spearman's sums of variables correlation coefficient, detailed by Sokal and Sneath (1963). It is felt that the correlation between groups is indicative of relative positions on a dendrogram, somewhat difficult to see from a single linkage dendrogram. However, there is one problem with the statistic. A between-group correlation calculated in this way may be lower (or higher) than any of the individual correlations between a member of group 1 and a member of group 2. These correlations sometimes, therefore, exceed the usual -1 to +1 range, as happened with the present data.
RESULTS AND DISCUSSION

The single linkage output include the original data, the individual correlation matrix, and a phenogram (dendrogram) in graphic and tabular form. Of this, only the phenogram, in Figure 6, will be presented. This is a very difficult diagram to interpret, as the "chaining" of populations defies establishing hierarchies. However, with the aid of the interpopulation correlations in Table 5, a rough hierarchy was established. It was found to be in good agreement with the dendrogram specified by the author. This rough hierarchy was made by forming groups as the correlation increased, much as the program itself does.

The individual correlations used in the linkage program are product-moment correlations, an example of which is calculated below for 2 individuals each with 4 attributes. First the individuals are arranged as two vectors, X and Y, the 4 elements of which are the scores on the 4 attributes. The cross products and sums of squares are then calculated as usual.

\[
r = \frac{(-2)(-2) + (-1)(-1) + \ldots + 1}{\sqrt{10}} = 1.
\]

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Individual 1(X)</th>
<th>Individual 2(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Average</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 6. MINT Dendrogram for single linkage. Individuals 1 to 8 (population A) are scattered. Notice the tendency to "chain," or construct strings of populations.
Table 5. Group correlations from the program GRPSIM. Individuals in these groups agreed with their respective, specified population individuals. Correlations above 1 are due to "reversal." See Sokal and Sneath, 1963.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-.66</td>
<td>.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-.56</td>
<td>-.23</td>
<td>.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-.40</td>
<td>.24</td>
<td>.91</td>
<td>.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-.31</td>
<td>-.08</td>
<td>.78</td>
<td>.38</td>
<td>.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-.26</td>
<td>-.15</td>
<td>1.35</td>
<td>.89</td>
<td>1.62</td>
<td>.99</td>
<td></td>
</tr>
</tbody>
</table>

The r value of 1 in the above example indicate that the two individuals are extremely similar.

Population A could not be recognized as a population with the simple linkage program, simply because its intrapopulation correlations were so low. Although these values ranged from -.91 to .81, most of them fell between -.4 and .4. All other populations had correlations centering around .8 and .9, the average increasing as the range of differing attributes increased.

After extensive testing of the programs involved, it was found that these low population A correlations were correct and unavoidable. The basic explanation is this: Remembering that all attribute means for population A were set at 20, no individuals "go anywhere" on these attributes and can therefore not be correlated with each other. Intuitively, correlation necessitates varying together, and if no varying occurs,
neither should a high correlation. These ideas are next illustrated by a simple case where e populations are examined, each with 2 individuals and 4 attributes per individual. These populations are diagramed in Figure 7, where probability circles are drawn to locate where an observation on that attribute will most likely fall. One circle occurs for each attribute; the axes are then for each of the individuals.

The first diagram shows that individuals of population 1 have identical expected values for all attributes. When a bivariate sample is taken for individuals 1 and 2 on attribute 1 and plotted on the diagram, the point falls at random within the first probability circle. When a similar sample is taken for any other attribute, it falls at random within its respective probability circle, all of which are identical for population 1. Thus, all points fall at random within the same probability space and will tend to give a low correlation.

The second diagram shows the probability circles for 3 identical attribute means and one different one; this being population 2. Here, a regression line can be fitted, limited by the dotted lines shown. A better correlation in this population, over the first, is inherent.

In population 3, diagram C, all attribute means differ and extend farther into space than they do for either of the other 2 populations. As the fitted regression line is more restricted here (the explained variability increases) there is a great tendency for a high correlation. Generating individuals, in any population, with the same attribute mean values determines the marked linearity in this case.
Individual 2

(A) Poor correlation

Individual 2

(B) Good correlation

Individual 2

(C) Excellent correlation

Figure 7. The correlation between 2 individuals. The correlation increases as the range of differing attributes increases. Variability of the regression line is shown by the dotted lines in Figures B and C. All circles are really of the same size, drawn larger only for illustration.
As the regression line fits the data better as the range of the attribute means increases, correlations should also increase. This happened very noticeably with the present data, the definitions below showing the relationship of correlation and regression.

**Definition 1:** Correlation coefficient \( \rho = \frac{\text{cov}(X,Y)}{(\sqrt{\text{V}(X)} \sqrt{\text{V}(Y)})} \).

**Definition 2:** Regression function \( E(Y|X,\mu,\sigma) = \frac{\text{cov}(X,Y)}{\text{V}(X)} = \frac{\rho \sqrt{\text{V}(Y)}}{\sqrt{\text{V}(X)}} \).

Orloci's MDISP program works on a cyclic principle, printing all of its information on the groups formed at any cycle, and terminating when all of the data is in one group. The output is in columns, headed as follows: (1) Groups; (2) Within-group dispersion; (3) Average within as a percent and (4) The individuals which form groups at any particular cycle. As cycles progress, more dispersion is allowed per group, i.e. group size increases. No dendrogram is printed out, but it is easy to construct one from the information provided. The dendrogram is found in Figure 8. The cycle levels shown there agree with those of the output. No groups were completely formed in cycles 1 to 4, so these are not shown. The basic statistic in this procedure is the distance between two individuals, with 4 attributes, is calculated. These are the same two individuals used previously for the MINT correlation coefficient illustration.
Figure 8. The MDISP dendrogram. Notice how E is out by itself. Other than this, groups formed much as expected.
The distance between individuals 1 and 2 = \((\sum_{e=1}^{p} (X_{e1} - X_{e2})^2)\), with \(p = 4\) here. This in turn is equal to \((3-6)^2 + (4-7)^2 + \ldots + (6-9)^2 = 27\). This distance would then be used to calculate a within-group dispersion for this pair, as described in the literature review. As individuals join this group, the within-group dispersion increases accordingly.

As noticed, the author-prescribed dendrogram and that of MDISP do not agree entirely. To find explanation for this, part of the MDISP program was used to print out distances between the 8 populations, now defined only by those means used to generate the data. Judging by these distances, found in Table 6, and their configurations in Figure 9, populations were not originally oriented in space as intended. In particular, the following inconsistencies were noticed.

1. Population E is a great distance from most populations, especially D, to which it should be closest.

2. Both A and D are very close to G.

Looking at the data in this light, the MDISP dendrogram is explainable. It is apparently more realistic than the author-prescribed one. Rough schematics of a dendrogram for the means-only run were constructed from a listing of pairs formed at increasing distances, as taken from Table 6.
Table 6. Population distances, using attribute means as the scores

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>33</td>
<td>56</td>
<td>62</td>
<td>153</td>
<td>151</td>
<td>59</td>
<td>104</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>147</td>
<td>161</td>
<td>257</td>
<td>251</td>
<td>126</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>62</td>
<td>240</td>
<td>167</td>
<td>126</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>153</td>
<td>37</td>
<td>27</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>230</td>
<td>138</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>52</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These schematics agreed with the group relationships implied by the full-data, MDISP dendrogram, implying good agreement between input and output relationships.

The information program, MINFO, has two sections of output. The first has the new group numbers, which old groups have fused to form this new group, and the individuals composing the new group. The second section contains the information data. In addition to group identification, it adds the mutual information in the new group, the mutual information in the old groups, and the change in the mutual information resulting from the fusion of the two old groups.

Mutual information will now be calculated for the two individuals previously used for method illustration. In the equation below, where $I(X)$ means the information in group $X$, both $I(1)$ and $I(2)$ equal zero,
Figure 9. Relative population distances. Note the relative position of E in all the stars.
as there is no information in an individual. Thus, the dissimilarity between individuals 1 and 2 is calculated by

\[ d_{1,2} = I(1,1) - I(1) - I(2) = I(1,2). \]

In equation form, where \( X_{ij} \) is the \( i \)th attribute of the \( j \)th individual, the mutual information in 1 and 2 together becomes

\[
\sum_{j} \sum_{i} \frac{X_{ij}}{X_{i}} \cdot \frac{S_{ij}}{X_{j}} - \log \left( \frac{(20+32) \cdot 3}{3+6} \right) \cdot \frac{20}{3} + \frac{(20+32) \cdot 6}{3+6} \cdot \frac{52 \cdot 9}{15} + \ldots + \frac{32}{32} = 0.094
\]

The dendrogram for this program, Figure 10, must be manually constructed and disagrees with both the prescribed and the MDISP dendrograms. Notice the position of population E and the relative levels at which groups form and/or fuse. To investigate the consistency of this dendrogram and the information input data, MINFO was manipulated so as to print out the initial mutual information among all population means, as was done with distances in MDISP. This mutual information matrix is in Table 7. Constructing schematics as before, these mutual information values were found to agree with the overall relationships suggested by the analysis of all 64 individuals.

It was concluded that the differences in the dendrograms of the three methods are most likely attributable to the way the different algorithms "see" the data, and not that one method is necessarily right or wrong. On the other hand, the display of the single linkage method was not very encouraging for its further use with this type of data.
Figure 10. Dendrogram for MINFO. Shows information levels at which groups first form.

Table 7. Mutual information between the population means

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>141</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>227</td>
<td>579</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>266</td>
<td>685</td>
<td>331</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>630</td>
<td>1034</td>
<td>1075</td>
<td>802</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>837</td>
<td>1293</td>
<td>986</td>
<td>287</td>
<td>1394</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>287</td>
<td>621</td>
<td>637</td>
<td>132</td>
<td>820</td>
<td>435</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>519</td>
<td>881</td>
<td>928</td>
<td>210</td>
<td>936</td>
<td>324</td>
<td>54</td>
</tr>
</tbody>
</table>
SUMMARY AND CONCLUSIONS

A short summary of the results of the 3 programs will now be given. The single linkage was the least informative of the 3 programs, for three reasons: (1) It chains populations; (2) It did not locate population A, and (3) it may cluster very different individuals as identical individuals. Although both of Orloci's programs located populations well, this author favors the MDISP program.

Jardine and Sibson (1971, p. 54) feel that chaining is not as serious a defect as it appears to be. They base this on a belief that "the existence of intermediates is often regarded as providing grounds for classifying together seemingly desperate OTU's at a lower level than might otherwise be chosen." This might be true, but, when the chains stretch populations in only one dimensions, alot of information on interpopulation relationships is lost. This information, to this author, is regained somewhat by a hierarchic dendrogram, as can be constructed from output of the other 2 programs.

The inability of the single linkage program to locate individuals of population A as a set demonstrates its second weakness. Although the individuals were very similar to one another in this set, the correlation-similarity option did not recognize them as such. This is not a desirable attribute, and leads one to be cautious of its use.
The third weakness of the single linkage program used here is related to the second. Using correlation as a similarity criterion tends to locate dissimilar individuals as being identical. To demonstrate this, recall the example in this paper where the correlation coefficient was calculated for 2 individuals with 4 attributes. There, the coefficient is 1, but the distance between the two individuals was later calculated to be $\sqrt{27}$. Further, data can be constructed to separate individuals by any large distance, yet the correlation will remain 1!

The information program did an excellent job of locating the individuals of the 8 populations. The dendrogram constructed from it was quite close to that of MDISP, thought to be the best for comparison. The main difference is that, in the MINFO program, E joined the ABC group before it joined the DFGH group, whereas it first joined the DFGH group in the MDISP program. However, there is not a great difference between the EA, ED, and EG informations, indicating that the E population is fairly close to the DFGH group.

There are two objections to the use of information analysis in clustering. First, it is best applied to non-continuous data (Williams, Clifford, and Lance, 1971) and doubtfully conclusive to reallocation. Secondly, it is not easy to deal with, because of its obscure geometry. It is doubtful that many researchers feel comfortable with it for this reason, and will tend to shy away from it.

The MDISP program has two nice features. First, it is very clearly written and easy to manipulate or add to if necessary. Secondly, the use of Euclidean-distance-similarity is easy to cope with and lends itself to the construction of data structures.
The MDISP program located the 8 populations exactly as the individuals were specified when generated. Although the MDISP dendrogram disagreed with the one specified by the author, it is now believed that the MDISP dendrogram is the more realistic of the two, and that the disagreement is not a defect of the program. A few recommendations for further exploration in this area will finish this paper. For simplicity, the recommendations will be listed.

1. Recall the 3 trials used to examine how close attribute means could be before single linkage could not distinguish populations. In further work, one might design a factorial set up to investigate several factors simultaneously. Combinations of attribute correlation matrices, attribute means and standard deviations, data types, numbers of attributes, and number of individuals might be examined.

2. Seven attributes are not necessary. Probably 4 or 5 would suffice, as would 4 or 5 populations.

3. Purely binary data could be examined.

4. One might investigate the close relationship between using correlation and using distance as a similarity criterion. Recall that the correlation is a function of the angle between these two individuals.
LITERATURE CITED


VITA

John C. Klingel

Candidate for the Degree of

Master of Science


Major Field: Applied Statistics

Biographical Information:

Personal Data: Born at East Cleveland, Ohio, September 17, 1948, son of Arthur R. and Helen M. Klingel; married Sarah K. Crowther, June 11, 1971; two dogs—Max, and Truffle.

Education: Graduated from Solon High in 1966; received a Bachelor of Arts degree from Hiram College, majoring in General Science, in 1970; completed requirements for the Master of Science degree at Utah State University in 1973.

Professional Experience: 1971 to 1973, involved in the set-up, analysis, and interpretation of many research projects while a graduate student at Utah State University.