Minimum Spanning Trees and Single Linkage Cluster Analysis

By J. C. Gower and G. J. S. Ross

Rothamsted Experimental Station

Summary

Minimum spanning trees (MST) and single linkage cluster analysis (SLCA) are explained and it is shown that all the information required for the SLCA of a set of points is contained in their MST. Known algorithms for finding the MST are discussed. They are efficient even when there are very many points; this makes a SLCA practicable when other methods of cluster analysis are not. The relevant computing procedures are published in the Algorithm section of the same issue of *Applied Statistics*.

The use of the MST in the interpretation of vector diagrams arising in multivariate analysis is illustrated by an example.

1. Background

In the analysis of multivariate data, the sampling units are commonly represented as points in a multidimensional space where the distances between pairs of points are defined as some function of the observed sample values. Distance can be defined in many ways, not all of which obey the rules of Euclidean geometry. If there are v quantitative variates and x_{ij} is the observed value of the jth variate for the ith sampling unit, the following are two common ways of defining the distance d_{pq} between the pth and qth units:

$$d_{pq}^2 = \sum_{k=1}^{v} (x_{pk} - x_{qk})^2,$$

$$d_{pq} = \sum_{k=1}^{v} |x_{pk} - x_{qk}|.$$

When x_{ij} is the mean of the jth variate in *population i* and the covariance matrix within all populations is known, Mahalanobis's generalized distance D can be used (Mahalanobis, 1936).

When some, or all, the variates are qualitative, coefficients measuring the similarity between two sampling units can be defined (see Sokal and Sneath, 1963). For example, when all variates are of presence/absence type, one of the simplest similarity measures, the simple matching coefficient, may be used. The similarity S_{pq} between individuals is then expressed as the ratio of matches to the number of variates compared; a match occurs when the variate is either present or absent in both individuals. In another similarity coefficient, "matches" between absent characters are ignored in both numerator and denominator. Similarities are not distances, because they take the value 1 when both individuals are identical, but they can be simply transformed into distances by formulae such as

$$d_{pq}^2 = 1 - S_{pq}.$$

Thus a doubly infinite set of distance definitions arises by using different measures of similarity and different ways of transforming these similarities into distances.

Distance in multivariate work may also be determined from direct experimental observation. For example, each member of a panel is asked to assess a pair of odours by scoring on a scale 0, if the odours are considered identical, to 5 if they have nothing in common. Distance could then be defined as the average score (over all members of the panel) for each pair of odours.

Although so far it is the sampling units which have been represented by points in multidimensional space, this is not the only such representation. In many applications the sampling variates may be represented by points. This is commonplace in factor analysis but a more simple example occurs when r_{jk} is the correlation between the jth and kth variates and the variates are represented by points whose distance apart d_{jk} is defined by:

$$d_{jk}^2 = 1-r_{jk}^2.$$

Multidimensional configurations are difficult to interpret because they cannot be easily visualized. Two broad classes of analysis have been used to give approximate representations of many dimensions in a few; these are vector analysis and cluster analysis.

Vector analysis methods are based on the calculation of latent roots and vectors and include principal components, canonical variate analysis and factor analysis. This type of analysis is familiar to statisticians, but it is not always realized that, even when a two-dimensional (say) representation gives a good fit to the multidimensional one, there may still be some quite serious distortion in the true relative distances as seen in two dimensions. In Section 5 we show, with an example, that this can happen and suggest one convenient way in which this distortion can be checked.

Cluster analysis is less familiar to statisticians, but is becoming more popular. It attempts to group the points in multidimensional space into (usually) disjoint sets which, it is hoped, will correspond to marked features of the sample. The grouped sets of points may themselves be grouped into larger sets, so that all the points are eventually hierarchically classified. This hierarchical classification can be represented diagrammatically (as in Fig. 2), when it is termed a *dendrogram*. It is usual to incorporate a scale into the dendrogram to indicate the similarity or distance level at which the various groups are supposed to join. There are many forms of cluster analysis, but we are only concerned here with one of the simplest (Single Linkage Cluster Analysis (SLCA) discussed in Section 3).

Some aspects of Operational Research are concerned with configurations of sets of points (usually in two-dimensional space). The points may represent towns joined by roads with given distances, or unjoined (conventionally represented by very long distances). Typical O.R. problems are to find the shortest distance between two towns or the shortest length of road required to join all the towns, or the shortest circuit including all towns. Methods based on mathematical programming and graph theory have been developed and used for solving such problems. We show below that some very simple parts of graph theory, concerned with so-called *trees*, are useful as a basis for SLCA (Section 4) and also for interpreting vector diagrams (Section 5). Computational aspects are set out in the Appendix and the relevant procedures for the algorithms in the algorithm section of this journal (Ross, 1969).

Although a distinction is usually made between O.R. and statistics, we believe that this distinction is artificial, or at least unprofitable. It is often argued that for a method to be statistical it should have some probabilistic basis, but many methods profitably used by practising statisticians do not have this basis. In others (for example, analysis of variance) it is arguable that the probabilistic features are not fundamental to the method. The methods discussed here are not probabilistic, although they could be cast in a probabilistic framework by considering different distributions of the points and of the distance statistics.

2. Minimum Spanning Trees

Suppose n points are given (possibly in many dimensions), then a tree spanning these points (or vertices) is any set of straight line segments joining pairs of points such that:

(i) no closed loops occur;
(ii) each point is visited by at least one line;
(iii) the tree is connected.

Fig. 1 is a simple example of a tree with integer segment lengths; thus segment BE is 5 units long. If, for example, vertices C and G were joined, a closed loop would be formed and the resulting figure would not be a tree. The length of a tree is the sum of the lengths of its segments so the tree in Fig. 1 has length $4+2+5+1+6+7+3 = 28$ units.

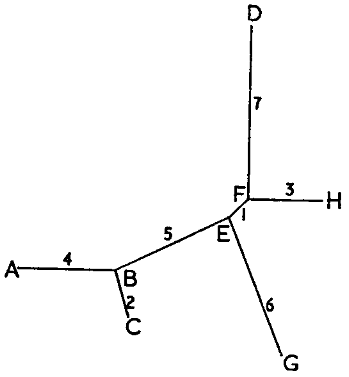

Fig. 1. A tree with eight vertices and the lengths of its segments.

When a set of n points and the lengths of all $\binom{n}{2}$ segments are given, the spanning tree of minimum length (the minimum spanning tree, MST) is often required. Algorithms to find the MST are frequently rediscovered because applications are very diverse; examples known to us are:

(*a*) Loberman and Weinberger (1957) required the shortest length of wire to connect a set of terminals.
(*b*) Prim (1957) was interested in finding the minimum length of telephone line to interconnect 49 cities.
(*c*) Kruskal (1956) drew attention to a mathematical discussion of the MST problem given by Borůvka (1926) (in Czech) which we have not read. Kruskal offered alternative algorithms and suggested that the MST problem was related to the travelling salesman problem.

(*d*) Obruča (1968) took up Kruskal's suggestion and put forward a method of iterating on the MST solution to find an approximate solution to the travelling salesman problem.

(*e*) Stillinger (1967) needed the MST to describe the physical configuration of sets of molecules as related to critical phenomena such as the change of state from liquid to gas.

(*f*) Florek *et al.* (1951) give a mathematical discussion of the MST and relate this to the taxonomic problem of describing the interrelationships of species. This type of application in taxonomy seems to have had a long history in Poland, originating with Czekanowski (1909).

(*g*) Edwards and Cavalli-Sforza (1964) were concerned with reconstructing an evolutionary tree given the gene frequencies derived from the different blood-groups for fifteen present-day human populations. They associated a likelihood with every possible evolutionary tree and, after defining a suitable metric, used a MST to give a first approximation to the tree with maximum likelihood.

(*h*) The MST problem, with many related problems, is familiar in O.R. A comprehensive review of the subject is, for example, given by Berge and Ghouila-Houri (1965).

The two most popular algorithms for finding the MST both operate iteratively; at any stage the segments belong to one of two sets—set A containing those segments assigned to the MST and set B, those not assigned.

The most popular algorithm is to assign iteratively to A the shortest segment in B which does not form a closed loop with any of the segments already in A. Initially A is empty and iteration stops when A contains $(n-1)$ segments. We call this algorithm I; it is given, among other possible algorithms, in the references under points (*a*), (*b*), (*c*), (*e*), (*f*) and (*h*) above.

It can now be seen that the tree in Fig. 1 is the MST. The joins specified by the above algorithm occur in the sequence:

(EF) (BC) (FH) (AB) (BE) (EG) (DF).

Algorithm II, which we shall need later, is given only by Prim (1957). We start with any one of the given points and initially assign to A the shortest segment starting from this point. The procedure is then to continue to add to A the shortest segment from B which connects to at least one segment from A without forming a closed loop amongst the segments already in A. As in algorithm I, iteration stops when there are $(n-1)$ segments in A.

The order in which the segments of the MST of Fig. 1 are found by Prim's algorithm if we start from A is:

(AB) (CB) (EB) (FE) (HF) (GE) (DF).

Starting from E, the order is:

(EF) (FH) (BE) (CB) (AB) (EG) (DF).

In both algorithms, when a choice of several equal segments of minimum length occurs, any one may be selected. In such cases there need not be a unique MST, but otherwise algorithms I and II must give identical results.

3. Single Linkage Cluster Analysis (SLCA)

This method was put forward by Sneath (1957) as a convenient way of summarizing taxonomic relationships in the form of dendrograms (taxonomic trees). The relationships between n samples are supposed to be expressed in terms of the taxonomic distances (measured on some acceptable scale) between every pair of samples. The method consists of a sorting scheme that determines clusters at a series of increasing distance thresholds ($d_1, d_2, \ldots$). The clusters at level d_i are constructed as follows:

Group the samples into disjoint sets by joining all segments of length d_i or less. Each set is said to form a *cluster* at level d_i. Thus all segments joining two clusters defined at level d_i will have lengths greater than d_i. Clearly many of the links of length $\leqslant d_i$ will be redundant; all that is required is that a chain of segments of length $\leqslant d_i$ joins all the members of a cluster.

If sorting is done at a greater distance threshold d_{i+1}, all clusters at level d_i remain but some may combine into larger clusters. Two clusters will combine when at least one link exists between them of length d where $d_i < d \leqslant d_{i+1}$. This property of requiring only one link for combination of groups explains the name "Single Linkage Cluster Analysis".

The dendrogram shows how clusters at level d_1 combine at level d_2 and so on at successive levels until all samples combine into a single cluster at the root of the tree.

The form of algorithm I shows that the clusters at any level d_i can be obtained from the MST by deleting all segments of length greater than d_i and therefore that the dendrogram can also be derived from the MST.

Fig. 2(a) shows the dendrogram derived from Fig. 1. Thus E and F are clustered at distance level 1 by SLCA and therefore occur on branches of the dendrogram which join at level 1. Similarly H clusters with E and F at level 3, etc. This Figure is also a tree but its branch-points do not correspond to any of the points given originally, and it has no metric properties except that, when two points are linked at level d (say) in the dendrogram, no segment joining the same two points in the MST can be longer than d units. For example, A and C are linked at level 4 in Fig. 2(a) and A is joined to C in Fig. 1 via B where AB is 4 and BC 2 units. This relationship between the MST construction and SLCA shows that SLCA is very similar to the taxonomic method of Florek *et al.* (1951), that is, application (f) mentioned above.

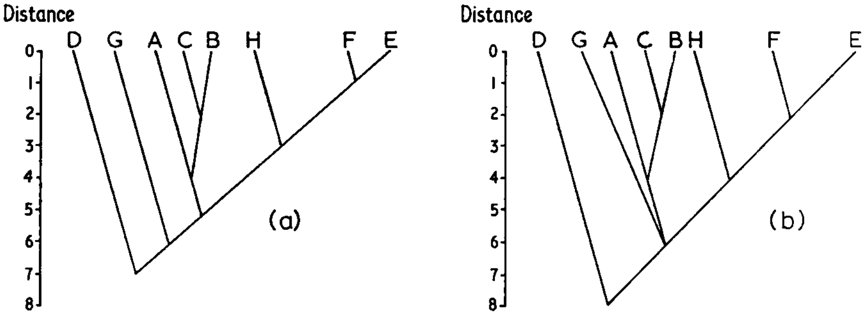

Fig. 2. Dendrograms derived from Fig. 1: (a) using continuous increments, (b) using increments 0 (2) 8.

In practice the threshold levels in a SLCA are not increased continuously but a constant increment (or decrement) δ is made. Because several links may join between two threshold levels L and $L+\delta$ some detail on the exact distances when clusters combine can be lost, so that a dendrogram obtained in this way may not be exactly the same as one derived directly from the MST. Fig. 2(b) shows how the dendrogram of Fig. 2(a) is distorted when links are considered only at 0, 2, 4, 6 and 8 units.

4. Using SLCA

Gower (1967) interpreted several methods of cluster analysis in terms of distance. These methods generally define clusters by maximizing some simple function of average interset distance and so tend to give fairly compact, roughly spherical clusters; SLCA can produce long clusters such that individuals belonging to two different clusters may be closer together than different members of the same cluster. For this reason SLCA is often disliked, but evidence of a continuous sequence of intermediate samples can be informative.

Unlike most other cluster analysis methods, SLCA obtains exactly the same results by agglomerating small clusters into larger ones (sorting upwards by increasing the threshold distances) as by dividing larger clusters into smaller ones (sorting downwards by decreasing the threshold distances). Algorithm I shows that monotonic transformations of the distance function all give the same MST. This is a very desirable property when the distances are defined rather arbitrarily, so that relative magnitudes are more meaningful than absolute values.

Algorithm II, described above, allows a SLCA of much larger samples than is possible for other types of cluster analysis (see the Appendix for some computational details). The method is therefore especially convenient for a preliminary analysis of very large multivariate samples. It may indicate that these can be reasonably split into smaller sub-samples which can then be analysed separately using more refined methods. Additional information on clusters suggested by SLCA can easily be obtained in supplementary analyses; for example, it is usual to evaluate quantities such as inter- and intra-cluster average distances to indicate the density of linkage. A table of the k nearest neighbours of each point ($k = 5$ is a convenient number) reveals possible alternative branches in event of ties, and allows compact clusters to be distinguished from long clusters; this table also helps in planning the display of the MST.

5. Interpretive Uses of the MST

The MST itself also helps in the interpretation of other cluster analysis methods. For example, close neighbours assigned to different clusters will be revealed, and the adequacy of the clusters can be judged.

The MST has also been found useful when used with vector diagrams that illustrate approximations in a few dimensions, to configurations in many dimensions.

The distances in Table 1 were derived from ten measurements on various skull characteristics of samples of white-toothed shrews from the Scilly and Channel Islands (Delany and Healy, 1966). Fig. 3 shows a plot of the first two canonical variate means; the distances in this diagram do not reproduce those of Table 1 exactly; to do this nine dimensions would be required. The two-dimensional approximation accounts for 89 per cent of the variance, but that there is still some distortion is readily apparent from examining the MST given in Fig. 3. Thus the Jersey and Sark races appear to be well separated from the other Channel Island and also from the Scilly Island races. The MST shows that the Jersey and Sark races are closer to the

TABLE 1

Generalized distances giving the distances between the canonical variate means of ten island races of white-toothed shrews (from Delany and Healy, 1966)

	Scilly Islands					*Channel Islands*			
	Tresco	*Bryher*	*St Agnes*	*St Martin's*	*St Mary's*	*Sark*	*Jersey*	*Alderney*	*Guernsey*
Bryher	1·88								
St Agnes	2·33	2·54							
St Martin's	2·26	2·97	3·22						
St Mary's	1·74	2·05	1·54	2·68					
Sark	2·93	4·00	4·01	4·51	3·84				
Jersey	3·30	4·52	4·10	3·46	3·56	3·37			
Alderney	10·73	10·89	11·28	10·01	10·54	10·99	10·44		
Guernsey	8·83	9·09	9·66	8·20	9·04	9·00	8·96	3·27	
Cap Gris Nez†	8·57	8·78	9·21	8·24	8·64	8·74	9·07	3·77‡	3·00

† A French mainland race.

‡ This figure was incorrectly printed as 2·77 in the original publication.

Tresco race than to each other (otherwise the join JS would be a link in a shorter tree), so that the separation from those of the Scilly Islands is illusory. The tree also shows that the Bryher race is closer to that of Tresco than to those of St Mary's and St Agnes, so that the Scilly Island cluster is more compact than it appears in the diagram.

It is remarkable, and immediately obvious from a glance at the MST, that the Tresco race is the closest race to each of the races in the Scilly Islands, Jersey, Sark group, with the exception of St Agnes. An examination of the distances in Table 1 shows that Tresco is second closest to the St Agnes race.

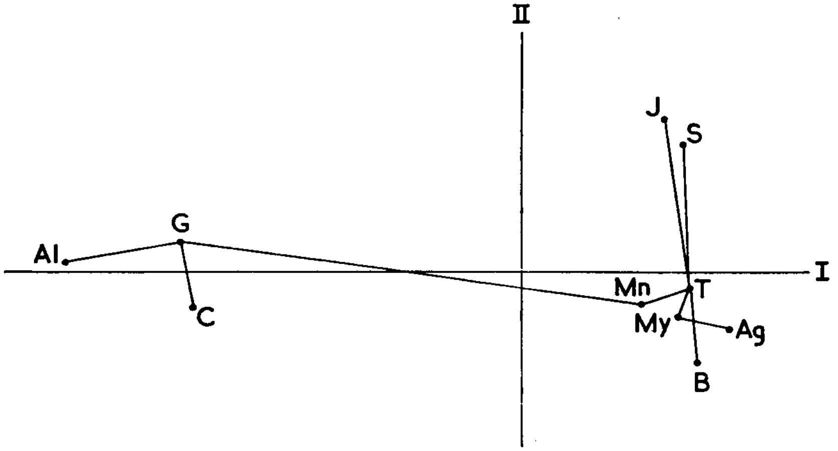

FIG. 3. The means of the first two canonical variates for ten island races of white-toothed shrews, with superimposed MST. Ag, St Agnes; Al, Alderney; B, Bryher; C, Cap Griz Nez; G, Guernsey; J, Jersey; Mn, St Martin's; My, St Mary's; S, Sark; T, Tresco. (Adapted from Delany and Healy, 1966.)

These findings might be derived directly from examining the distances in Table 1, and studying the means of the third canonical variate. When there are very many points (n, say) in the vector diagram, the distance matrix becomes too large to be assimilated and it is impractical to record all $\binom{n}{2}$ distances. Information from the higher-order axes cannot always be incorporated conveniently, and the MST is then a useful ancillary technique.

Acknowledgement

Table 1 and Fig. 3 were derived from Delany and Healy (1966), whom we thank for permission to use their data.

References

Berge, C. and Ghouila-Houri, A. (1965). *Programming, Games and Transportation Networks.* London: Methuen and Co. Ltd.

Borůvka, O. (1926). On a minimal problem, Prace Moravoke. *Prdovedecky Spolecnosti*, **3**.

Czekanowski, J. (1909). Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenzblatt der Deutschen Gesellschaft für Anthropologie*, XL, S, 44–47.

Delany, M. J. and Healy, M. J. R. (1966). Variation in the white-toothed shrews (*Crocidura* spp.) in the British Isles. *Proc. Roy. Soc.* B, **164**, 63–74.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification*, pp. 67–76. London: The Systematics Association, Publication No. 6.

Florek, K., Lukaszewicz, J., Perkal, H., Steinhaus, H. and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, **2**, 282–285.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, **23**, 623–637.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and a travelling salesman problem. *Proc. Amer. Math. Soc.*, **7**, 48–50.

Loberman, H. and Weinberger, A. (1957). Formal procedures for connecting terminals with a minimum total wire length. *J. Assoc. Comp. Mach.*, **4**, 428–237.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, **12**, 49–55.

Obruča, A. K. (1964). Algorithm mintree. *Comp. Bull.*, **8**, 67. (See also comments, corrections and modifications in *Comp. Bull.*, **8**, 109; and in *Comp. Bull.*, **9**, 18.)

—— (1968). Spanning tree manipulation and the travelling salesman problem. *Comp. J.*, **10**, 374–377.

Prim, R. C. (1957). Shortest connection matrix network and some generalizations. *Bell System Tech. J.*, **36**, 1389–1401.

Ross, G. J. S. (1969). Algorithms AS 13–15. *Appl. Statist.*, **18**, 103–110.

Shepherd, M. J. and Willmott, A. J. (1968). Cluster analysis on the Atlas computer. *Comp. J.*, **11**, 57–62.

Sneath, P. H. (1957). Computers in taxonomy. *J. gen. Microbiol.*, **17**, 201–226.

Sokal, P. R. and Sneath, P. H. (1963). *Principles of Numerical Taxonomy.* San Francisco and London: Freeman.

Stillinger, F. H. (1967). Physical clusters, surface tension, and critical phenomena. *J. Chem. Phys.*, **47**, 2513–2533.

APPENDIX

Computation

The computation of the MST

Obruča (1964) published an Algol procedure for algorithm I but it is more convenient to use Prim's (1957) algorithm II to find the MST and hence compute a SLCA. Prim's algorithm is faster and requires each interpoint distance once only, so that the distance matrix need not be stored (but must still be computed). Hence the limiting factor is time rather than space.

The MST is computed as follows:

Three lists are formed to contain, for each point P:

List 1. An indicator which is 1 if P belongs to group A, and 0 otherwise.

List 2. For members of group A, the reference number of the point to which P was linked when it joined group A. For members of group B, the reference number of the point in group A nearest to P.

List 3. For all points the distance between P and the point referred to in list 2.

Initially point 1 is assigned to group A. Let Q be the latest addition to A. Then the distance PQ is calculated for each member of B, and if it is less than the value recorded in list 3, Q and the new distance are substituted for the values in lists 2 and 3. Simultaneously the minimum value of the distances recorded in list 3 for members of B is found, and the next point Q is determined. The new point Q is then assigned to A. The process terminates when all points belong to group A.

The order of printing the links should aim at helping users construct the MST; building the tree from several hundred links in random order can be very time-consuming. A convenient standard printing order is given by selecting a current vertex and printing any set of links joining this vertex to a terminal vertex. Steps are then retraced until a vertex is found from which unprinted links stem; this is the new current vertex. Initially the current vertex can be taken as point 1. All the information for this process is in list 2.

The computation of the SLCA

SLCA may be computed from the MST list as follows. A distance threshold δ is provided, and if the minimum distance $d_{\min}$ is known, sorting may begin at L_0 the largest multiple of δ which is less than $d_{\min}$. A list H is formed of all links whose lengths lie between L and $L+\delta$.

Define a list G containing the group members, marking the final member of each group with an indicator (the members of each group are assumed contiguous). G consists initially of all points as single groups.

Take each link in turn from list H and find the end-points in list G. Amalgamate the two groups in which the points are found, and shift down the intervening groups where necessary. When list H is exhausted print the current grouping. Continue until all links have been used.

The superiority of the MST method based on algorithm II over other published methods of computing the SLCA can be summarized as follows:

(i) distances are required once only; repeated scanning of the distance matrix is not required, therefore it need not be stored;

(ii) a hierarchical system of agglomerative clusters is easily constructed;

(iii) the MST reveals the relationships between the clusters, the links which join clusters and the internal structure of each cluster.

Thus Shepherd and Willmott (1968) describe algorithms for SLCA which require the full distance matrix at every level of clustering and which fail to show even the hierarchical nature of clustering at successive stages. They then have to use ill-defined concepts of multiple linkage to provide the information on structure which is very simply derived from the MST. Their method of output also requires considerable extra work when interpreting the results.

A method for printing the dendrogram

If the output characters include "underline" and "vertical bar" a dendrogram may be printed directly. The information required is obtained during sorting, and is packed into a list each element of which indicates two print characters according to the following code:

Code 0 = Space, space	Code 1 = Underline, space
Code 2 = Underline, vertical bar	Code 3 = Underline, underline
Code 4 = Space, vertical bar	Code 5 = Underline, vertical bar

At each increase of L the list G is processed as follows: Let s = previous print indicator, t = next print indicator, u and v be switches. Initially $t = 3$ for all points. Shift up the print word 3 places before adding t.

Table 2 is a decision table for selecting the value of t, and changing u and v as necessary. A — symbol indicates that the test or setting is to be ignored. The tree is printed in the order defined by G at the final stage.

TABLE 2

Decision table for selecting print indicators (t) for direct printing of dendrogram

Is point first of a group? ($u = 0$?)	*Is point last of a group?* (*indicated in G*)	$s = 2$ or 3?	$s = 3$?	$v = 0$?	*Then set*		
					t	u	v
Yes	Yes	—	—	—	3	0	—
	No	—	Yes	—	1	1	1
			No	—	0	1	—
No	Yes	—	—	Yes	3	0	—
				No	2	0	0
	No	No	—	Yes	0	1	—
				No	4	—	—
		Yes	—	Yes	1	1	1
				No	5	1	—

In the Algorithm section of this journal Ross (1969) gives algorithms, in Algol, for the following three purposes:

(i) Algorithm AS 13. Prim's Algorithm II, to find the MST.
(ii) Algorithm AS 14. To print the MST.
(iii) Algorithm AS 15. To compute a SLCA from the MST and print the dendrogram.

These algorithms, being in a machine-independent language, do not take advantage of the efficient use of storage that can be gained by packing the lists.

Limitations

Cluster analysis programs should be able to handle many objects. If the size of the collection is too small, the clusters observed may have little meaning, except as accidental aggregations of random observations.

In any method of SLCA either the data matrix or the distance matrix must be held in the machine store because random access is required to distances which must be either looked up or evaluated. If N is the number of objects and T is the number of test characters, the storage required for the data matrix is proportional to NT; for the distance matrix it is proportional to N^2. Whether packing is used or not, if N is sufficiently large the distance matrix will require more store than the data matrix, and thus for the largest jobs the data matrix is best used (although the elements of the distance matrix in random order may be stored on magnetic tape for use by other routines). For example, the original program (not based on MST's) used on the Orion computer at Rothamsted could handle 400 objects, storing the whole distance matrix, whereas the same store can hold a data matrix of 625 objects with 256 tests, or 1,250 objects with 128 tests.

Computing time is an important limitation on the size of a cluster analysis. Because MST requires computation of the distance matrix, whether it is stored or not, time will depend on N^2. This compares favourably with other types of cluster analysis, having no sounder theoretical basis, but whose times depend on N^3, or even 2^N (see some of the methods discussed by Gower, 1967).