PSYCHOMETRIKA—VOL. 43, NO. 1
MARCH, 1978

U-STATISTIC HIERARCHICAL CLUSTERING

ROY G. D'ANDRADE

UNIVERSITY OF CALIFORNIA, SAN DIEGO

A monotone invariant method of hierarchical clustering based on the Mann-Whitney U-statistic is presented. The effectiveness of the complete-link, single-link, and U-statistic methods in recovering tree structures from error perturbed data are evaluated. The U-statistic method is found to be consistently more effective in recovering the original tree structures than either the single-link or complete-link methods.

Key words: clustering, non-parametric.

Hierarchical clusters consist of nested sets of objects such that any pair of sets are either disjoint or one includes the other. Hierarchical clusters may be represented by a rooted tree structure and correspond to a distance metric which satisfies the ultrametric inequality [Johnson, 1967; Hartigan 1975]. Figure 1 gives an example of a tree structure for five objects (a, b, c, d, e). This tree structure fits a set of distance measurements in which the two closest pairs are (a, b) and (c, d); the next closest pairs are (e, c) and (e, d); while (a, c), (a, d), (a, e), (b, c), (b, d) and (b, e) are the most distant pairs. Each numbered node of the tree represents a cluster of objects with the root node representing the most inclusive cluster, and the terminal nodes representing single-object clusters.

In the social and biological sciences tree structures are potential representations for a number of different kinds of phenomena, such as historically related languages, biological and folk taxonomies, phrase structure grammars, and part-whole relationships [Kruskal and Carroll, 1969, pp. 639–671]. Since in most empirical situations the structure of the set of objects is unknown, a structure must be constructed from the given set of measurements. When there is no measurement error, and when the proximity values among objects directly correspond to a tree structure, then any one of a number of clustering algorithms will be equally effective. At present the best known and most widely used clustering techniques are the single-link method (also called the *minimum distance, nearest neighbor* and *connectedness method*) and the complete-link method (also called the *maximum distance, furthest neighbor* and *diameter method*).

The complete-link method consists of:

a. Searching a symmetric matrix of distance measurements to determine the least distant pair of objects.
b. Once found, classing these objects together into a cluster.
c. Constructing a new matrix of distance measurements by selecting from the measurements of the two objects with all other objects in the matrix the most distant measurement. This maximum distance measurement then becomes the measurement which represents the relation of the cluster to the other objects.
d. Repeating the process until all objects and clusters of objects are merged into a single cluster.

Requests for reprints should be sent to Roy G. D'Andrade, Department of Anthropology, University of California, San Diego, La Jolla, CA 92093.

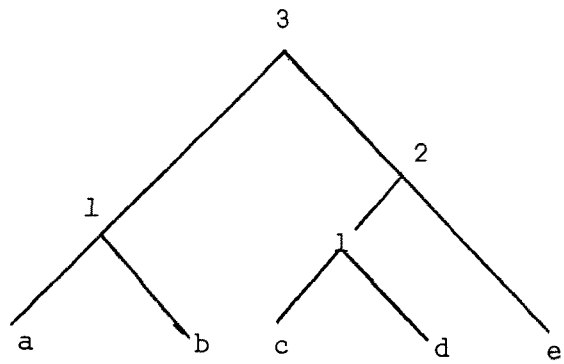


FIGURE 1
Tree structure

The single-link or minimum distance method uses the same operations as the complete-link method except that the minimum distance measurement is selected to represent the proximity relation of the new cluster-object to every other cluster-object.

Neither the single-link nor the complete-link procedure necessarily constructs the best possible tree for a given set of input measures. However, Johnson has shown that if the original data are ultrametric, then both methods will yield the same results [Johnson, 1967]. Unfortunately, for empirical data both methods usually do not give the same results.

With respect to the problem of how to arrive at a single best solution, Johnson [1967, pp. 253–254] makes the following comments.

> ". . . in methods like our Minimum and Maximum Methods, the merging of the two clusters depends upon a single similarity value (viz., the least or greatest in the appropriate set). [Sokol and Sneath] suggest that, for greater robustness of the solution, it may sometimes be desirable to use some sort of average value instead. As we have already noted, to base such a procedure upon averages of the more obvious types is to lose the invariance, sought here, under monotone transformation of the similarity values. . .
>
> "Nevertheless, when this seems desirable, the methods described here can be (and, indeed, have been) modified to yield solutions intermediate between those obtained by these two extreme methods. J. D. Carroll [Note 1] has suggested an average method based upon medians which, of course, do have the desired property of monotone invariance. The main problem, in the case of medians, is the choice of an appropriate procedure for dealing with the ambiguities that tend to arise when two or more of the initial similarity estimates are tied."

Developing robust procedures is a basic problem in hierarchical clustering. A major drawback of the single-link and complete-link methods is that while they are monotone invariant, cluster formation is fairly sensitive to measurement error. The procedure presented in this paper appears to be less sensitive to measurement error than either the single-link or complete-link methods. This procedure is based upon a non-parametric measure of association, the U-statistic, as defined for the Mann-Whitney test [see Seigel 1967].

To illustrate how the U-statistic is applied to cluster analysis, we begin with a situation in which a matrix of distance measurements have been grouped by some method to the point where there are three clusters. (In the example given below the distance measurements have been selected so that almost any technique of clustering will yield these three clusters.) The first cluster consists of the objects a, b, and c, the second cluster of the objects d, e, f, and g, and the third cluster of the objects h and i. Table 1 presents the distance measurements for this example.

The immediate problem is to decide how to proceed with the clustering. From the data given in Table 1 different clusters will be formed depending on whether the maxi-

TABLE 1

Distance Measurements

		Cluster I			Cluster J				Cluster K	
		a	b	c	d	e	f	g	h	i
Cluster I	a	0	30	45	80	95	50	88	55	61
	b	30	0	40	97	46	92	83	62	50
	c	45	40	0	76	96	85	69	61	81
Cluster J	d	80	97	76	0	10	30	40	60	57
	e	95	46	96	10	0	20	35	56	72
	f	50	92	85	30	20	0	25	78	58
	g	88	83	69	40	35	25	0	73	60
Cluster K	h	55	62	61	60	56	78	73	0	10
	i	61	50	81	57	72	58	60	10	0

mum, minimum, or median method is used. Table 2 presents the distance measurements for between cluster pairs of objects. Using the maximum distance method, clusters J and K would be grouped together. Using the minimum distance method, clusters I and J would be grouped together. Using the median method, clusters J and K would be grouped together.

The U-statistic procedure is a method used to determine, for two samples of ordinal scores, which sample has the overall greater value. As a notational convenience to illustrate computation of the U-statistic, let 'I × J' represent the set which is composed of all the distance measurements occurring between those objects where one object comes from Cluster I and the other object comes from Cluster J. To determine, given any number of clusters, which two clusters are closest together we calculate whether a randomly drawn distance measurement from one pair of clusters is likely to be less than a randomly drawn distance measurement from another pair of clusters. In the example above there are three pairs of clusters: I & J; I & K; and J & K. To determine whether a randomly drawn distance measurement from I × J would be likely to be less than a randomly drawn distance measurement from J × K all of the distance measurements from the I × J set are compared to all the distance measurements from the set J × K. The U-statistic is computed by taking the number of times the distances from the I × J set are less than the distances of the J × K set. If the I × J distances are less than the J × K values in the majority of the comparisons, then clusters J & K are U-closer than clusters I & J. Table 3 presents these comparisons for the distance measurements for I × J, I × K, and J × K.

Using this method we find clusters I & K are U-closer than I & J, clusters J & K are U-closer than I & J, and clusters I & K are very slightly U-closer than clusters J & K. Using 'C' to stand for the U-statistic measure of closeness, $C(I, K) > C(J, K) > C(I, J)$.

In the example given above, the clustering procedure began with three fairly large and

TABLE 2

Rank Ordered Distance Measurements

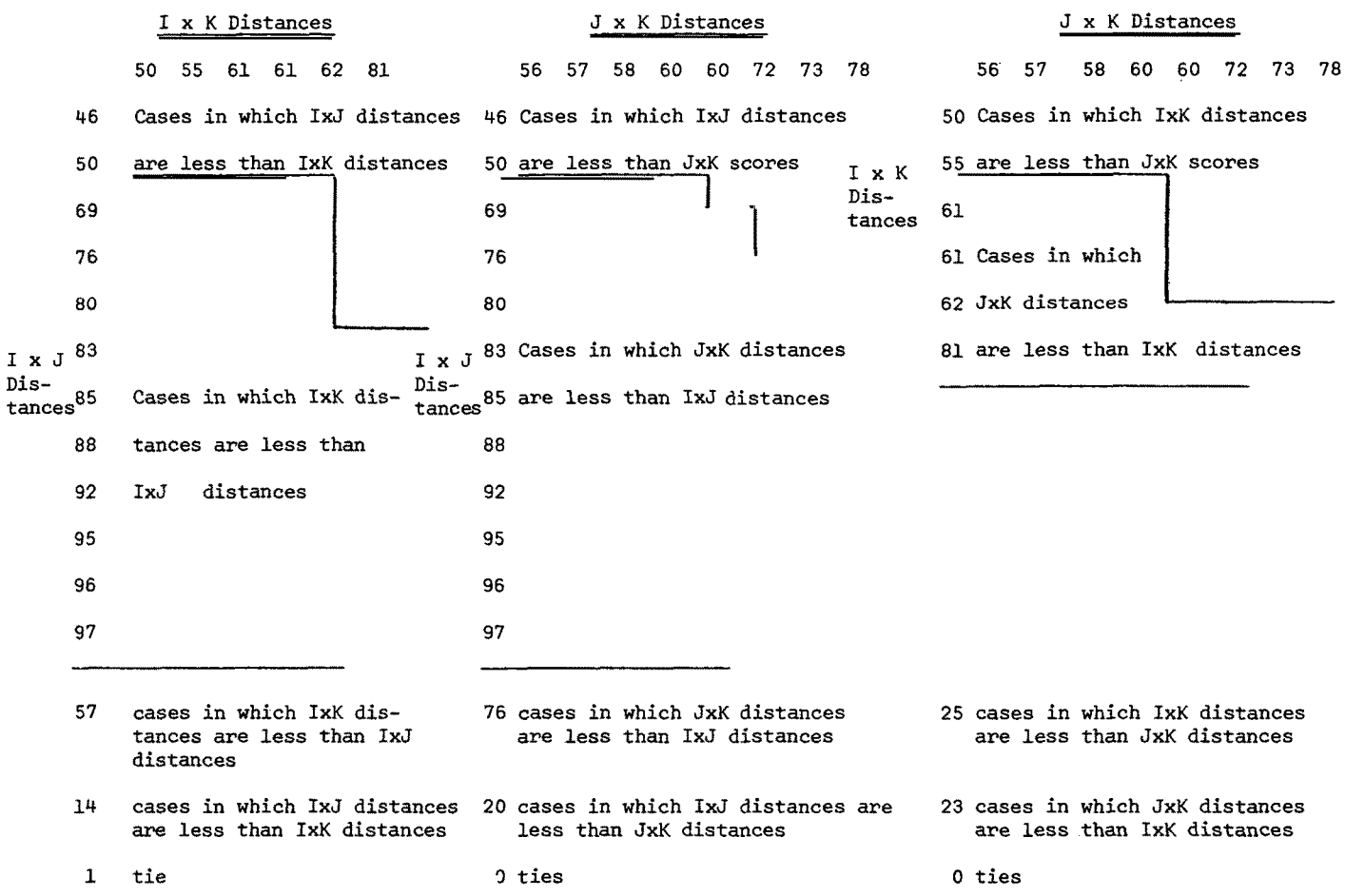
	Cluster I & Cluster J		Cluster I & Cluster K		Cluster J & Cluster K	
	Distance	Pair of Objects	Distance	Pair of Objects	Distance	Pair of Objects
least → minimum distance	46	b & e	50	b & i	56	e & h
	50	a & f	55	a & h	57	d & i
	69	c & g	61	c & h	58	f & i
			median →			
	76	c & d	61	a & i	60	d & h
					least → median	
	80	a & d	62	b & h	60	g & i
median →	83	b & g	81	c & i	72	e & i
	85	c & f			73	g & h
	88	a & g			least → 78 maximum distance	f & g
	92	b & f				
	95	a & e				
	96	c & e				
	97	b & d				

already formed clusters. In an actual cluster analysis the procedure begins with the matrix of distance measurements between the initial objects. Unlike the single-link and complete-link methods, the U-statistic cluster procedure is too tedious to carry out by hand for all but the smallest data sets. A computer program for the U-statistic method has been written in Algol. This program begins with a cluster list, the initial clusters consisting of just the original objects. Using the U-statistic procedure on the input data matrix of distance measures, every pair of clusters on the cluster list is compared in order to find the closest pair of clusters. These two clusters are then combined into a single cluster and placed on the cluster list, and the old pair of clusters removed from the cluster list. U-statistic ties between two different pairs of clusters are broken by selecting the pair of clusters which have the least distant pair of original objects (i.e., the complete-link method). Clusterings are printed out at each merger. This process is repeated until all clusters are merged into a single cluster. From the history of the clustering process a tree structure may be drawn.

The U-statistic method yields what Boorman and Oliver term a 'bare' tree, in that there are no direct measurement related evaluation levels for cluster nodes [Boorman & Oliver, 1973]. However, a value may be given to each node by counting the maximum number of nodes below that node: terminal nodes have a value of zero; nodes which directly precede only terminal nodes have a value of 1; nodes which directly precede a node of value 1 but no node of higher value have a value of 2; etc. The *cluster rank distance* between any two objects at terminal nodes is then the value of the lowest common ancestral node connecting these two objects.

TABLE 3

Comparison of Set x Set Distance Measurements



Evaluation of Goodness of Fit

It should be mentioned that the U-statistic procedure is not the only monotone invariant technique which uses more than a single value as a basis for cluster merging. Hubert has presented a procedure, called the objective function method, which is based on the calculation of a monotone invariant goodness of fit index. This index is composed of the ratio of the number of actually occurring 'discrepancies' to the number of possible 'discrepancies.' A 'discrepancy' occurs at a given level of clustering when the distance value for some pair of objects contained within a cluster is larger than the distance value for some pair of objects belonging to distinct clusters [Hubert, 1973]. At each level of clustering the goodness of fit index is calculated for the merger of every pair of clusters; the pair of clusters with the best (i.e., smallest) goodness of fit figure are then merged into a higher level cluster.

Hubert's objective function differs from the U-statistic procedure in that decisions about the merger of clusters are based on evaluations of the entire distance matrix rather than just the object distances between clusters. Since the count of discrepancies is based on all pairs of objects in which one pair of objects come from within a potential cluster and the other pair comes from distinct clusters, the goodness of fit index can be affected by anomalous values not directly connecting the clusters being considered. Because the objective function method counts more discrepancies when considering the formation of a large cluster than when considering the formation of a small cluster, a bias is introduced which is not completely adjusted for in the denominator of the index, which takes into account the total number of possible discrepancies. This sometimes results in the merger of small clusters when large clusters are closer together according to the U-statistic measure (or any other pair based measure of closeness). Since the objective function is not a pair based comparison procedure, it is not surprising that under certain conditions it yields outcomes which differ from those of the U-statistic procedure.

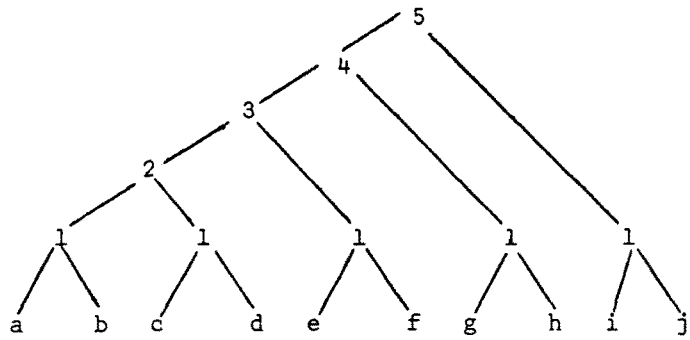
The objective function index, in slightly modified form, has proved useful in comparing the effectiveness of different hierarchical clustering procedures. To generalize the objective function index to measure the relation between a matrix of distance values and a hierarchical clustering structure, a discrepancy is defined as a pair of pairs of objects where the distance value for one pair of objects is greater than the distance value for the other pair of objects but the cluster rank of the first pair of objects is less than the cluster rank of the second pair of objects. A 'concordancy' is defined as a pair of pairs of objects where the distance value for one pair of objects is less than the distance value for the other pair of objects and the cluster rank of the first pair of objects is less than the cluster rank of the second pair of objects [Hubert, 1974]. Putting aside those pairs with ties on either distance values or cluster rank, the number of possible discrepancies equals the number of actual discrepancies, D, plus the number of actual concordancies, C; the objective function index equals $D/(C + D)$. This measure is very closely related to Goodman and Kruskal's gamma, which is $(D - C)/(D + C)$ [Baker and Hubert, 1975].

One advantage to using gamma is that in comparing input distance values and cluster rank, gamma has a direct probability interpretation: given randomly selected object pairs untied on either cluster rank or distance values, gamma equals the probability of a consistent ranking on both scales minus the probability of an inconsistent ranking. When gamma is equal to 1.0 then there are no discrepancies. Using gamma as an evaluation statistic, Hubert found that the complete-link method gave cluster structures which generally had a better fit to the input measures than the single-link method [Hubert, 1974].

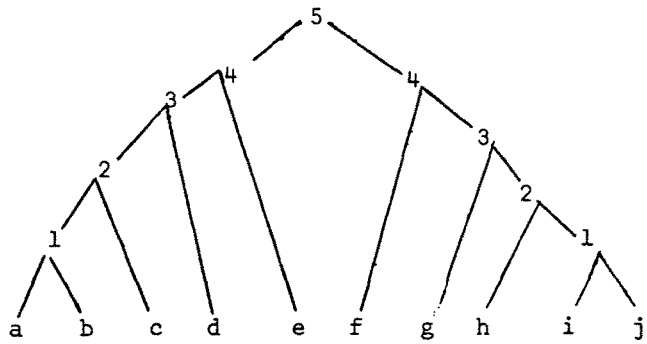
Evaluation of the U-Statistic, Complete-Link and Single-Link Methods of Cluster Analysis

In order to compare the U-statistic clustering method to the single-link and complete-link methods with respect to their effectiveness in recovering known tree structures from

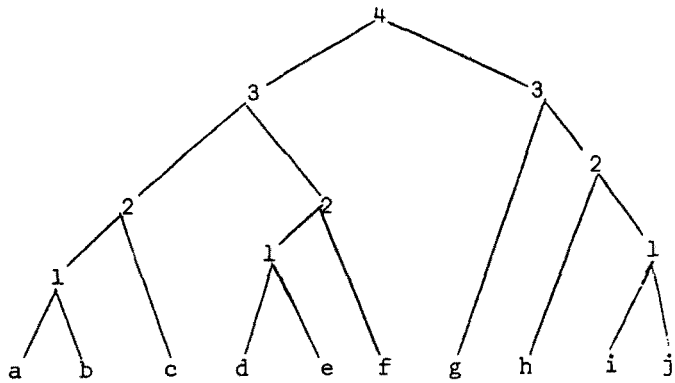
I



II



III



IV

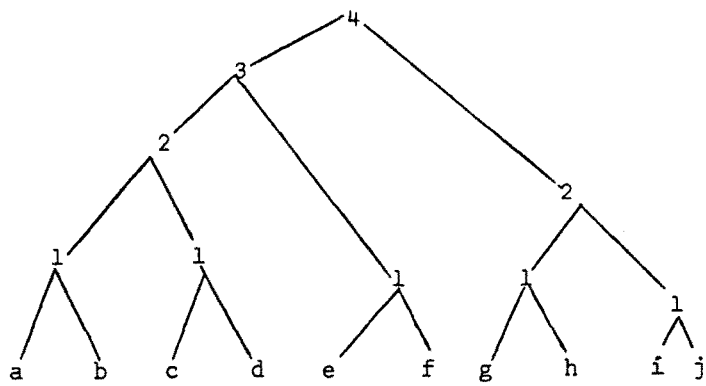


FIGURE 2
Test tree structures

TABLE 4

Means and Standard Deviations for Gamma Between Original Tree Distances and Cluster Rank Distances, 40 Replications per cell.

	Cluster Method					
	U-Statistic		Complete-Link		Single-Link	
Level of Error	Mean Gamma	S.D.	Mean Gamma	S.D.	Mean Gamma	S.D.
Low	.94	.14	.88	.22	.82	.30
Low-Medium	.88	.19	.66	.33	.61	.34
High-Medium	.72	.34	.49	.36	.47	.41
High	.49	.35	.43	.40	.18	.39
Overall Mean	.75		.61		.52	

error perturbed data, four different trees of ten objects were constructed from which cluster rank distances were taken (see Figure 2). To each of the four original distance matrices 'high,' 'high mid,' 'low mid,' and 'low' degrees of normally distributed random error were added, making a total of sixteen different conditions. The cluster rank distances range from 1 to 5, with a standard deviation of 1.35 for Tree 1, 1.21 for Tree 2, .95 for Tree 3, and 1.08 for Tree 4. For a 'high' degree of error, the standard deviation for the error term was 1.40, for a 'high mid' 1.14, for a 'low mid' .90, and for 'low' .60.

The U-statistic, single-link and complete-link methods were run for ten trials in each of the sixteen conditions. The accuracy with which each method recovered the original tree structure from the error perturbed data was evaluated using gamma to measure the degree of concordance between the original error free distance measures and the cluster rank

TABLE 5

Means and Standard Deviations for Gamma Between Original Tree Distances and Cluster Rank Distances, 40 Replications Per Cell.

	U-Statistic		Complete-Link		Single-Link	
Data Set	Mean Gamma	S.D.	Mean Gamma	S.D.	Mean Gamma	S.D.
Tree 1	.88	.19	.64	.34	.75	.28
Tree 2	.71	.34	.63	.32	.47	.41
Tree 3	.64	.35	.46	.47	.35	.45
Tree 4	.78	.34	.71	.36	.50	.47
Overall Mean	.75		.61		.52	

distances. Hubert's use of gamma was modified in order to take account of the fact that in the construction of a 'bare' tree by means of the U-statistic, cluster rank distances are not strictly comparable in disjoint sub-trees. That is, given two pairs of objects from disjoint subtrees, the relation between the cluster rank distances may be changed if new objects are introduced into the cluster analysis. When the two pairs are not disjoint, the relation between cluster ranks will remain the same, no matter how many new objects are introduced, since one pair must come from a cluster which is a proper subset of the cluster from which the other pair comes. In the modified calculation of gamma, only those pairs of pairs of objects which are not from disjoint sections of a tree are tested for discrepancies. The results are given in Tables 4 and 5. No interaction between type of tree and level of error was detected for any of the clustering methods.

These results indicate that the U-statistic method is consistently more effective in recovering the original data structures than either the single-link or complete-link methods. The complete-link method is more effective than the single-link method for all comparisons except Tree 1 which consists of 'chained' pairs, supporting the previous findings of Hubert [1973] and Baker and Hubert [1975].

Viewing the contemporary situation concerning the many competing varieties of cluster analysis, it appears unlikely that the relations between different methods and data types will be untangled solely by formal analyses. A promising approach to understanding how these different methods operate is to compare the effectiveness of different methods across a variety of data types. Using such an approach, it is the conclusion of this paper that a method which uses all the available measurements concerning the pair-wise relations between clusters, such as the U-statistic procedure, provides a more robust method for obtaining hierarchical clustering than single-value methods such as the complete-link method or single-link method.

REFERENCE NOTE

1. J. D. Carroll, personal communication, 1977.

REFERENCES

Baker, F. B. & Hubert, L. J. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association,* 1975, *70,* 31–38.

Boorman, S. A. & Oliver, D. C. Metrics on spaces of finite trees. *Journal of Mathematical Psychology,* 1973, *10,* 25–59.

Hartigan, J. A. *Clustering algorithms.* New York: John Wiley & Sons, 1975.

Hubert, L. J. Monotone invariant clustering procedures. *Psychometrika,* 1973, *38,* 47–62.

Hubert, L. J. Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association,* 1974, *69,* 698–704.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika,* 1967, *32,* 241–253.

Kruskal, J. B. & Carroll, J. D. *Geometric models and badness of fit functions. Multivariate analyses.* San Francisco: Academic Press, 1969.

Seigel, S. *Non-parametric statistics.* New York: McGraw-Hill, 1956.

Sokal, R. R. & Sneath, P. H. A. *Principles of numerical taxonomy.* San Francisco: Freeman, 1963.

Manuscript received 1/28/76
First revision received 9/13/76
Final version received 5/11/77