A DYNAMIC PROGRAMMING ALGORITHM FOR CLUSTER ANALYSIS

Robert E. Jensen

University of Maine, Orono, Maine

(Received July 25, 1968)

This paper considers the problem of partitioning N entities into M disjoint and nonempty subsets (clusters). Except when both N and $N—M$ are very small, a search for the optimal solution by total enumeration of all clustering alternatives is quite impractical. The paper presents a dynamic programming approach that reduces the amount of redundant transitional calculations implicit in a total enumeration approach. A comparison of the number of calculations required under each approach is presented in Appendix A. Unlike most clustering approaches used in practice, the dynamic programming algorithm will always converge on the best clustering solution. The efficiency of the dynamic programming approach depends upon the rapid-access computer memory available. A numerical example is given in Appendix B.

CONSIDER the problem of partitioning N entities into M disjoint and nonempty subsets (or clusters) in such a manner that entities within clusters are 'more homogeneous' than entities between clusters. Methods of clustering or grouping entities according to multiple characteristics possessed by each entity appear in the literature under various synonyms, e.g., cluster analysis, grouping methods, classification theory, numerical taxonomy, and clump theory. Clustering problems cut across various disciplines. Entities to be grouped together may be persons, plants, traits, institutions, structures, fields, stars, cities, companies, etc. Considerable attention is given to this matter in the behavioral and life sciences. Many references are provided at the end of this paper. The objective of the paper is to introduce a clustering algorithm using dynamic programming.

Cluster analysis searches for hidden similarities among different entities that are not immediately obvious. It is similar to, but not exactly like, discriminant analysis. In discriminant analysis the problem is how best to discriminate between given populations that are assumed to be separable in an unambiguous manner by some external criterion. The discriminant research task is one of finding partitions or boundaries that separate individual items in an optimal manner according to a discriminant function.

In cluster analysis, on the other hand, the problem is one of finding groupings whose populations are not known in advance, i.e., the objective is to discover 'natural' clusters of the items based upon some internal cri-

terion. The basic procedure is, first of all, to determine whether the entities fall into clusters on the basis of multiple characteristics, and then to delineate the degree of homogeneity (or compactness) among the clusters.

Criterion of Homogeneity

Implicit in cluster analysis is a 'distance' measure, d_{ij}, between each pair of entities $i, j = 1, 2, \cdots, N$. Let x_{cj} denote the value of characteristic c of entity j for $c = 1, 2, \cdots, p$ characteristics or properties under study. One pairwise distance measure in cluster analysis is the familiar Euclidean metric

$$d_{ij} = [\sum_{c=1}^{c=p} (x_{ci} - x_{cj})^2]^{1/2}. \tag{1}$$

This metric may be standardized by first substracting characteristic means from raw observations and dividing by characteristic standard deviations. Investigators have also transformed Euclidean distances to Mahalanobis generalized distances. Somewhat related measures of distance or association commonly used in cluster analysis are coefficients of association and correlation. Still other distances that might be used are the Chebychev metric

$$d_{ij} = \max_c |x_{ci} - x_{cj}| \tag{2}$$

and the 'taxicab' metric

$$d_{ij} = \sum_{c=1}^{c=p} |x_{ci} - x_{cj}|. \tag{3}$$

Other similarity and distance concepts are discussed in JENSEN [1968a].

Suppose there are n_k entities in the set g_k of entities contained within cluster k of a given alternative for partitioning N entities into M clusters. In the ensuing discussion it will be assumed that

$$W = \sum_{k=1}^{k=M} T(g_k) \tag{4}$$

is the criterion measure of clustering homogeneity, where

$$T(g_k) = (1/n_k) \sum_{i<j \epsilon g_k} (d_{ij}^2) \tag{5}$$

is the 'transition cost' of cluster k. When d_{ij} is a Euclidean metric, the W value can be shown to be pooled within-groups sums of squares for the M groups. The objective in this case is one of finding the clustering alternative that minimizes the pooled within-groups sums of squares, and accordingly maximizes the between-groups sums of squares. Use of the W minimization criterion is extremely common in cluster analysis, although other approaches have also been suggested.

Inefficiency of Total Enumeration

The number of ways in which N entities can be partitioned into M nonempty subsets is given by Stirling's numbers of the second kind, the formula

for which in closed form is

$$S(N, M)=\frac{1}{M!}\sum_{K=0}^{K=M}(-1)^{M-K}\binom{M}{K}K^N, \tag{6}$$

where

$$\binom{M}{K}=\frac{M!}{K!(M-K)!}. \tag{7}$$

In most clustering problems, total enumeration of all feasible clustering alternatives for the optimal solution is out of the question, even with our largest electronic computers. Interestingly, the difficulty is usually one of computing speed rather than rapid-access storage. The number of feasible clustering alternatives may be astronomical for relatively small clustering problems. For example, the number of ways $N=25$ entities can be partitioned into $M=10$ groups is $S(25, 10)=1{,}203{,}163{,}392{,}175{,}387{,}500$.

The inefficiency of total enumeration has led to a gallimaufry of linkage clustering approaches that are computationally more efficient. Many of the methods that have been suggested are reviewed and illustrated in Jensen [1968b]. To date, however, the author knows of no clustering methods (other than total enumeration) that guarantee convergence on the optimal clustering solution. Most linkage methods merely search for the best solution among a small subset of clustering alternatives. Under the various approaches that have been suggested, there is no assurance that the subset includes the optimal solution among all $S(N, M)$ alternatives.

The purpose of this paper is to introduce a means of viewing the clustering problem in dynamic programming form using the W minimization criterion. The dynamic programming algorithm assures convergence on the optimal solution without having to enumerate all clustering alternatives. From a computational standpoint the algorithm eliminates many redundant calculations implicit in total enumeration. Unfortunately, it also requires additional rapid-access storage, and like many other dynamic programming applications, may not be practical in very large problems because of added search time in auxiliary storage. However, for problems of certain dimensions the dynamic programming method suggested here may be practical and yield substantial computer savings.

A DYNAMIC PROGRAMMING MODEL FOR CLUSTER ANALYSIS

Model Formulation

The recursion formula for our dynamic programming formulation (forward algorithm) may be written

$$W_K^*(z)\equiv\begin{cases}0, & \text{for } K=0,\\ \min_y\,[T(z-y)+W_{K-1}^*(y)], & \text{for } K=1, \cdots, M_0,\end{cases} \tag{8}$$

where $M \equiv$ number of disjoint and nonempty subsets into which N entities are to be partitioned,
$K \equiv$ index of stage variable,
$M_0 \equiv M$ if $N \geqq 2M$, and $N-M$ if $N<2M$,
$z \equiv$ state variable representing a given set of entities at Stage K,
$y \equiv$ state variable representing a given set of entities at Stage $K-1$,
$z-y \equiv$ subset of all entities contained in z that are not contained in y,
$T(z-y) \equiv$ is the same as defined previously in (5).

Values of $W_K^*(z)$, as defined in (8), represent the minimum W criterion value for the optimum way to partition entities contained in z into K nonempty and mutually exclusive subsets. This formulation assumes the W minimization criterion. Further explanation of relations in (8) will follow. Initially we will explain how states and feasible arcs connecting these states can be automatically generated. This will be followed by a discussion of how to reduce further the amount of computation necessary. A comparison of computational efficiencies in the various alternative approaches is subsequently provided.

Total Feasible Connecting Arcs and States in a Network

All $S(N, M)$ clustering alternatives can be classified according to various *distribution forms* of M clusters. For example, suppose $N=4$ entities are to be partitioned into $M=2$ clusters. In this case there are $S(4, 2)=7$ clustering alternatives, each of which has one of the following distribution forms:

(i) Distribution form {3} {1}, where three entities are assigned to one cluster and a single entity is assigned to the other cluster.

(ii) Distribution form {2} {2}, where two entities are assigned to each cluster.

The various clustering alternatives in this simple case are shown below for entities 1, 2, 3, and 4.

(i) Alternatives having distribution form {3} {1}:

(1 , 2 , 3), (4)
(1 , 2 , 4), (3)
(1 , 3 , 4), (2)
(2 , 3 , 4), (1)

(ii) Alternatives having distribution form {2} {2}:

$$(1\ ,\ 2), (3\ ,\ 4)$$
$$(1\ ,\ 3), (2\ ,\ 4)$$
$$(1\ ,\ 4), (2\ ,\ 3)$$

In the ensuing discussion it is assumed that components of distribution forms are arranged in descending order, e.g., {3} {1} rather than {1} {3}.

The number of distribution forms may be substantially less than the number of clustering alternatives. For example, there are $S(24, 6) = 6{,}090{,}236{,}036{,}084{,}530$ ways to partition 24 entities into 6 clusters, but these alternatives fall into only 199 unique distribution forms (*see* Note 1).

Explicit generation of distribution forms is unnecessary in our dynamic programming formulation. The concept of distribution form, however, was used in developing the dynamic programming state-generating subroutine. The maximum number of entities, $\max(K)$, at Stage K is equal to the maximum sum of distribution form components from Stages 1 through K inclusively. The minimum number, $\min(K)$, is the minimum sum of these components.

The value of $\max(K)$ at Stage K is

$$\max(K) \equiv N - M + K. \tag{9}$$

If N is an even multiple of M, then

$$\min(K) \equiv K(N/M). \tag{10}$$

Otherwise, when N is not evenly divisible by M, we have

$$\min(K) \equiv \begin{cases} ([N/M]+1)K, & \text{for } 1 \leqq K \leqq N - M[N/M], \\ N - (M-K)[N/M], & \text{for } N - M[N/M] < K \leqq M, \end{cases} \tag{11}$$

where $[N/M]$ denotes the integer portion of N/M. In our previous illustration of Distribution Form {3} {1} we derive: $\max(1) = 3$, $\max(2) = 4$, $\min(1) = 2$, $\min(2) = 4$.

At Stage K of the dynamic programming process, the number of states (*see* Note 2) that can be formulated is

$$NS(K) = \begin{cases} 1, & \text{for } K = 0 \\ \sum_{L=\min(K)}^{L=\max(K)} \binom{N}{L} & \text{for } K = 1, \cdots, M_0. \end{cases} \tag{12}$$

Between successive stages, states are connected by arcs. For instance, consider the example states shown in Fig. 1. As a necessary condition, feasible arcs cannot exist between a state in Stage K and a state in Stage $K+1$ if any entity contained in the Stage K state is not contained in the Stage $K+1$ state for $1 \leqq K \leqq M_0 - 1$ (*see* Note 3). For example, an arc does not exist between States 12 and 5 in Fig. 1, since entity 5 is contained in State 12 and not in State 5. At $K = 0$ there is a dummy state that con-

tains no entities. One arc exists from this state to each of the Stage 1 states.

As defined previously, let $z-y$ represent the set of all entities contained in z that are not contained in y. The transition cost along an arc between a state with y entities and a state with z entities, if the arc exists, is $T(z-y)$. For example, State 5 contains entities 1, 7, 8 that are not contained in State 13, i.e., $(z-y)=(1, 7, 8)$. The transition cost along the interconnecting arc is then $T(1, 7, 8)=(\frac{1}{3})(d_{17}^2+d_{18}^2+d_{78}^2)$, as defined in (5). Similarly, the transition cost on the arc between State 13 and State 10 is $T(5, 7, 8)=(\frac{1}{3})(d_{57}^2+d_{58}^2+d_{78}^2)$.

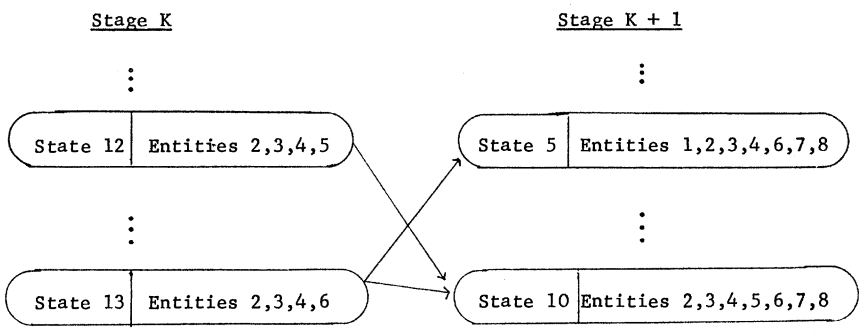

Figure 1

In our dynamic programming formulation, let TFA represent the total number of feasible arcs in the entire network, where

$$TFA \equiv NS(1)+\sum_{K=1}^{M_0-1} TA(K). \tag{13}$$

There are $NS(1)$ arcs connecting the dummy origin with the $NS(1)$ states in Stage 1, where $NS(1)$ is defined in (12). The value $TA(K)$, representing the total number of feasible arcs between Stage K and Stage $K+1$, is given for $K=1, \cdots, M_0$ by

$$TA(K) \equiv \sum_{L=\min(K)}^{L=\max(K)} \sum_{J=1}^{J=\max(K+1)-\min(K)} FA(L, J), \tag{14}$$

where

$$FA(L, J)=\begin{cases} \binom{N}{L}\binom{N-L}{J} & \text{if } \min(K+1) \leqq N+J \leqq \max(K+1), \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Note that M_0, as defined in (8), is the final stage of the dynamic programming process. It would be possible to let $M_0=M$ always. However, in simpler problems where $N<2M$ there must always be at least $N-M+1$ single-entity clusters, i.e., clusters that contain one entity. Single-entity clusters add nothing to W defined in (4). When Stage M_0 is reached under dynamic programming, only single-entity clusters will remain to be

formed. Hence the process may be terminated at Stage M_0 and all remaining entities are then assumed to comprise single-entity clusters.

In (14) and (15), L denotes the number of entities among a class of feasible states at Stage K. There are $\binom{N}{L}$ such states containing L entities since $\binom{N}{L}$ is the number of combinations of N entities taken L at a time. Also in (14) and (15), J depicts the number of entities to be combined with L entities to form a new state at Stage $K+1$. Obviously, the condition $\min(K+1) \leqq L+J \leqq \max(K+1)$ must be satisfied for a state containing $L+J$ entities to exist at Stage $K+1$. If $L+J$ satisfies this condition, there are $\binom{N-L}{J}$ unique combinations of $N-L$ entities that may be added to L entities J at a time.

Reduced Network Formulation

Although the dynamic programming formulation given in (8) can be recursively applied to a network of *TFA* feasible arcs, it is possible in most cases to reduce the number of feasible arcs that must be evaluated because various symmetries create redundancies. In the majority of problems there will be a number of redundant feasible arcs that can easily be eliminated from consideration without affecting convergence on the optimal solution, i.e., certain feasible arcs may, in fact, represent the same clustering alternative.

We will now indicate how redundant feasible arcs can be eliminated quite simply. Let NAT represent the number of arcs remaining after such eliminations, i.e., NAT should be viewed as the maximum number of feasible arcs required to solve the problem using (8). For our purposes, let

$$NAT \equiv NS(1) + \sum_{K=1}^{M_0-1} NA(K). \tag{16}$$

The value $NA(K)$, representing the number of arcs between Stage K and Stage $K+1$, is given by

$$NA(K) \equiv \sum_{L=\min(K)}^{L=\max(K)} \sum_{J=1}^{J=\max(K+1)-\min(K)} A(L, J), \tag{17}$$

where

$$A(L,J)\begin{cases} \binom{N}{L}\binom{N-L}{J} & \text{if } L \neq J \\ \frac{1}{2}\binom{N}{L}\binom{N-L}{J} & \text{if } L = J \\ 0 \text{ otherwise.} \end{cases} \quad \text{and} \quad \begin{cases} \text{Condition I: } \min(K+1) \leqq N+J \leqq \max(K+1). \\ \text{Condition II: } (M-K)J+L \geqq N. \end{cases} \tag{18}$$

Further discussion of these derivations will follow.

We proceed by way of illustration. Suppose $N=7$ entities are to be partitioned into $M=3$ clusters. There are four possible distribution forms as shown below:

$$\begin{array}{lll}\{5\} & \{1\} & \{1\},\\ \{4\} & \{2\} & \{1\},\\ \{3\} & \{3\} & \{1\},\\ \{3\} & \{2\} & \{2\}.\end{array}$$

In this case, $M_0=M=3$ and

$$\begin{array}{lll}\max(1)=5, & \max(2)=6, & \max(3)=7,\\ \min(1)=3, & \min(2)=5, & \min(3)=7,\\ NS(1)=91, & NS(2)=28, & NS(3)=1,\\ TA(1)=602, & TA(2)=28, & \\ NA(1)=427, & NA(2)=28. & \end{array}$$

Because there are 7 entities to be taken $L=3$, 4, and 5 at a time, respectively, at Stage 1, the number of states at Stage 1 is

$$NS(1)=\binom{7}{3}+\binom{7}{4}+\binom{7}{5}=91.$$

There is one arc leading in from the dummy origin to each of these states.

Since $\min(1)=3$ and $\max(1)=5$, the relevant values of L at Stage 1 are $L=3, 4, 5$. Also, since $\max(2)-\min(1)=6-3=3$, the relevant values of J are $J=1, 2, 3$. Using (15), we may then compute for $K=1$ the following feasible arcs:

$$FA(3,1)=0,\ FA(3,2)=\binom{7}{3}\binom{4}{2}=210 \text{ arcs},$$

$$FA(3,3)=\binom{7}{3}\binom{4}{3}=140 \text{ arcs},$$

$$FA(4,1)=\binom{7}{4}\binom{4}{1}=105 \text{ arcs},\ FA(4,2)=\binom{7}{4}\binom{4}{2}=105 \text{ arcs},$$

$$FA(4,3)=0,$$

$$FA(5,1)=\binom{7}{5}\binom{2}{1}=42 \text{ arcs},\ FA(5,2)=0,\ FA(5,3)=0.$$

$$TA(1)=602 \text{ arcs}.$$

Similarly, for $K=2$ we obtain

$$FA(5,2)=\binom{7}{5}\binom{2}{2}=21 \text{ arcs},\ FA(6,1)=\binom{7}{6}\binom{1}{1}=7 \text{ arcs},$$

$$TA(2)=28 \text{ arcs}.$$

Hence the total number of feasible arcs in the entire network is

$$TFA = NS(1) + TA(1) + TA(2) = 91 + 602 + 28 = 721 \text{ arcs.}$$

It will now be shown how 175 of these feasible arcs are redundant and can be eliminated, leaving fewer alternatives to be evaluated in the dynamic programming algorithm.

In general, whenever $L = J$ for a set of $\binom{N}{L}\binom{N-L}{J}$ feasible arcs, $\frac{1}{2}\binom{N}{L}\binom{N-L}{J}$ of these arcs are redundant and may be ignored provided the arcs to be removed are selectively chosen. This is reflected in (18). For example, when $L = J = 3$ at Stage 1 of our illustration such redundancies arise. Consider the two feasible arcs shown in Fig. 2. Each of these two alternatives results in a partitioning of entities $1, \cdots, 6$ into clusters $(1, 2, 3)$

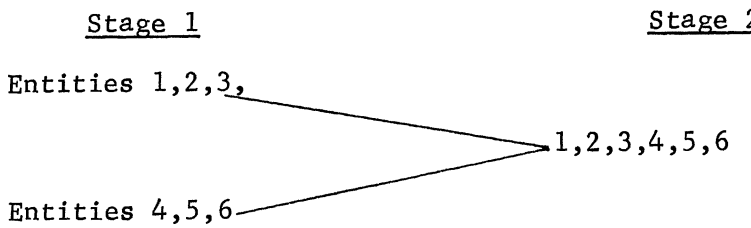

Figure 2

and (4, 5, 6). Hence any one of the arcs shown in Fig. 2 may be ignored as long as the other arc remains. At Stage 1 it was shown previously that there are $FA(3, 3) = 140$ feasible arcs when $L = 3$ and $J = 3$. However, $\frac{1}{2}\binom{7}{3}\binom{4}{3} = 70$ of these arcs may be selectively eliminated as illustrated for just two arcs above.

Another form of redundancy may also arise. For example, at Stage 1 there are $FA(4, 1) = 105$ feasible arcs when $L = 4$ and $J = 1$. In this instance, however, it would be necessary to add 2 entities between Stage 2 and Stage 3, thereby giving rise to a distribution form $\{4\}\ \{1\}\ \{2\}$. However, this distribution form is equivalent to $\{4\}\ \{2\}\ \{1\}$, which corresponds to the $L = 4$ and $J = 2$ alternative at Stage 1. In other words, all $FA(4, 1) = 105$ arcs can be ignored, since these alternatives are ultimately equivalent to the $FA(4, 2) = 105$ arcs. Condition II in (18) eliminates this form of redundancy. In our formulation, the number, J, of entities added between Stage K and $K+1$ need only be greater than or equal to the number of en-

tities to be added in each succeeding stage along any path from this arc to the final stage, as is inherent in the Condition II requirement $(M-K)J+L \geqq N$.

Thus, in our illustration we may compute the following according to (18) for $K=1$:

$$A(3,1)=0,\ A(3,2)=\binom{7}{3}\binom{4}{2}=210 \text{ arcs},\ A(3,3)=\tfrac{1}{2}\binom{7}{3}\binom{4}{3}=70 \text{ arcs},$$

$$A(4,1)=0,\ A(4,2)=\binom{7}{4}\binom{3}{2}=105 \text{ arcs},\ A(4,3)=0,$$

$$A(5,1)=\binom{7}{5}\binom{2}{1}=42 \text{ arcs},\ A(5,2)=0,\ A(5,3)=0.$$

$$NA(1)=427 \text{ arcs}.$$

Similarly, for $K=2$ we have

$$A(5,2)=\binom{7}{5}\binom{2}{2}=21 \text{ arcs},\ A(6,1)=\binom{7}{6}\binom{1}{1}=7 \text{ arcs},\ NA(2)=28 \text{ arcs}.$$

Thus, the maximum number of feasible arcs that must be evaluated in the dynamic programming formulations becomes

$$NAT=NS(1)+NA(1)+NA(2)=91+427+28=546 \text{ arcs},$$

which is $70+105=175$ arcs less than $TFA=721$.

COMPUTATIONAL EFFICIENCY

COMPUTATIONAL EFFICIENCY of dynamic programming relative to a total-enumeration approach in solving clustering problems depends upon the size and type of computer available. For complex problems, comparative efficiency of dynamic programming increases as the amount of rapid-access storage is increased.

One means of contrasting dynamic programming and total-enumeration approaches is to compare the number of transitional calculations required for problems of various dimensions. There are $S(N, M)$ unique clustering alternatives, as defined in (6). Under a total-enumeration approach, each of these alternatives must generally be evaluated separately in terms of M transitional calculations, since there are M distinct clusters. However, in particular cases where $N<2M$, every clustering alternative must contain $N-M+1$ single-entity clusters for which transition costs are zero. Hence, for purposes of comparison it will be assumed that M_0 transitional calculations are required for each alternative, where $M_0=M$ except when $N<2M$

such that $M_0 = N - M$ (*see* Note 4). Let NTT denote the number of total transitional calculations under total enumeration, i.e.,

$$NTT \equiv M_0 : S(N, M). \tag{19}$$

The number of transitional calculations required under dynamic programming is no greater than NAT defined in (16). Appendix A makes a comparison between selected NTT and NAT values, and shows that the ratio of NAT to NTT decreases significantly as the magnitude of the problem is increased. The reason for this can be simply explained. Suppose at Stage K of the dynamic programming process there exists a state

TABLE I

SOME CLUSTERING ALTERNATIVES FOR $N = 7$ ENTITIES IN $M = 3$ CLUSTERS

Alternative	Transitional calculations under total enumeration
1	$T(1,2,3) + T(4,5) + T(6,7)$
2	$T(1,2,4) + T(3,5) + T(6,7)$
3	$T(1,2,5) + T(3,4) + T(6,7)$
4	$T(1,3,4) + T(2,5) + T(6,7)$
5	$T(1,3,5) + T(2,4) + T(6,7)$
6	$T(1,4,5) + T(2,3) + T(6,7)$
7	$T(2,3,4) + T(1,5) + T(6,7)$
8	$T(2,3,5) + T(1,4) + T(6,7)$
9	$T(2,4,5) + T(1,3) + T(6,7)$
10	$T(3,4,5) + T(1,2) + T(6,7)$

containing entities $1, 2, \cdots, n$, where $n \leqq N$. In dynamic programming, we record in memory the optimal way to partition these n entities into K nonempty and mutually exclusive subsets. In subsequent stages for all clustering alternatives in which the same n entities are partitioned into K clusters, dynamic programming makes efficient use of this recorded information, i.e., it is unnecessary each time to recompute all feasible ways to partition these particular entities into K clusters.

For example, suppose $N = 7$ entities are to be partitioned into $M = 3$ clusters. In this case there are $S(7, 3) = 301$ unique clustering alternatives, some of which are listed in Table I for illustrative purposes. Under total enumeration, these 10 alternatives are separately evaluated, each evaluation requiring three transitional calculations as shown in Table I, i.e., these 10 alternatives alone require 30 transitional calculations under total enumeration. Using dynamic programming, however, it takes only 20 tran-

sitional calculations to determine the best partition of entities 1, 2, 3, 4, and 5 into $K=2$ clusters. This optimal solution, say alternative 5 for which $T(1, 3, 5)+T(2, 4)=W_2^*(1, 2, 3, 4, 5)$, is then recorded in memory. Subsequently it requires only one additional transitional calculation to compute $W_2^*(1, 2, 3, 4, 5)+T(6, 7)$, since it is known that alternative 5 is as good as all other alternatives listed above. Hence, from these 10 alternatives alone, the dynamic programming algorithm eliminates $30-21=9$ redundant transitional calculations required in a total-enumeration approach.

TABLE II

THE RATIOS *NAT/NTT* AND *TFA/NTT* FOR SELECTED VALUES OF *N* AND *M*

N	*M*	*NAT/NTT*	*TFA/NTT*
7	3	0.6046512	0.7984496
9	4	0.2875483	0.3591698
12	5	0.0520958	0.0590863
14	7	0.0159710	0.0177106
16	9	0.0060417	0.0065683
18	9	0.0005621	0.0006022
20	11	0.0001931	0.0002049
22	13	0.0000121	0.0000909
24	11	0.0000006	0.0000006
25	10	0.0000001	0.0000001

It is interesting to note the computational efficiency of dynamic programming when redundant arcs are not eliminated. The total number of transitional calculations required in this instance is equal to *TFA*, the total number of feasible arcs as defined in (13). The ratios of *NAT/NTT* and *TFA/NTT* are compared in Table II for selected values of *N* and *M*. These results indicate that, as the magnitude of the problem is increased, the relative gains in efficiency that are due to elimination of redundant arcs are reduced. In absolute terms, however, the difference in *NAT* and *TFA* may be substantial, e.g., when $N=25$ and $M=10$ the values of *NTT*, *TFA*, and *NAT* are as follows:

$$NTT = 12{,}031{,}633{,}921{,}753{,}871{,}000,$$
$$TFA = \quad 1{,}484{,}135{,}596{,}032,$$
$$NAT = \quad 1{,}412{,}092{,}133{,}376.$$

In this case there are $TFA - NAT = 72{,}043{,}462{,}656$ redundant arcs, which is a large number in absolute terms even though it is relatively small in comparison to the tremendous number, NTT, of transitional calculations under total enumeration. *Even when arc redundancy is not eliminated, the number of transitional calculations required under dynamic programming is substantially less than the number of such calculations required under total enumeration in large problems.*

A major portion of the computer memory required under dynamic programming will be for recording optimal policies and $W_K^*(z)$ values for each state in successive stages of the process. The maximum number of states at any stage in our formulation is $NS(1)$, as defined in (12), i.e., the number of states at Stage $K=1$ is greater than the number of states at any other stage. For this reason, values of $NS(1)$ are also listed in Appendix A for various clustering problems. These values provide a guideline for initial-state memory requirements.

Note that, as $K>1$ increases, the number of states at Stage K continually diminishes in our dynamic programming formulation. Thus, even though slow-access auxiliary computer storage may be required in early stages of a large clustering problem, it becomes possible at some point to stay within the bounds of rapid-access storage.

Since the dynamic programming approach requires more computer memory than is needed under a total-enumeration approach, the point at which dynamic programming ceases to be more efficient than total enumeration depends upon the computer at hand. Clearly, in extremely large clustering problems our largest computers may currently be insufficient for practical usage of either dynamic programming or total enumeration, thereby forcing the analyst to satisfice with a more practical linkage approach.

APPENDIX A

Comparison of the Number of Transitional Calculations Required

This appendix compares the number of transitional calculations required when N entities are positioned into M subsets by two methods, total enumeration and dynamic programming. The comparison is shown in Table III, which also shows the maximum number of initial states. In the table,

NTT = number of transitional calculations using total enumeration,
NAT = number of transitional calculations using dynamic programming,
$NS(1)$ = maximum number of initial states.

The data in the table are significant to 16 digits.

TABLE III

COMPARISON OF TRANSITIONAL CALCULATIONS REQUIRED WHEN N ENTITIES ARE PARTITIONED INTO M SUBSETS

N	M	NAT	NTT	NAT/NTT	NS(1)
6	3	266.	270.	0.9851852	50.
7	3	546.	903.	0.6046512	91.
7	4	868.	1050.	0.8266667	91.
8	3	1730.	2898.	0.5969634	210.
8	5	2352.	3150.	0.7466667	154.
9	4	8937.	31080.	0.2875483	420.
9	6	5538.	7938.	0.6976568	246.
10	3	14377.	27990.	0.5136477	837.
10	5	47377.	212625.	0.2228195	837.
10	7	11715.	17640.	0.6641156	375.
11	4	101915.	583000.	0.1748113	1914.
11	6	128139.	897435.	0.1427836	1474.
11	8	22792.	35640.	0.6395062	550.
12	3	164528.	259578.	0.6338288	3784.
12	5	359305.	6897000.	0.0520958	3718.
12	7	317977.	3136980.	0.1013641	2497.
12	9	41470.	66825.	0.6205761	781.
13	4	893997.	10130120.	0.0882514	7722.
13	6	1367756.	55927872.	0.0244557	7007.
13	8	734097.	9498060.	0.0772892	4082.
13	10	71435.	117975.	0.6055096	1079.
14	3	1225849.	2366910.	0.5179111	14898.
14	5	4416673.	200375175.	0.0220420	15808.
14	7	5514861.	345304960.	0.0159710	12896.
14	9	1594047.	25675650.	0.0620840	6461.
14	11	117572.	198198.	0.5932048	1456.
15	4	9404764.	169423800.	0.0555103	32071.
15	6	13647639.	2524159638.	0.0054068	30706.
15	8	14220651.	1516394880.	0.0093779	22803.
15	10	3283723.	63313250.	0.0518647	9933.
15	12	186200.	319410.	0.5829498	1925.
16	3	10256682.	21425058.	0.4787237	58634.
16	5	39143824.	5480952750.	0.0071418	64142.
16	7	41549536.	22973178228.	0.0018086	58514.
16	9	34712368.	5745489750.	0.0060417	39186.
16	11	6461186.	144684540.	0.0446571	14876.
16	13	285328.	496860.	0.5742624	2500.
17	4	80183040.	2777349160.	0.0288703	127704.
17	6	160735520.	105034499388.	0.0015303	127704.
17	8	150774592.	163327960224.	0.0009231	109140.
17	10	80650144.	19308339050.	0.0041770	65518.
17	12	12210420.	3101116 20.	0.0393743	21760.
17	14	424932.	749700.	0.5668027	3196.
18	3	114890112.	193317030.	0.5943093	249509.
18	5	373034240.	144790477725.	0.0025764	260168.
18	7	475470080.	1382237383800.	0.0003440	249356.
18	9	537128192.	955578561795.	0.0005621	199121.
18	11	179232496.	58737034356.	0.0030514	106743.
18	13	22263040.	629273190.	0.0353790	31161.

TABLE III—*Continued*

18	15	617253.	1101600.	0.5603241	4029.
19	4	813709568.	45038667800.	0.0180669	519061.
19	6	1433783552.	4158489610674.	0.0003448	518092.
19	8	1346515968.	13678008027840.	0.0000984	480301.
19	10	1358829056.	4295673304065.	0.0003163	354502.
19	12	382638080.	164268659100.	0.0023293	169746.
19	14	39310832.	1217887650.	0.0322779	43776.
19	16	877116.	1581408.	0.5546424	5016.
20	3	891744512.	1741819338.	0.5119615	988095.
20	5	3969178624.	3746030452500.	0.0010596	1045874.
20	7	5413683200.	78004878319564.	0.0000694	1026665.
20	9	4178830848.	108101543802525.	0.0000387	910385.
20	11	3303106304.	17107581865374.	0.0001931	616645.
20	13	787697408.	427480622660.	0.0018427	263929.
20	15	67436576.	2261646000.	0.0298175	60439.
20	17	1222270.	2223855.	0.5496177	6175.
21	4	6926114816.	726036280200.	0.0095396	2069024.
21	6	12800937984.	159514076776824.	0.0000802	2088043.
21	8	15794270208.	1060088122776672.	0.0000149	2014760.
21	10	14502334464.	711871322912750.	0.0000204	1694990.
21	12	7736217600.	61497378271602.	0.0001258	1048554.
21	14	1568720384.	1045128031500.	0.0015010	401908.
21	16	112697808.	4049722320.	0.0278285	82138.
21	18	1673749.	3070305.	0.5451410	7525.
22	3	7478620160.	15684238350.	0.4768239	3913681.
22	5	30373318656.	95689109560275.	0.0003174	4183401.
22	7	46388747264.	4219336659772078.	0.0000115	4157067.
22	9	44255850496.	111776697318052 75.	0.0000040	3913450.
22	11	48578887680.	4029107509573143.	0.0000121	3096491.
22	13	17502892032.	202471756816327.	0.0000864	1744413.
22	15	3031001856.	2419311602400.	0.0012528	600347.
22	17	183904544.	7020710235.	0.0261946	110033.
22	19	2256254.	4169550.	0.5411265	9086.
23	4	68900225024.	11665370299000.	0.0059064	8343779.
23	6	132073127936.	5993819147900427.	0.0000220	8375657.
23	8	173967933440.	779356401592 03177.	0.0000022	8242832.
23	10	124235808768.	95934012973134553.	0.0000013	7507361.
23	11	133476322816.	53506764398462098.	0.0000026	6690171.
23	12	121977307136.	18393790508323213.	0.0000066	5546358.
23	14	38344118272.	617662582264039.	0.0000621	2842202.
23	16	5695942656.	5336527890848.	0.0010673	880946.
23	18	293639168.	11824426845.	0.0248333	145475.
23	20	2998556.	5578650.	0.5375057	10879.
24	3	81905188864.	141189602418.	0.5801078	16241036.
24	5	308366475264.	2425003917476250.	0.0001272	16761940.
24	7	412071165952.	221742246962631730.	0.0000019	16719436.
24	9	504107368448.	10856031689346 51800.	0.0000005	16240760.
24	11	392918925312.	694101822653530590.	0.0000006	14197785.
24	12	455965933568.	299162455089099040.	0.0000015	12236805.
24	13	296667709440.	757771966371 41740.	0.0000039	9740661.
24	15	81515184128.	1762382180223658.	0.0000463	4540361.
24	17	10433765376.	11276649558398.	0.0009253	1271601.
24	19	459577344.	19403695850.	0.0236850	190026.
24	21	3933920.	7363818.	0.5342229	12926.
25	2	33554304.	33554430.	0.9999962	16777215.
25	4	589187973120.	187085158955240.	0.0031493	33308496.
25	6	1145371099136.	222158502000014510.	0.0000052	33523808.
25	8	1561267798016.	552178976894 6948000.	0.0000003	33306240.
25	10	1412092133376.	120316339217538 71000.	0.0000001	31746288.
25	12	1311826247680.	434715144941849 5200.	0.0000003	26434576.
25	14	700081242112.	285539213969857150.	0.0000025	16777190.
25	15	373814198272.	42993946553471159.	0.0000087	11576890.
25	16	168504524800.	4739896455263929.	0.0000356	7119490.
25	18	18666364928.	22929750164510.	0.0008141	1807755.
25	20	706177792.	31100973753.	0.0227060	245480.
25	22	5100550.	9601350.	0.5312326	15250.

APPENDIX B

A Numerical Illustration

A network illustration is shown in Fig. 3. In order to simplify the network display, a small problem of partitioning $N=5$ entities into $M=3$ clusters is considered. For purposes of applying the dynamic programming algorithm, some redundant arcs shown in Fig. 3 can be eliminated. First of all, transition costs along all arcs

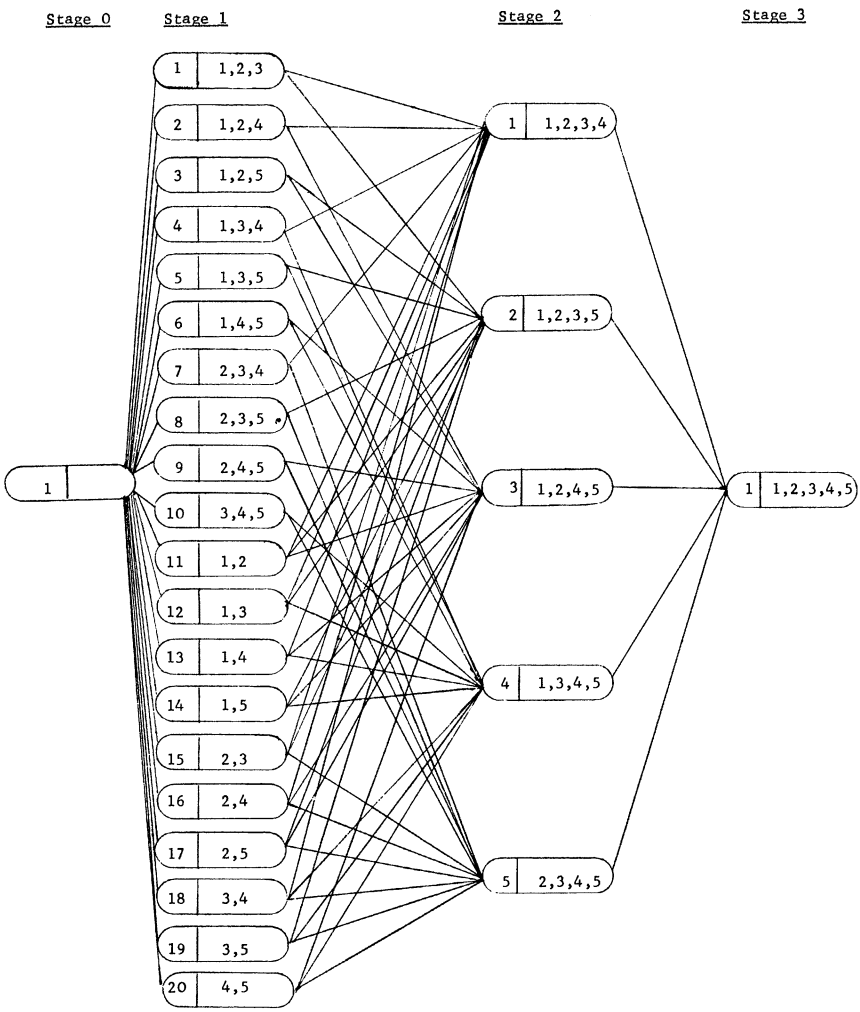

Fig. 3. The network of states when $N=5$ entities are to be partitioned into $M=3$ clusters.

from Stage 2 to Stage 3 are zero, since only one entity can be added to any state after Stage 2. Hence, $M_0 = N - M = 5 - 3 = 2$, as defined in (8), is the last necessary stage of the process.

According to (9), (10), and (11) we derive:

$$\max(1) = 5-3+1 = 3,\ \max(2) = 5-3+2 = 4,$$
$$\min(1) = ([5/3]+1)1 = 2,\ \min(2) = ([5/3]+1)2 = 4.$$

Then, according to (12), we have

$$NS(0) = 1,\ NS(1) = \binom{5}{2} + \binom{5}{3} = 20,\ NS(2) = \binom{5}{4} = 5.$$

Thus, there are 26 states between Stage 0 and Stage 2 inclusively.

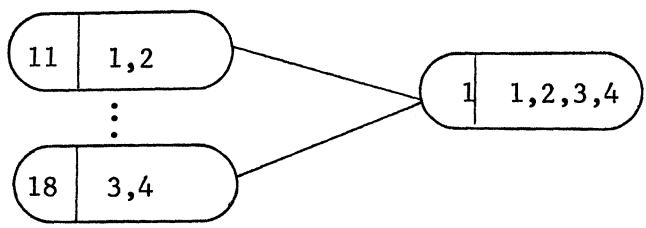

Figure 4

In Fig. 3 it is evident that some of the arcs between Stage 1 and Stage 2 are redundant. For instance, the clusters (1, 2) and (3, 4) of entities 1, 2, 3, and 4 result from each alternative shown in Fig. 4. Hence, one of the two above alternatives can be ignored.

Although $TFA = 70$ arcs are technically feasible between Stage 0 and Stage 2, it is possible to eliminate 15 of these arcs. Using (18) we compute the following for Stage 1:

$$A(2, 2) = \tfrac{1}{2}\binom{5}{2}\binom{3}{2} = 15 \text{ arcs},\ A(3, 1) = \binom{5}{3}\binom{2}{1} = 20 \text{ arcs},\ NA(1) = 35 \text{ arcs}.$$

We then have a total number of arcs $NAT = NS(1) + NA(1) = 20 + 35 = 55$ arcs. It is necessary to make certain that the 15 feasible arcs eliminated between Stage 1 and Stage 2 are redundant. Various elimination alternatives are available, one of which is to eliminate the arcs from between Stages 1 and 2 shown in Fig. 5.

For illustrative purposes, assume that there are $p = 2$ characteristics under study for each of the $N = 5$ entities, where characteristic values are as follows: $x_{11} = 1$, $x_{12} = 3$, $x_{13} = 5$, $x_{14} = 4$, $x_{15} = 1$, $x_{21} = 1$, $x_{22} = 4$, $x_{23} = 5$, $x_{24} = 4$, $x_{25} = 2$. The entities are plotted on the graph in Fig. 6. Using (1) we may compute:

$$d_{12}^2 = 13,\ d_{13}^2 = 32,\ d_{14}^2 = 18,\ d_{15}^2 = 1,\ d_{23}^2 = 5,$$
$$d_{24}^2 = 1,\ d_{25}^2 = 8,\ d_{34}^2 = 2,\ d_{35}^2 = 25,\ d_{45}^2 = 13.$$

From (5) we may then compute the relevant transition costs as shown in Table IV.

Then, accordingly to (8), we may compute for:

Stage 0. $W_0^*(0) = 0$.

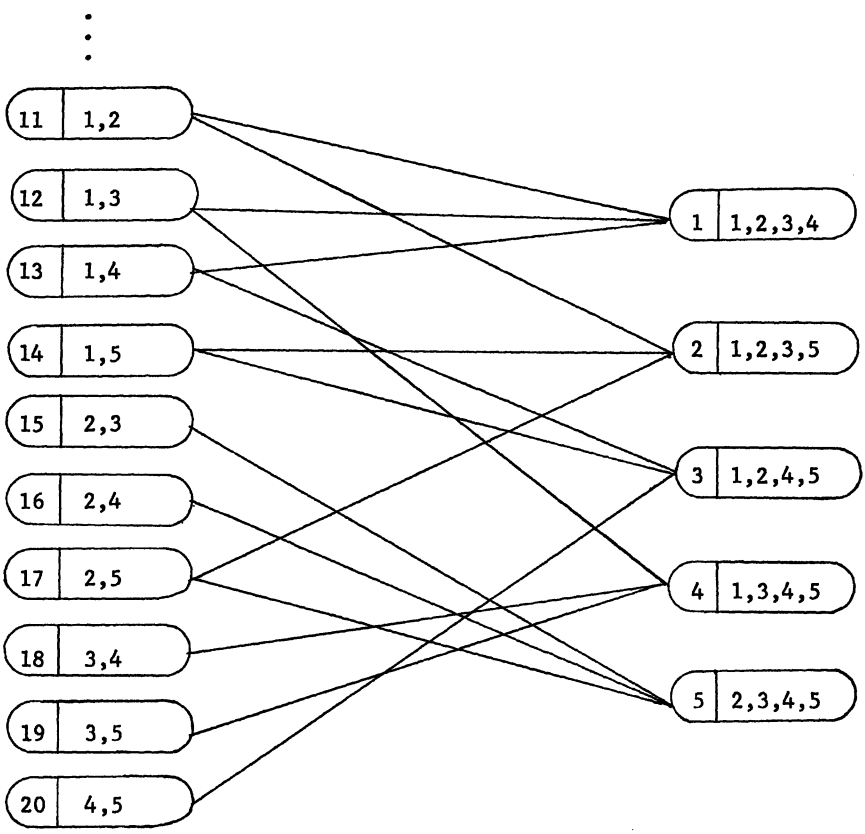

Fig. 5. Arcs to be eliminated between stages 1 and 2.

Stage 1. The values shown in Table V.

Stage 2. The values shown in Table VI.

At Stage $M_0 = 2$ the process in (8) is terminated. At this point the minimum $W_2^*(z)$ value is found to be 1.00 corresponding to the optimal clustering policy of (1, 5) and (2, 4) for entities 1, 2, 4, and 5. The remaining entity, entity 3, is then

assumed to comprise a single-entity cluster such that the optimal clustering solution is (1, 5), (2, 4), and (3) having a distribution form {2} {2} {1}. The minimum W value attainable is $W_2^*(1, 2, 4, 5) = 1.00$.

In a very small problem such as the one illustrated above, dynamic programming is no more efficient than total enumeration of all clustering alternatives. The

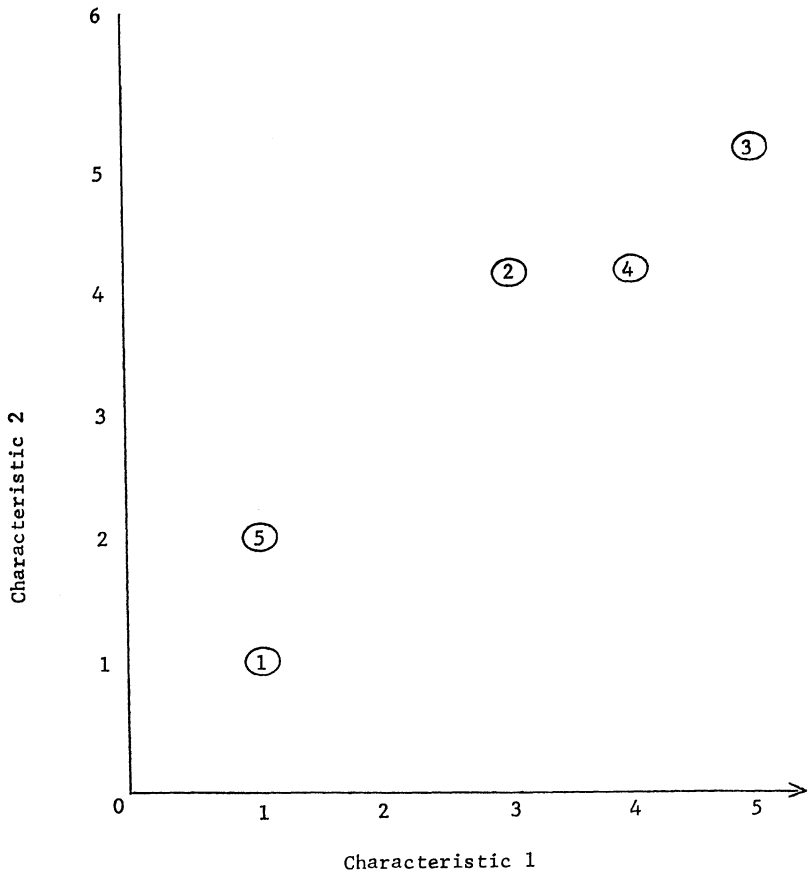

Fig. 6. Graph of $N=5$ entities.

purpose of the illustration was merely to display how networks can be automatically formed in more complicated problems and to illustrate the process in (8).

NOTES

1. Interested readers may obtain a FORTRAN program from the author that will generate all distribution forms for specified values of N and M.
2. A FORTRAN subprogram that will generate all states at each stage is available from the author on request.
3. An alternate approach would be to assign infinite transition costs to arcs between infeasible states.

TABLE IV

TRANSITION COSTS

$T(1) = 0$
$T(2) = 0$
$T(3) = 0$
$T(4) = 0$
$T(5) = 0$

$T(1,2) = \frac{1}{2}(d_{12}^2) = 6.50$
$T(1,3) = \frac{1}{2}(d_{13}^2) = 16.00$
$T(1,4) = \frac{1}{2}(d_{14}^2) = 9.00$
$T(1,5) = \frac{1}{2}(d_{15}^2) = 0.50$
$T(2,3) = \frac{1}{2}(d_{23}^2) = 2.50$
$T(2,4) = \frac{1}{2}(d_{24}^2) = 0.50$
$T(2,5) = \frac{1}{2}(d_{25}^2) = 4.00$
$T(3,4) = \frac{1}{2}(d_{34}^2) = 1.00$
$T(3,5) = \frac{1}{2}(d_{35}^2) = 12.50$
$T(4,5) = \frac{1}{2}(d_{45}^2) = 6.50$

$T(1,2,3) = (1/3)(d_{12}^2+d_{13}^2+d_{23}^2) = 16.67$
$T(1,2,4) = (1/3)(d_{12}^2+d_{14}^2+d_{24}^2) = 10.67$
$T(1,2,5) = (1/3)(d_{12}^2+d_{15}^2+d_{25}^2) = 7.33$
$T(1,3,4) = (1/3)(d_{13}^2+d_{14}^2+d_{34}^2) = 17.33$
$T(1,3,5) = (1/3)(d_{13}^2+d_{15}^2+d_{35}^2) = 13.00$
$T(1,4,5) = (1/3)(d_{14}^2+d_{15}^2+d_{45}^2) = 10.67$
$T(2,3,4) = (1/3)(d_{23}^2+d_{24}^2+d_{34}^2) = 2.67$
$T(2,3,5) = (1/3)(d_{23}^2+d_{25}^2+d_{35}^2) = 32.67$
$T(2,4,5) = (1/3)(d_{24}^2+d_{25}^2+d_{45}^2) = 7.33$
$T(3,4,5) = (1/3)(d_{34}^2+d_{35}^2+d_{45}^2) = 13.33$

TABLE V

VALUES FOR STAGE 1

$W_1^*(1,2,3) = T(1,2,3)+W_0^*(0) = 16.67$
$W_1^*(1,2,4) = T(1,2,4)+W_0^*(0) = 10.67$
$W_1^*(1,2,5) = T(1,2,5)+W_0^*(0) = 7.33$
$W_1^*(1,3,4) = T(1,3,4)+W_0^*(0) = 17.33$
$W_1^*(1,3,5) = T(1,3,5)+W_0^*(0) = 13.00$

$W_1^*(1,4,5) = T(1,4,5)+W_0^*(0) = 10.67$
$W_1^*(2,3,4) = T(2,3,4)+W_0^*(0) = 2.67$
$W_1^*(2,3,5) = T(2,3,5)+W_0^*(0) = 32.67$
$W_1^*(2,4,5) = T(2,4,5)+W_0^*(0) = 7.33$
$W_1^*(3,4,5) = T(3,4,5)+W_0^*(0) = 13.33$

$W_1^*(1,2) = T(1,2)+W_0^*(0) = 6.50$
$W_1^*(1,3) = T(1,3)+W_0^*(0) = 16.00$
$W_1^*(1,4) = T(1,4)+W_0^*(0) = 9.00$
$W_1^*(1,5) = T(1,5)+W_0^*(0) = 0.50$
$W_1^*(2,3) = T(2,3)+W_0^*(0) = 2.50$

$W_1^*(2,4) = T(2,4)+W_0^*(0) = 0.50$
$W_1^*(2,5) = T(2,5)+W_0^*(0) = 4.00$
$W_1^*(3,4) = T(3,4)+W_0^*(0) = 1.00$
$W_1^*(3,5) = T(3,5)+W_0^*(0) = 12.50$
$W_1^*(4,5) = T(4,5)+W_0^*(0) = 6.50$

TABLE VI

VALUES FOR STAGE 2

$$W_2^*(1,2,3,4)=\min\,[T(4)+W_1^*(1,2,3),\ T(3)+W_1^*(1,2,4),\ T(2)+W_1^*(1,3,4),\ T(1)+W_1^*(2,3,4),\ T(2,3)+W_1^*(1,4),\ T(2,4)+W_1^*(1,3),\ T(3,4)+W_1^*(1,2)]=T(1)+W_1^*(2,3,4)=2.67$$

$$W_2^*(1,2,3,5)=\min\,[T(5)+W_1^*(1,2,3),\ T(3)+W_1^*(1,2,5),\ T(2)+W_1^*(1,3,5),\ T(1)+W_1^*(2,3,5),\ T(1,3)+W_1^*(2,5),\ T(2,3)+W_1^*(1,5),\ T(3,5)+W_1^*(1,2)]=T(2,3)+W_1^*(1,5)=3.00$$

$$W_2^*(1,2,4,5)=\min\,[T(5)+W_1^*(1,2,4),\ T(4)+W_1^*(1,2,5),\ T(2)+W_1^*(1,4,5),\ T(1)+W_1^*(2,4,5),\ T(1,2)+W_1^*(4,5),\ T(2,4)+W_1^*(1,5),\ T(2,5)+W_1^*(1,4)]=T(2,4)+W_1^*(1,5)=1.00$$

$$W_2^*(1,3,4,5)=\min\,[T(5)+W_1^*(1,3,4),\ T(4)+W_1^*(1,3,5),\ T(3)+W_1^*(1,4,5),\ T(1)+W_1^*(3,4,5),\ T(1,4)+W_1^*(3,5),\ T(1,5)+W_1^*(3,4),\ T(4,5)+W_1^*(1,3)]=T(1,5)+W_1^*(3,4)=1.50$$

$$W_2^*(2,3,4,5)=\min\,[T(5)+W_1^*(2,3,4),\ T(4)+W_1^*(2,3,5),\ T(3)+W_1^*(2,4,5),\ T(2)+W_1^*(3,4,5),\ T(3,4)+W_1^*(2,5),\ T(3,5)+W_1^*(2,4),\ T(4,5)+W_1^*(2,3)]=T(5)+W_1^*(2,3,4)=2.67$$

4. M is defined initially in (8); it is discussed further in the paragraph following (15).

ACKNOWLEDGMENT

I AM grateful for the helpful suggestions of the referee of this paper that helped to improve both its conceptual formulation and clarity.

BIBLIOGRAPHY

BARTON, D. E. AND F. N. DAVID, 1956, "Spearman's 'Rho' and The Matching Problem," *British Journal of Statistical Psychology* **9,** 69–73.

CLARK, P. J., 1952, "An Extension of the Coefficient of Divergence for Use with Multiple Characters," *Copeia* **2,** 61–64.

COCHRAN, W. G., 1954, "Some Methods for Strengthening the Common Chi Square Tests," *Biometrics* **10,** 417–51.

COLE, L. C., 1957, "The Measurement of Partial Interspecific Association," *Ecology* **38,** 226–33.

———, 1949, "The Measurement of Interspecific Association," *Ecology* **30,** 411–24.

COX, D. R., 1957, "Note on Grouping," *Journal of the American Statistical Association* **52,** 543–47.

DANIELS, H. H., 1950, "Rank Correlation and Population Models," *Journal of the Royal Statistical Society, Series B* **12,** 171–81.

DAVID, S. T., M. G. KENDALL, AND A. STUART, 1951, "Some Questions of Distribution in the Theory of Rank Correlation," *Biometrika* **38,** 131–40.

DICE, L. R., 1945, "Measures of the Amount of Ecological Association Between Species," *Ecology* **26,** 297–302.

DURBIN, J. AND A. STUART, 1951, "Inversions and Rank Correlation Coefficients," *Journal of the Royal Statistical Society, Series B* **13,** 303–09.

EDWARDS, A. W. F. AND L. L. CAVALLI-SFORZA, 1965, "A Method for Cluster Analysis," *Biometrics* **21,** 362–75.

EHRLICH, P. R., 1961, "Has the Biological Species Concept Outlived Its Usefulness?" *Systematic Zoology* **10,** 167–76.

FISHER, R. A., 1936, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics* **7,** 179–88.

FISHER, W., 1958, "On Grouping for Maximum Homogeneity," *Journal of the American Statistical Association* **53,** 789–98.

FORGY, E., 1965, "Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications," *Biometrics* **21,** 758.

FORTIER, J., 1962, "Contributions to Item Selection," *Technical Report Number 2,* Laboratory for Quantitative Research in Education, Stanford University, Stanford, California.

——— AND H. SOLOMON, 1966, "Clustering Procedures," Pp. 493–506, *Multivariate Analysis,* Edited by P. R. KRISHNAIAH, Academic Press, N. Y.

FOSTER, F. G. AND A. STUART, 1954, "Distribution-Free Tests in Time Series Based on the Breaking of Records," *Journal of the Royal Statistical Society, Series B* **16,** 1–22.

FRASER, D. A. S., 1957, *Nonparametric Methods in Statistics,* New York: John Wiley & Sons, Inc.

FRECHET, M., 1953, "Emile Borel, Initiator of the Theory of Psychological Games and Its Application," *Econometrica* **21.**

FRIEDMAN, H. P. AND J. RUBIN, 1957, "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association* **62.**

GREEN, P. E., R. E. FRANK, AND P. J. ROBINSON, 1967, "Cluster Analysis in Test Market Selection," *Management Science* **13,** 13–387–400.

HOLZINGER, K. J. AND H. H. MARMON, 1941, *Factor Analysis,* Chicago Press, Chicago, Illinois.

JACCARD, P., 1908, "Nouvelles Recherches sur la distribution florale," *Bull. Soc. Vand. Sci. Nat.* **44,** 223–70.

JENSEN, R. E., 1968a, "Clustering Algorithms for the Determination of Compact Entity Configurations—Part I," Working Paper No. 18. Department of Accounting and Finance, Michigan State University, East Lansing, Michigan.

———, 1968b, "Clustering Algorithms for the Determination of Compact Entity Configurations—Part II," Working Paper No. 19, Department of Accounting and Finance, Michigan State University, East Lansing, Michigan.

KAHL, J. A. DAVIS, 1955, "A Comparison of Indexes of Socio-Economic Status," *American Sociological Review* **20,** 317–25.

KENDALL, M. G., 1966, "Discrimination and Classification," *Multivariate Analysis*, Edited by P. R. KRISHNAIAH, Academic Press, New York, 1966, pp. 165–84.

———, 1938, "A New Measure of Rank Correlation," *Biometrika* **30,** 81–93.

———, S. F. H. KENDALL, AND B. B. SMITH, 1938, "The Distribution of Spearman's Coefficient of Rank Correlation in a Universe in Which All Rankings Occur an Equal Number of Times," *Biometrika* **30,** 251–73.

KING, B., 1967, "Stepwise Clustering Procedures," *Journal of the American Statistical Association* **62,** 86–101.

KULCZNSKI, S., 1927, "Die Pflanzenassoziationen der Pieninen [In Polish, German Summary], *Bull. Intern. Acd. Pol. Sci. Lett. cl. Sci. Math. Nat.* **13** (Sci. Nat.), 1927 (Supl. 2), pp. 57–2–3.

MAHALANOBIS, P. C., 1936, "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Science in India* **2,** 49–55.

MCQUITTY, L. L., 1966a, "Single and Multiple Hierarchical Classification by Reciprocal Pairs and Rank Order Types," *Educational and Psychological Measurement* **26,** 253–65.

———, 1966b, "Improved Hierarchical Syndrome Analysis of Discrete and Continuous Data," *Educational and Psychological Measurement* **26,** 577–82.

———, 1964, "Capabilities and Improvements of Linkage Analysis as a Clustering Method," *Educational and Psychological Measurement* **24,** 441–56.

———, 1963, "Rank Order Typal Analysis," *Educational and Psychological Measurement* **23,** 55–61.

———, 1962, "Multiple Hierarchical Classification of Institutions and Persons with Reference to Union—Measurement Relations and Psychological Well-Being," *Educational and Psychological Measurement* **22,** 513–31.

———, 1961, "Typal Analysis," *Educational and Psychological Measurement* **21,** 677–96.

———, 1960, "Hierarchical Syndrome Analysis," *Educational and Psychological Measurement* **20,** 293–304.

———, 1957, "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevances," *Educational and Psychological Measurement* **17,** 207–29.

———, 1956, "Agreement Analysis: Classifying Persons by Predominant Patterns of Response," *British Journal of Statistical Psychology* **9,** 5–16.

MICHENER, C. D. AND R. R. SOKAL, 1957, "A Quantification of Systematic Relationships and Phylogenetic Trends," *Proceedings of the Xth International Congress of Entomogy* **1,** 409–15.

MORISHIMA, H. AND H. OKA, 1960, "The Pattern of Interspecific Variations in the Genus *Oryza*: Its Quantitative Representation by Statistical Methods," *Evolution* **14,** 153–65.

MORRISON, D. G., 1967, "Measurement Problems in Cluster Analysis." *Management Science* **13,** 13–775–80.

OLDS, E. J., 1949, "The 5% Significance Levels for Sums of Squares of Rank Differences and a Correction," *Annals of Mathematical Statistics* **20.**

RAO, C. R., 1952, *Advanced Statistical Methods in Biometric Research*, Wiley, New York.

———, 1948, "The Utilization of Multiple Measurements in Problems of Biological Classification," *Journal of the Royal Statistical Society, Series 13* **10,** 159–93.

ROGERS, D. J. AND T. T. TANIMOTO, 1960, "A Computer Program for Classifying Plants," *Science* **132,** 115–18.

ROZEBOOM, W. W., 1965, "Linear Correlations Between Sets of Variables," *Psychometrika* **30.**

RUBIN, J., 1967, "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem," *Journal of Theoretical Biology* **15,** 103–44.

RUSSELL, P. F. AND T. R. RAO, 1940, "On Habitat and Association of Species of Anopheline Larvae in South-eastern Madras," *J. Malor Inst. India* **3,** 153–78.

SIEGAL, S., 1956, *Nonparametric Statistics for the Behavioral Sciences,* New York: McGraw-Hill.

SMIRNOV, E. S., 1960, "Taxonomic Analysis of a Genus," [In Russian with an English Summary], *Zhurnal Obshenye Biologii* **21,** 89–103.

SNEATH, P. H. A., 1957, "The Application of Computers to Taxonomy," *Journal of General Microbiology* **17,** 201–226.

SOKAL, R. R. AND P. H. A. SNEATH, 1963, *Principles of Numerical Taxonomy,* San Francisco, W. H. Freeman and Company.

———, 1961, "Distances as a Measure of Taxonomic Similarity," *Systematic Zoology* **10,** 70–79.

——— AND C. D. MICHENER, 1958, "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin* **38,** 1904–38.

SORENSEN, T., 1948, "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons," *Biol. Skr.* **5,** 1–34.

SORIA, V. J. AND C. B. HEISER, JR., 1961, "A Statistical Study of Relationships of Certain Species of the *Solanum Nigrum Comples,*" *Economic Botomy* **15,** 245–55.

SPEARMAN, C., 1913, "Correlations of Sums and Differences," *British Journal of Psychology* **5,** 417–26.

TRYON, R. C., 1939, *Cluster Analysis,* Edwards Brothers, Ann Arbor, Michigan.

WALSH, J. E., 1962, *Handbook of Nonparametric Statistics,* Princeton, New Jersey, D. Van Nostrand.

WARD, J. H., 1963, "Hierarchical Grouping to Optimize an Objective Function," *Journal of American Statistical Association* **58,** 236–44.

WILKS, S., 1960, "Multidimensional Statistical Scatter," pp. 486–503, *Contributions to Probability and Statistics,* Edited by I. OLKIN, Stanford University Press, Stanford, California.

———, 1935, "On the Independence of k Sets of Normally Distributed Statistical Variables," *Econometrica* **3,** 309–26.

YULE, G. U. AND M. G. KENDALL, 1950, *An Introduction to the Theory of Statistics,* Hafner, New York.