AN ESSAY IN THE USE OF ASSOCIATION AND DISSOCIATION MEASURES IN PHYTOSOCIOLOGICAL CLASSIFICATION

BY J. ARAOZ, G. SARMIENTO AND M. MONASTERIO

Departamento de Computación, Facultad de Ciencias, Universidad Central de Venezuela, and Facultad de Ciencias, Universidad de los Andes, Mérida, Venezuela

INTRODUCTION

Numerical taxonomy, especially the techniques of clustering, are being increasingly utilized in the resolution of classificatory problems in different fields of biological research. In this paper the results of applying an original clustering method to phytosociological data will be discussed. It is a polythetic and non-hierarchical classificatory system, centred upon the floristic similarity between stands or samples. A simple algorithm is used to obtain an initial classification, which is then improved by means of an association and a dissociation function until a sufficiently satisfactory classification is attained.

The present application of this clustering method concerns tropical savanna vegetation at Los Llanos Biological Station, Venezuela. The vegetation was originally sampled for an earlier study by means of 380 2×2 m quadrats, selected by stratified randomization within a 190 ha stand. In each quadrat, presence of vascular species was recorded. The general features of vegetation and environment, together with details about sampling procedure and floristic composition, have been described by Sarmiento & Monasterio (1969).

The clustering was performed on the IBM 360–40 computer of the Faculty of Sciences in the Central University of Venezuela and the computer programmes for this were written by J. Araoz.

METHODOLOGY

General principles

The basic principles of this clustering methodology were proposed by Araoz & Varsavsky (1967), Araoz (1968) and Varsavsky (1969). Araoz & Varsavsky (1967) reported a 'second order' method of numerical taxonomy for revising previously obtained classifications. The classes of the 'first order' classifications originated from grouping all the points whose distance to randomly selected centres were smaller than a previously fixed value. The 'second order' method classified the points according to their permanency in groups in several clusterings obtained by the successive application of the 'first order' procedure.

Araoz (1968) introduced an association and a dissociation function to evaluate the result of a classification, and proposed a system of numerical taxonomy where a preliminary classification is obtained by any method and the clusters obtained are improved by the application of certain criteria of class homogeneity based on these two functions.

By use of the 'second order' method to obtain the preliminary classification, a first application was made to a socio-cultural classification of 141 countries on the basis of twelve attributes, giving satisfactory results.

In this phytosociological analysis the basic classificatory procedure remains unchanged, but the greater complexity of the data, 380 quadrats (points) with fifty species (attributes), made necessary some modifications of programming and certain methodological refinements. The classification is accomplished in two steps. A preliminary clustering is first obtained by the repeated application of the 'first order' method and the selection among the classifications obtained from this procedure of that classification exhibiting the greatest taxonomic value according to Varsavsky's (1969) definition. Then, in a second step, this classification is improved by the use of the association and dissociation functions.

The starting point for the preliminary classification is the selection of a coefficient of similarity between two quadrats, i.e. a point-to-point 'distance' d in the multidimensional attribute space. This need not be a distance with all its metric attributes, but may be a 'quasi-metric' similarity coefficient in the sense of Williams & Dale (1964). A threshold of similarity u is then fixed, according to certain characteristics of the data and to the desired properties of the classification.

Then a point x (a quadrat in this case) is selected at random within the whole set, grouping in a first cluster all points y with a distance to x equal to or less than u:

$$d_{(y,\,x)} \leqslant u$$

After the exclusion of this first cluster, the same procedure is repeated on the remaining set of points, another point being selected randomly and a second cluster formed around it of all the points localized at a distance equal to or less than u. The same procedure is followed until all points have been clustered.

For this classification $\mathscr{C}$, its taxonomic value $TV_{(l)}$ is calculated according to Varsavsky's (1969) formula:

$$TV_l = HG_{(E)} - \sum_{C \varepsilon \mathscr{C}} HG_{(C)}$$

where E represents the whole set of points for classification, C the resulting classes and $HG_{(x)}$ the global heterogeneity of set X.

For points defined by binary attributes this value is given by:

$$HG_{(x)} = pN \log N - \sum_{i=1}^{p} (n_i \log n_i + (N - n_i) \log (N - n_i))$$

where p is the number of attributes, N is the number of points and n_i is the number of points in set X possessing the attribute i.

The preceding clustering procedure is repeated several times and that classification which has the greatest taxonomic value is selected.

The second part of the classificatory process starts with the application of the association and dissociation functions, which are defined for this application in the following way.

Given a finite set E, the association A of a point x belonging to E, with a set C included in E, is defined as the proportion of points 'similar' to x included in C. Two points are considered as 'similar' when their distance from each other is equal to, or less than, a pre-established similarity value,

$$A(x, C) = \frac{n(C \cap V_x)}{n(V_x)}$$

where: $n(C)$ = number of points in C and V_x = set of points similar to x.

The dissociation D of a point x with a class C is similarly defined as the proportion of points 'dissimilar' to x included in C; two points are considered 'dissimilar' when their distance from each other is equal to or greater than a pre-established dissimilarity value

$$D(x, C) = \frac{n(C \cap W_x)}{n(W_x)}$$

where W_x = set of points dissimilar to x.

A and D can take values between 0 and 1, and they show the following properties:

(a) $A(x, E) = 1$
$D(x, E) = 1$

(b) If $C_1 \cap C_2 = \varnothing$
$A(x, C_1 \cup C_2) = A(x, C_1) + A(x, C_2)$
$D(x, C_1 \cup C_2) = D(x, C_1) + D(x, C_2)$

(c) If $C_1 \subset C_2$
$A(x, C_1) \leqslant A(x, C_2)$
$D(x, C_1) \leqslant D(x, C_2)$

The association between two classes C_i and C_j can be measured by their average association AA defined in this way:

$$AA(C_i, C_j) = 1/2 \frac{\sum_{x \varepsilon Cf} A(x, C_j)}{n(C_i)} + \frac{\sum_{x \varepsilon Cj} A(x, C_i)}{n(C_j)}$$

Similarly, their average dissociation AD can be defined in the same way.

For a classification to be satisfactory the association of every point with its own class must be 'high' and the dissociation must be 'low', where 'high' and 'low' are defined by conventional thresholds of the association and dissociation functions. A second condition to be fulfilled is that the $AA(C_i, C_i) \geqslant a$ and the $AD\ (C_i, C_i) \leqslant b$, where a and b are also pre-established thresholds. Finally one can expect that the AA between different classes will be small when the clustering is satisfactory.

To obtain the final classification, the association and dissociation of each point with every class of the preliminary clustering is calculated together with the AA and AD between each pair of clusters. If some misclassified points appear, i.e. points with 'high' dissociation with their own class or with 'high' association and 'low' dissociation with some other class, they must be moved to the class with which they present the best association and dissociation values. In the same way if two clusters show a 'high' AA and a 'low' AD with each other, they must be grouped together into a single cluster.

If a point can be classified in several clusters, i.e. if it has high association with them, we can employ other classificatory criteria, as cluster size, or any other criterion considered appropriate to the aims of the research.

If a certain point has a high dissociation with every class with which it is associated, this point is considered as 'non-classified'. This concept of non-classified points inside a classification has been introduced by Araoz & Varsavsky (1967) as a mean of characterizing those points which cannot be clearly included in any class, i.e. points present as

'noise' in the space between the classes. However, the association function as used by Araoz (1968) allows us to recognize the classes with which a non-classified point is more related, notwithstanding the fact that the point does not belong to these classes.

Through these manipulations a classification can be obtained in which each point has both a dissociation with its own class smaller than the pre-established threshold of dissimilarity and an association greater than the similarity threshold. Generally several intermediate steps will be necessary. After each step, consisting in the movement of certain ill-located points, all associations and dissociations are recalculated, in order to evaluate the results of the changes that have been made.

Selection of distances and thresholds

The methodology allows various possibilities of choice in each of its steps. Thus, for example, the first problem to solve is the selection of the point-to-point distance to be employed. Ecological literature shows a wealth of coefficients of similarity between stands which can be utilized in this context (see Dagnelie 1960). The coefficients which have been more frequently employed in phytosociology are those of Jaccard and Sørensen, both non-metrical and monotonic with each other. Either of these two coefficients can be used as a distance measure, because they give a real picture of the inter-point similarity. Other coefficients, e.g. the Euclidian distance, are not suitable for this purpose, because they take into account only the non-common attributes between two points, disregarding the common ones.

We have selected the complement of Sørensen's coefficient of similarity as the measure of the distance between two points i and j, that is:

$$d_{ij} = 1 - \frac{2a_{ij}}{a_{ii} + a_{jj}}$$

where a_{ij} are the attributes common to the points i and j, a_{ii} are the total attributes exclusive of point i and a_{jj} are the total attributes exclusive of point j.

As the distance values range between 0 and 1, one could arbitrarily fix any small value as a threshold of similarity, considering any pair of points lying apart at an equal or smaller distance to be similar enough to be included within the same cluster. However, if we demand certain convenient formal characteristics from the preliminary clustering, such as a small number of clusters each one with a roughly equivalent number of points, it seems advisable to select the threshold according to the real distribution of distances. In this way the configuration of the whole set of points in the distance hyperspace can be taken into account.

For each of the 380 points, the total number of neighbours lying at increasing distances from 0·25 to 0·80, taken at 0·05 units intervals was calculated. These figures showed that for a distance less than 0·35 a too great number of points had only very few neighbours; whilst for a distance greater than 0·40 nearly all the points were neighbours of each other. It follows from these figures that a similarity threshold less than 0·35 will lead to the formation of an excessively high number of few-membered clusters; on the other hand, a threshold greater than 0·40 will lead to a few clusters, each one with numerous members. For every point the number of neighbours situated in the distance range between 0·35 and 0·40, taken at 0·01 units intervals, was then calculated. In this manner the distance value of 0·37 was selected as the similarity threshold giving rise to the most convenient number of clusters. Fig. 1 shows the distribution of the neighbouring points for five different distances

Association and dissociation thresholds

In calculating the association and dissociation of each point with every cluster, those points whose distance was equal to or less than 0·37 were considered as 'similar', using therefore in the definition of these functions the same similarity value as was employed in the first clustering. On the other hand, the value of 0·60 was selected for defining 'dissimilar' points in the dissociation formula; this value was chosen by applying the same principles and procedures as were used in deciding the similarity threshold.

It remains to decide the criteria for considering a dissociation value as 'low' and an association value as 'high', and thus assessing the worth of the whole classification. Two basic criteria were followed. A point was considered as well classified when its dissociation with its own class was equal to 0; that is to say, a class was accepted as good when it did not have dissimilar points ($d_{ij}<0{\cdot}60$). The second criterion was that, when a point did not have dissimilar points in several classes, it had to be included in that class having the highest association with it.

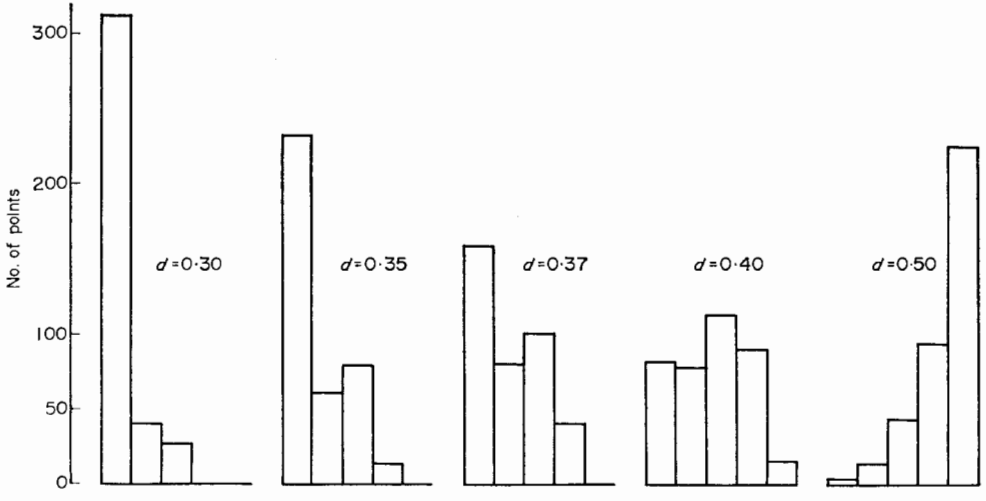

FIG. 1. Variation in the number of neighbours with increasing distances. The five classes represent the number of points with 0–10, 11–20, 21–50, 51–100 and more than 100 neighbours.

COMMUNITY ANALYSIS

Classification

The final classification of the 380 quadrats on the basis of fifty species (those with a frequency greater than 5%) gave rise to twenty-four classes, plus eleven isolated points and forty-two non-classified elements. The number of points within each class ranged from two to forty-nine.

Seven steps were necessary to obtain this result from the preliminary clustering. In each step some points were moved from one class to another with which they also showed a null dissociation but a higher association. Each running of the programme took about 10 min of computer time.

The eleven isolated points can be considered as one-membered classes whose elements do not have any neighbours at distances equal to or less than the similarity threshold ($d = 0{\cdot}37$). The non-classified points are quadrats which do not fulfil the criteria used for

class formation, either by having dissimilar points in all classes or by not having similar points in those classes in which they do not have dissimilar points.

The classification and the interclass relationships may be depicted by a network (Fig. 2) where the noda represent the classes and the interconnecting segments indicate the association and dissociation between them. In Fig. 2 the noda are not connected at all when these

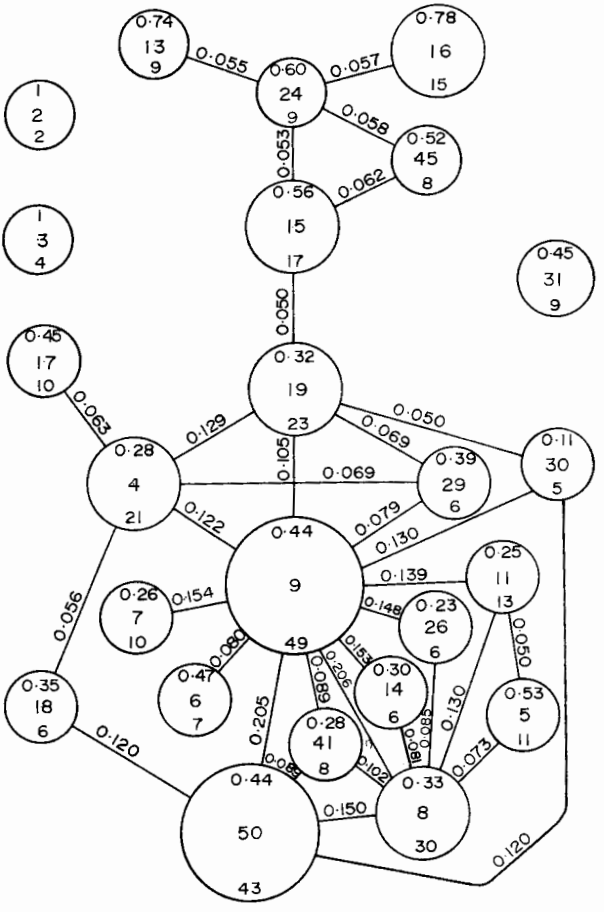

FIG. 2. Network representing the interrelationships between the classes. Two noda are interconnected when their AA are equal to or greater than 0·05 and their AD equal to or less than 0·03. The central figure inside each circle is the number of the class, the upper figure is the AA of the class and the lower figure is the number of points within the class. The figures on the connecting segments are the AA between the classes.

values are too low ($AA < 0{\cdot}05$) or too high ($AD > 0{\cdot}03$) respectively. Two of these twenty-four classes are completely isolated: classes 2 and 3, which have an AA $(C_i, C_i) = 1$, that is to say, their members do not have any similar point in the other classes. Another class, 31, is not linked with any other at the forementioned association level. The remaining twenty-one classes are more or less interrelated forming the network showed in Fig. 2.

The main feature evident in the overall pattern of clustering is a central core integrated

by thirteen closely interrelated clusters showing multiple and narrow interconnections. From this central core emerge, on one side, three classes attached to it by only one connection (classes 6, 7 and 17) and, on the other side, a more diffuse grouping of five classes (13, 15, 16, 24 and 45) linked to the core by a unique association between classes 15 and 19.

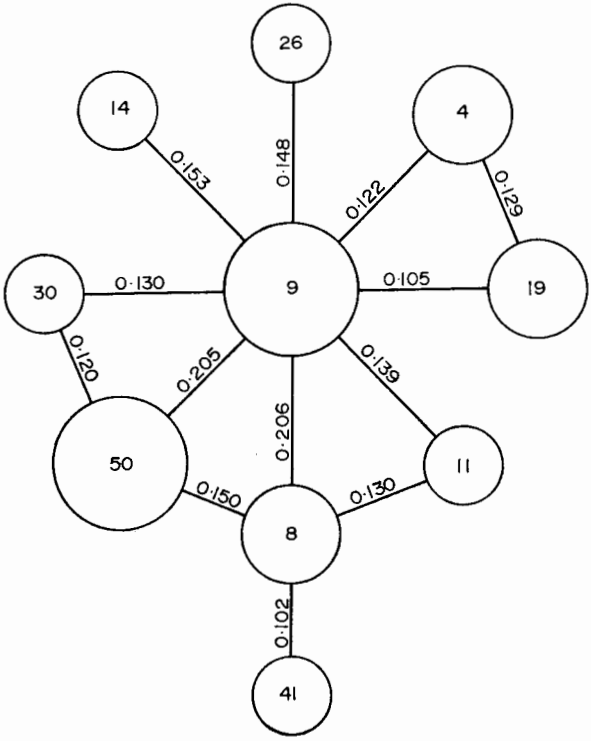

FIG. 3. Representation of the interclass relationship when only classes with an *AA* greater than 0·10 are interconnected.

If the level of association required for linking two clusters is raised to $AA > 0{\cdot}10$, the network shown in Fig. 3 is obtained, where only a part of the central core remains, reduced to a solid cluster of seven classes, with three additional classes (14, 26 and 41) linked to it by only one connection, whilst all other interconnections have disappeared. As is shown in these two representations, a great majority of points, 312 out of 380, are grouped in a more or less diffuse cluster, having in turn a central core of greater density around class 9 (ten classes with a total of 204 points); fifteen points are included in three isolated classes; the remaining fifty-three points form one-membered isolated classes or are non-classified elements.

Ecological interpretation

To evaluate the ecological significance of the classes formed in the clustering process, it is necessary to analyse their homogeneity with reference to some environmental and vegetational factors. In a previous paper (Sarmiento & Monasterio 1969) the following factors were considered: land form, topographic position, soil parent material, depth of

lateritic cuirass, development of soil profile and characteristics of each horizon, dominant species, total cover of the herbaceous layer and distance to the nearest grove. The successive subdivisions produced in an association-analysis separated groups differing in one or several of these features. Using this ecological interpretation as a reference we are able to evaluate the results of the present method.

The first order division in the association-analysis sets apart two different types of savannas: those of the $+$*Trachypogon vestitus* Anders. branch occurring in lower sites, on yellow-reddish loams, which always have this grass as a dominant species; the $-$*T. vestitus* branch is composed of savannas having *T. plumosus* (Humb. & Bonpl.) Nees as a dominant species and occurring in higher sites, on soils developed from red-yellowish sandy loams. In the present clustering this main subdivision is readily apparent; seven clusters are exclusively composed of $+$*T. vestitus* quadrats, and twelve clusters of $-$*T. vestitus* quadrats, and in only five classes are the groups intermingled. As shown in Fig. 4, the seven former classes are either isolated (classes 3 and 31) or they form the upper part of the clustering configuration (classes 13, 15, 16, 24 and 45).

Among the five 'mixed' clusters, three different cases can be considered. In class 5, quadrats of groups *A*, *J* and *K* of the association-analysis are put together. Significatively, these three groups, although belonging to different branches of the first order subdivision, are quite similar ecologically, because they include the savannas occurring on the hard lateritic cuirass outcrops. Class 19, as can be seen from Figs. 2 and 4, serves as the connecting link, between the central core, mainly occupied by the $-$*T. vestitus* savannas, and the upper part clusters, exclusively formed by the $+$*T. vestitus* quadrats. Therefore, it can be considered as a transitional class. The third case is that of classes 17, 18 and 30, where the savannas occurring on shallow soils of the $+$*T. vestitus* group (group *E*) are united with certain groups of the other branch occurring either on shallow soils (group *Q*), or in the vicinity of groves (groups *N* and *M*) or on deeper soils (group *V*). Here, apparently, another operative factor, scarcely detectable with our limited ecological sampling, must be at work.

A remarkable fact evident in most of the clusters grouping the $+$*T. vestitus* quadrats, is the co-occurrence within the same cluster of points belonging to the association-analysis groups *H* and *G*, *H* and *E*, *G* and *E*. This fact is a natural consequence of the great floristic similarity between them, but it also demonstrates the danger of applying a unique method of phytosociological analysis to complex data, and thereby overemphasizing certain factors or species to the detriment of others which may be at least equally significant in certain respects.

Without attempting a thorough assessment of the ecological significance of each cluster composing the central core of closely interrelated classes, we can draw certain general conclusions about them. In the first place, Fig. 4 shows how in each cluster a high proportion of points belongs to one or another of the final groups of the association-analysis. For instance, in the two largest classes, 9 and 50, twenty-six of the forty-nine members of class 9 belong to group *S* and fourteen to group *Q*, and of the forty-three points of class 50, twenty-four belong to group *Q*. Another feature observable in the clustering is that in general the most closely related association-analysis groups are represented in the same cluster. This is the case of groups *S* and *Q* in class 9 and of groups *L* and *M*, *V* and *W*, *S* and *R* in other clusters. In the third place points belonging to the same association-analysis group tend to be concentrated in the more closely related clusters. Thus, for example, group *S* is concentrated in clusters 8, 9, 11 and 26, all united by strong association links. The same can be said of groups *Q* and *M*.

Two additional features appearing in a comparison of the results obtained with the two methods refer to the single-membered classes and the negative, right hand parts of the association-analysis hierarchies. Most of the one-membered classes obtained in the previous paper, as the classes *B*, *D*, *F*, *I*, *P*, *T* and *U*, also appear in the clustering as isolated one-membered classes; or where this is not the case, they appear among the non-classified points. This fact emphasizes the exceptional or atypical character of these

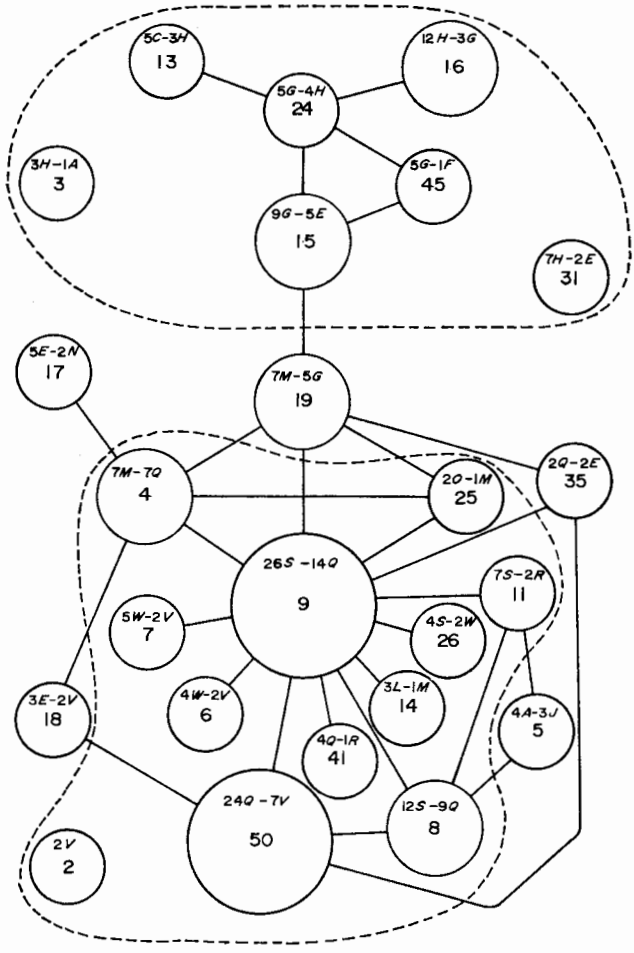

FIG. 4. Correspondence between the groups obtained by the clustering method and those produced by association-analysis (Sarmiento & Monasterio 1969). All the groups belonging entirely to one of the branches produced by the first association-analysis subdivision (based on *Trachypogon vestitus*) are enclosed by a broken line. The indication in the upper part of each circle refers to the number of points belonging to the two classes of the association-analysis with greatest representation inside this group.

points. On the other hand, most of the non-classified points in the clustering belong to the negative, right hand border of the association-analysis hierarchy. This feature confirms effectively that the 'minus' branches of the association-analysis are more heterogeneous and less well defined than the 'plus' branches, and this fact becomes accentuated

towards the right hand half of the hierarchies. For example, of the twenty-seven points forming group *W*, at the right edge of the $-T.$ *vestitus* branch, only 10, and of the twenty-one points of its neighbouring group *V* only fourteen, were classified.

Summarizing now the preceding observations about the ecological significance of this classification, we note the correspondence with the results obtained previously with the association-analysis, improving it in certain cases, as when the lateritic cuirass savannas are grouped together in a unique cluster; but generally the results are less clear or the agglutinating ecological factor less evident than in the preceding hierarchical classification.

DISCUSSION

The clustering method applied in this phytosociological research has produced a classification that reveals some interesting aspects not apparent in a direct observation of the raw data. Probably the main advantage obtained with the clustering is an effective representation of the complex relationships existing between the classified points. Indeed, one can easily grasp from the clustering the main features characterizing the phytosociological hyperspace in which the data were placed.

In this particular application of this clustering method to the analysis of Los Llanos Biological Station's savannas, the final classification shows clearly the high degree of floristic similarity between the quadrats, which makes it difficult to trace sharply defined limits between groups or classes. This is shown both by the high association between the clusters and by the numerous non-classified quadrats which form a loose nebula of points surrounding the more compact clusters.

On the other hand, the existence of many atypical quadrats isolated as one-membered classes is also evident, producing a discordant note inside the highly interrelated set formed by the majority of the points. The clustering also picks out the centre of gravity of the sample, represented by the central core of clusters still more closely associated with each other, and so gives an indication of the principal directions of divergence which radiate from this central core.

In this manner an evident element of order and patterning emerges from the clustering and this is precisely the main function we may demand from a cluster analysis. On the contrary, the system is not fully adequate to obtain clear-cut groups or to detect major operative ecological agents responsible for these discontinuities. In this respect a hierarchical and monothetic system such as Williams & Lambert's assocation-analysis is obviously more useful and pertinent.

This methodology has certain advantages compared with previous clustering principles and procedures which have been applied to the same sort of problems (see Sokal & Sneath 1963, for a review of currently used methods of cluster analysis). Employed as an exploratory tool for revealing the overall pattern hidden in a complex set of data, the system here presented has two important, and to a certain measure original, properties.

In the first place the system possesses a remarkable flexibility, given by the numerous choices the ecologist can make during the process, adjusting each step to his pre-established aims and desiderata. Each step allows a wide range of possibilities, first the selection of the distance function to be employed, and then the fixing of the similarity and dissimilarity thresholds, the definition of the criteria of association and dissociation to be used, and the formulation of the desired properties to be demanded from the clusters. Lying completely within the field of classification of qualitative data using unloaded attributes,

this method offers more a pattern of procedure rather than a crystallized recipe to be applied blindly. Only the nature of the raw data and the finality of the clustering set limit the flexibility of the methodology.

The second original property is the close interaction it allows between the user and the results, and this is obtained without abandoning the strict objectivity secured by a reproducible and clearly established methodology. In effect, the ecologist can manipulate the results as they are being produced, until he is sufficiently satisfied with the end product. By simplifying or modifying the clusters according to his previously established criteria, he can model to a wide degree the obtained result. In this way it is possible to give more or less emphasis to one or another property of the data, reducing or increasing the number of clusters and varying their composition.

These two attributes, flexibility and ease of handling, are without doubt especially useful in phytosociological and ecological research, because in these fields it is important to explore the main characteristics of a complex set of data than to produce a classification.

ACKNOWLEDGMENTS

The authors wish to express their warm gratitude to Dr O. Varsavsky who inspired and encouraged both the research in numerical taxonomy and its application to ecological problems. We are also indebted to Dr M. Bemporad, Chair of the Computation Department of the Central University, who made every effort to facilitate this work.

SUMMARY

An original method of cluster analysis is applied to the classification of a stand of tropical savannas in the Venezuelan llanos. The vegetation sample consists of 380 presence/absence quadrats, where fifty species were recorded and further employed as attributes in the clustering.

The first step is to obtain several independent preliminary clusterings by grouping the quadrats in classes according to their distances to randomly selected centres; then the clustering having the highest taxonomic value is selected as the basis for the subsequent steps.

By defining an association and a dissociation function between points and classes and between classes, and by fixing certain criteria that must be fulfilled by each class, the preliminary clustering is improved. This process is accomplished by moving points to the classes with which they show the best association and dissociation values, and this is continued until a satisfactory final classification is obtained.

The remarkable flexibility of the procedure, together with the close interaction between the ecologist and the results that it allows during the performing of the classification, suggest that this method may be a useful exploratory tool to elucidate the main elements of pattern and order hidden in a complex set of phytosociological data.

REFERENCES

Araoz, J. (1968). Asociación en taxonomía numérica. *Acta cient. venez.* **19,** 187–92.

Araoz, J. & Varsavsky, O. (1967). *Un método de segundo orden para taxonomía numérica.* Asociación Venezolana para el Avance de la Ciencia, XVII Convención Anual, Resúmenes, 135.

Dagnelie, P. (1960). Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bull. Serve. Carte phytogéogr.* Sér. B, **5**, 7–71.

Sarmiento, G. & Monasterio, M. (1969). Studies on the savanna vegetation of the Venezuelan llanos. I. The use of association-analysis. *J. Ecol.* **57**, 579–98.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* Freeman, San Francisco.

Varsavsky, O. (1969). *Entropía y Taxonomía Numérica.* U.C.V., Facultad de Ciencias, Departamento de Computación, Publ. 69–01.

Williams, W. T. & Dale, M. B. (1964). Fundamental problems in numerical taxonomy. *Adv. bot. Res.* **2**, 35–68.

(*Received* 3 *December* 1969)