Cluster Analysis of Simulated Vegetation Data

- H. van Groenewoud -

SUMMARY

Two sets of simulated, randomly distributed, vegetation data of different complexity (3 and 50 species) were analysed with both a "single linkage" and a "sum-of-squares" clustering method.
The 3-species data showed a strong clustering with both methods. Only the "sum-of-squares" method showed strong clustering in the data with 50 species. The "single linkage" clustering resulted in "chaining", pointing to a more continuously variable vegetation.
It is emphasized that many clustering methods result in strong clustering even when the data are random.

INTRODUCTION

Cluster analysis has been used as a multivariate numerical method for classifying a variety of objects from bacteria and plant communities to archaeological artifacts. The following remarks refer mainly to the classification of plant communities but may be equally valid for other objects.
Consider a multidimensional test-space in which the mutually orthogonal axes represent the species (with as many axes as there are species). Within this test-space each individual (sample plot, relevé, stand) is represented by one point. The species quantities form the coordinates of the locations of the points in this test-space.

There is no single definition of a cluster. A fairly general definition of cluster analysis and the one used in this paper is: The detection and identification of groups of individuals (sample plots, etc.) that resemble each other more than members of other groups; indicating that natural discontinuities occur in the distribution of the parent population.
The often unstated assumption, on the scale of vegetation patterns considered here, is that these discontinuities are caused by:
a) habitat discontinuity;
b) similarities or dissimilarities in the ecological ranges of species along continuously variable habitat gradients;
c) chemical relations among species (allelopathy, commensalism).

A distinction has to be made between the actual spacial distribution of the vegetation on the surface of the earth about which one wishes to draw conclusions, and the distribution of the points representing sample plots or stands in the multidimensional test-space. Inadequate or biased sampling of the first can result in artificial discontinuities in the second.
If each species had the same chance of success in each sample plot (as measured by cover percentage, biomass, etc.) a random distribution would result. Also the action of many levels of different operational habitat factors on the vegetation can be expected to result in random distribution patterns.
Since a certain degree of clustering characterizes random populations it appears necessary to redefine clustering analysis as the detection and identification of groups of individuals that resemble each other more than members of other groups to a degree greater than can be expected from random distributions.

It is the purpose of this paper to present, and discuss in the light of the foregoing, the results of two types of cluster analysis of two sets of hypothetical randomly distributed "vegetation" data.

METHODS

To obtain data with a known random distribution, sets of random numbers between 0 and 100 were generated. These sets formed the coordinates for random points in the test-space and represented cover percentages of plant species in a number of sample plots. Hypothetical data were thus created for two different plant populations.
One set represented the data from 25 sample plots (or relevés, or stands) each containing 3 species (3 sets of 25 random numbers). The other represented 40 sample plots each containing 50 species (50 sets of 40 random numbers). The Euclidean distance was used as a measure of dissimilarity.

Two clustering methods were choosen for this study, representing two fundamentally different approaches to the formation of clusters. Both are polythetic agglomerative methods but the first one uses a single linkage and the second a sum of squares approach, resulting in differently shaped clusters even when the same data are used.
The single linkage method is based on graph theory and is described by van GROENEWOUD & IHM (1974). This method can sometimes result in starshaped or oddly shaped, elongated clusters (IHM 1978, p. 470).
The sum of squares approach is described by ORLOCI (1975). It selects fusions that minimize the within group sum of squares. This approach results in more or less compact spherical clusters.

RESULTS AND DISCUSSION

The results of the clustering analysis are shown in the form of dendrograms. Figure 1a shows the single linkage clustering of the 25 plot, 3 species data and figure 1b the results of the sum of squares clustering analysis of the same data.

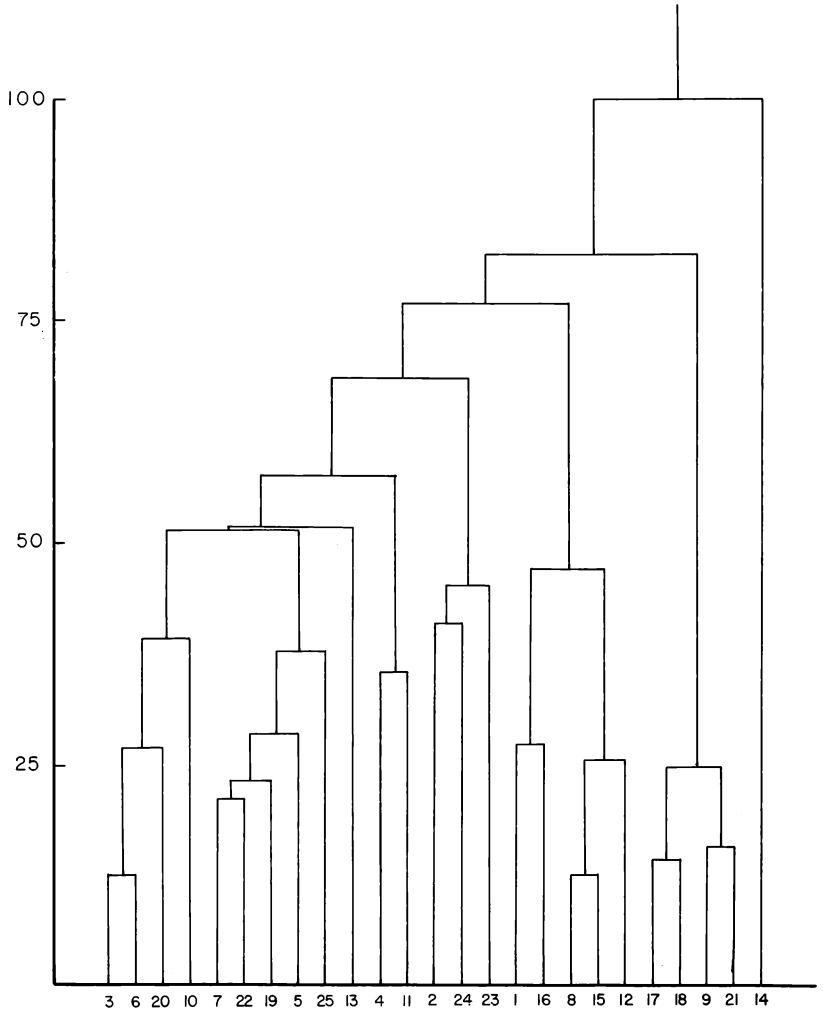

Fig. 1a. Dendrogram depicting the "single linkage" cluster analysis of a randomly distributed population (25 plots, 3 species).
Vertical scale shows the distance at which individuals or groups are linked, expressed as a percentage of the greatest distance.

A comparison of the two dendrograms quickly shows:
a) that essentially the same clusters can be recognized by either method; the clusters (1,16,8,15,12), (9,21,17,18), (2,24,23), (3,6,20,10), (4,11), and (5,25,7,22,19) can be recognized in both dendrograms, but
b) that the clusters resulting from the sum of squares analysis are more distinct than those resulting from the single linkage method. In the single linkage method the clusters are linked at much higher levels than in the sum of squares analysis, especially the group (17,18,9,21) is linked to the rest at a much different level. Also plot 14 has a single linkage with the rest in the first method but not in the sum of squares analysis.

Figure 2a shows the dendrogram representing the single linkage clustering of the 40 plot, 50 species data and figure 2b the sum of squares clustering of the same data. A different pattern appears in each.
The clusters are quite clearly defined in the dendrogram depicting the sum of squares analysis, but identical ones are not recognizable in the single linkage method. If clusters can be recognized at all in the single linkage method they are quite different from the ones recognized by the sum of squares method.

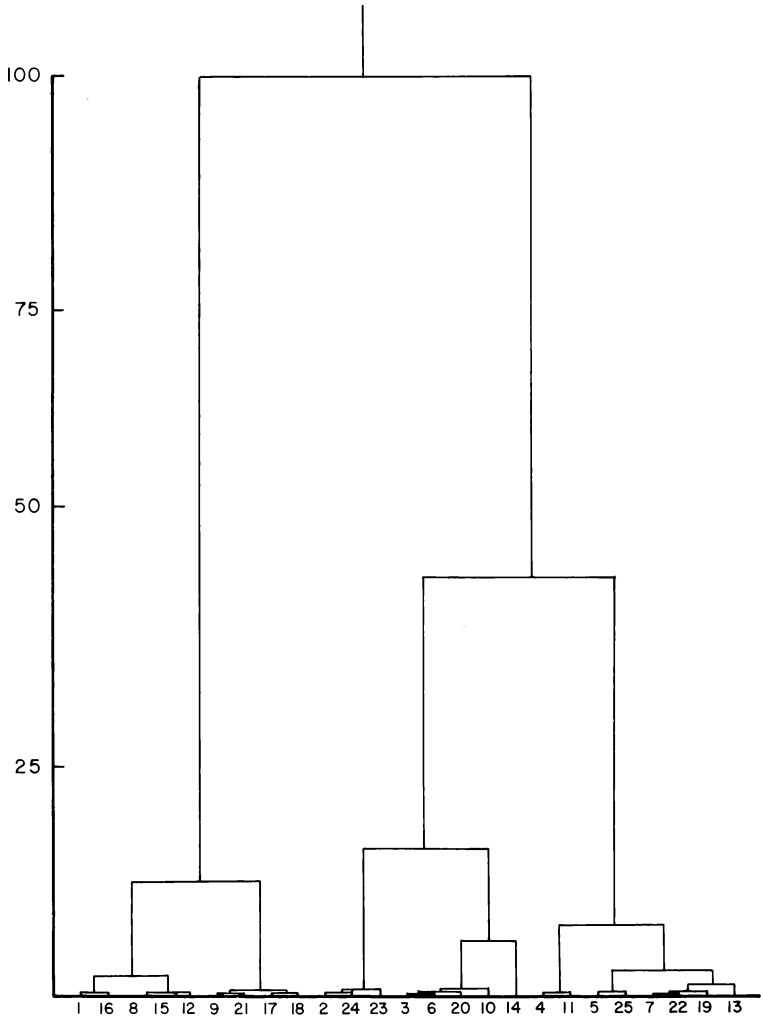

Fig. 1b. Dendrogram depicting the "sum of squares" cluster analysis of the same population as in figure 1a.
Vertical scale shows the within-cluster mean squares expressed as a percentage of the sample mean square.

In the single linkage method practically all points are joined at gradually greater distances and appear to form a continuum. There is thus a distinct difference in the results of the analysis of the two sets of data due to their different dimensionality. The ordering of the sample plots into clusters in the low dimensional test-space (3 species) is quite distinct and essentially the same for both clustering methods. This however, is not true for the high dimensional test-space (50 species). Here the single linkage method indicates a more or less continuous vegetation while the sum of squares shows strong clustering.
Considering that the data had a random distribution, a judgement based on the division of the vegetation into distinctly different communities based on the sum of squares clustering method can be misleading. The same is true for the single linkage method in the case of a vegetation type with very few species.

The type of chaining effect apparent in the single linkage method appears to be dependent on both the dimensionality and the distribution of the points in the test-space. Chaining sometimes may appear to be more due to the methodology followed than to the vegetation structure (LANCE & WILLIAMS, 1967). In my opinion, however, even methods that further the chaining effect would show definite clustering if true discontinuities were present.

The single linkage method emphasizes the formation of clusters less than does the sum of squares method. However, there does not seem to be a method for testing for discontinuities in the distribution of the points in the test-space. Mahalanobis' "generalized distance" or D^2 and Hotelling's T (both sorts

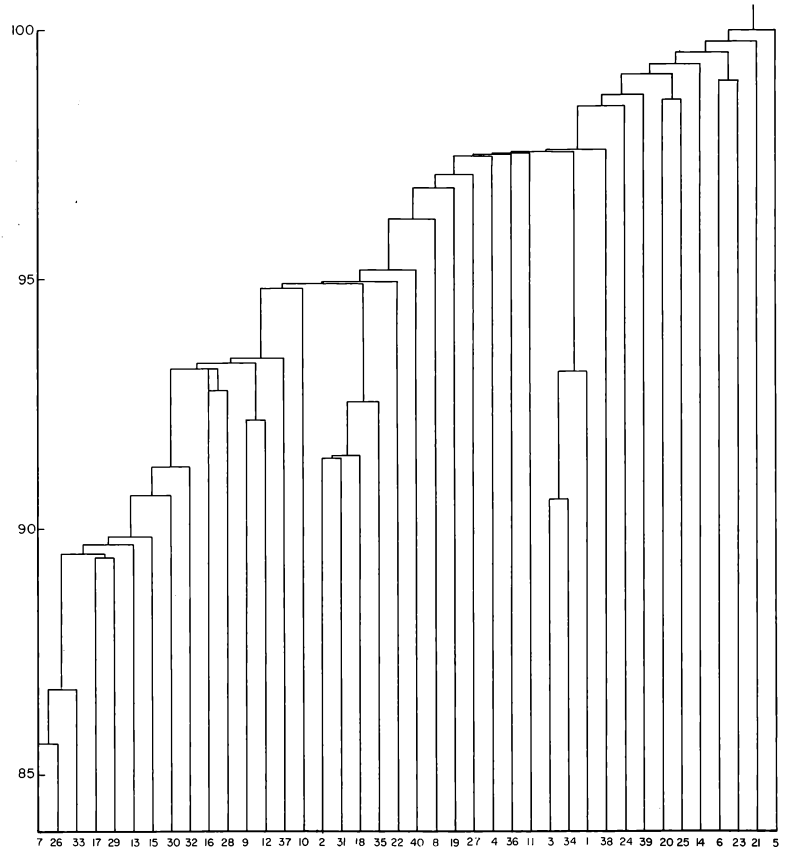

Fig. 2a. Dendrogram depicting the "single linkage" cluster analysis of a randomly distributed population (40 plots, 50 species).
Vertical scale shows the distance at which individuals or groups are linked, expressed as a percentage of the greatest distance.

of multidimensional t-tests) can be calculated among the clusters. This would give an indication of the statistical significance of the clusters but it would not answer the question of discontinuity posed before. The segments of a continuum, treated this way, would also indicate significant differences among them.

A comparison of published dendrograms depicting the results of cluster analyses of vegetation data, with the dendrograms shown in this paper makes it obvious that some test for discontinuities is necessary to enable conclusions to be drawn about the distinctness of the clusters.

GOODALL (1966, 1971) proposed a probalilistic approach based on the nul hypothesis that the samples are derived from a single population in which the attributes (species) are not correlated. His method, however, does not yet appear to have been used in vegetation analysis.

OBSERVATIONS AND RECOMMENDATIONS

1) At least some methods of cluster analysis cannot distinguish between clustering due to randomness of distribution or to true discontinuities.

2) High dimensionality of the test-space affects the single linkage clustering results and tends towards "chaining".

3) The sums of squares method of clustering results in dendrograms that show strong clustering, even with random distributions.

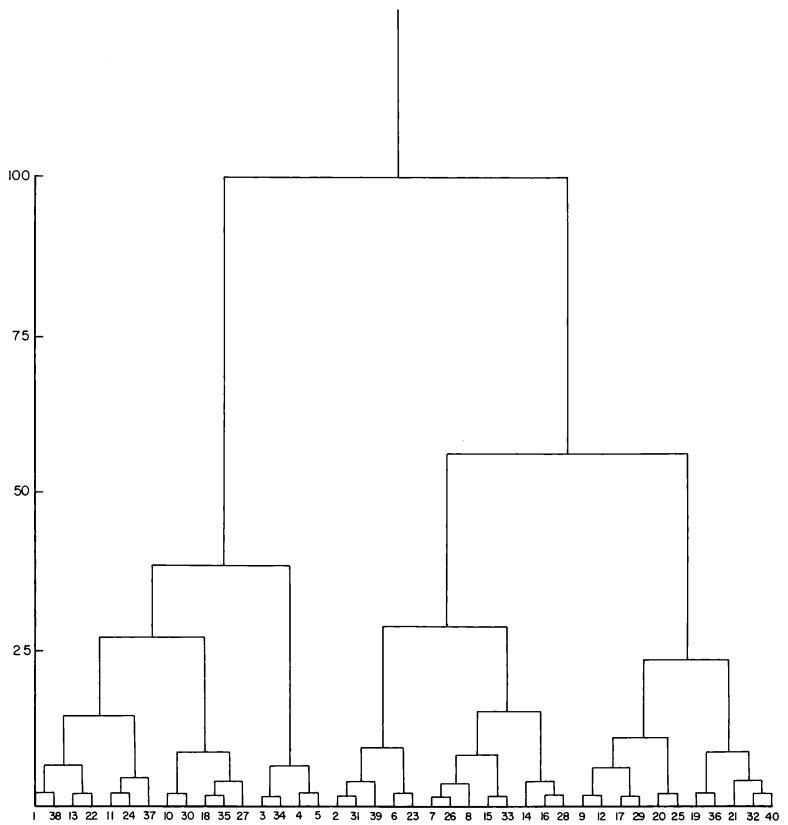

Fig. 2b. Dendrogram depicting the "sum of squares" cluster analysis of a randomly distributed population (40 plots, 50 species).
Vertical scale shows the within-cluster mean squares expressed as a percentage of the sample mean square.

4) Tests should be applied, to be sure that clustering is not due to randomness of distribution.

5) GOODALL's probalilistic approach and other similar methods should be tested on vegetation data with different structures.

LITERATURE

GOODALL, D.W. (1966): A new similarity index based on probability. - Biometrics 22: 882-907.

- (1971): Cluster analysis using similarity and dissimilarity. - Biometr.-Praxim. 11: 34-41.

GROENEWOUD, H. van & IHM, P. (1974): A cluster analysis based on graph theory. - Vegetatio 29(2): 115-120.

IHM, P. (1978): Statistik in der Archäologie. - Rheinland Verlag, Köln. 619 pp.

LANCE, G.N. & WILLIAMS, W.T. (1967): A general theory of classificatory sorting strategies. I. Hierarchical systems. - Comp. J. 9: 373-380.

ORLOCI, L. (1975): Multivariate analysis in vegetation research. - Dr. W. Junk. The Hague. 276 pp.

Anschrift des Verfassers:

Dr. H. van Groenewoud
Maritimes Forest Research Centre
P.O. Box 4000
Fredericton, N.B.
E3B 5P7
Canada