COENOSES 3 (2): 69-78, 1988



# VARIABLE CENTERED METHODS AND COMMUNITY CLASSIFICATION

H.M. André, Musée royal de l'Afrique centrale, Entomology section, B-1980 Tervuren and Laboratoire d'Ecologie et de Biogéographie, Université Catholique de Louvain, Place Croix du Sud, 5, B-1348 Louvain-la-Neuve.

**Keywords:** Community, Classification, Variable centered algorithms, Hierarchy, Similarity

**Abstract:** Variable centered methods of classifications are briefly described and classified into three types: multiple partitioning, initial clusters and dynamic clusters methods. A program for each has been tested (OSUCL4, TABORD and PARTS). Inasmuch as biotic community composition is studied and subjected to comparison and classification, the author proposes to use consistently variable centered classification algorithms, particularly the multiple partitioning methods such as OSUCL4. Special attention is given to the hierarchical interpretation of this method, initially designed as nonhierarchical, by preventing the dichotomous process usually associated with hierarchical procedures. Tests also indicate that the use of percentage dissimilarity give ecologically most easily interpretable results.

*If classificatory procedures are to be used rather than played with, ecologists must choose one or a very few methods and use them consistently. The choice will unavoidably be somewhat arbitrary.*
Pielou (1977 : 331)

## Introduction

Community ecology or biocenotics concerns assemblages of plants or animals living together in a given area or physical habitat. There are many possible approaches to classifying such assemblages. These emphasize different kinds of community characteristics as the bases of classification. For instance, phytosociologists may emphasize the floristic composition of vegetation or its physiognomy. Because of these various approaches, many schools have evolved in phytosociology, each with distinctive methodologies and emphases. Their history is reviewed by Whittaker (1962) and Shimwell (1971). Classification of animal communities is much more recent and, most often, workers emphasize community composition.

Hereafter, biotic communities will be characterized by their composition. Indeed, even if taxon-free approaches are useful in special cases (cf. Orlóci and Stofella 1986), the full species composition of communities is thought to better express their relationships to one another and to the environment than any other characteristic (Westhoff and Maarel 1973, Whittaker 1973). The approach implies that ecologists abstract from field data a formal description of communities and recognize community-types. The characteritics of the community-type may be generalized. As emphasized by Whittaker (1973), such a definition will describe a *class*, a grouping of samples, by their shared characteristics and, accordingly, a community, as a class-concept, will be inescapably an abstraction. Most often, the class concept is associated with the notion of 'average community' represented, in mathematical terms, by the centroid of the class (see below). The centroid, as any other abstraction, does not occur in nature, but this does not mean that it is meaningless as stated by Beals (1973). The conflicting views regarding abstract vs. concrete approaches is classical (cf. Barkman 1970, 1973). Specific methods have been designed for ecologists interested in defining concrete sociological groups (see a recent example in André 1984, 1985b).

Even if species abundances in relevés or samples, are accepted as the primary data for community classification, many methods remain available to ecologists. Different requirements and associated methods have been extensively discussed in numerous publications, informal vs. formal, exclusive vs. polythetic, monothetic vs. polythetic, qualitative vs. quantitative (see e.g. Pielou 1977, 1984, Orlóci 1978, Gauch 1982, Greig-Smith 1983, Legendre and Legendre 1983, 1984). In comparison, the problem of selecting hierarchical vs. reticulate methods has been given little attention. My aim in writing this paper is threefold: (1) to draw attention to the assumptions implicit in hierarchical methods, (2) to test comparatively different "variable centered classification" algorithms, and (3) to propose the generalization of the one which is compatible with (1).

## Hierarchical structure: pros and cons

As noted by Goodall (1986), classification implies nothing about any possible relationships between classes. The classes are recognized as discrete groupings - that is all. However in a hierarchical classification, the classes at any level are subdivisions of classes at a higher level. The results are expressed in the form of a dendrogram, i.e. an acyclic connected graph. This means that a hierarchical procedure does not yield a classifi-

cation directly, but it remains to be derived by cutting accross the dendrogram. Three major solutions for selecting a level - or stopping rules - are presented and discussed by Lambert and Williams (1966) and Pielou (1977). However, the resulting dendrogram has also an advantage in that a simple analysis may be viewed on several levels at once. Hierarchical procedures are also supposed to be better known, less cumbersome and more widely used in ecological work than the others (Lambert and Williams 1966). Most have, however, a major drawback, since they contain no provision for reallocation of entities which may have been poorly classified at an early stage of the analysis. Possibly this can be done in follow-up analyses by other methods.

The mathematical or technical reasons notwithstanding, the question arises as to whether hierarchical classifications are ecologically meaningful. Goodall (1986) stated that there is little reason to assume that differences between associations will fit into a hierarchical stucture. This statement is not supported by the past history of phytosociology, a history rich in numerous hierarchical systems proposed well before the use of modern quantitative methods (see Whittaker 1962, Shimwell 1971). Goodall's assertion is denied by Gauch and Whittaker (1981) who discuss the theory of hierarchical classification in application to ecological communities. Furthermore, Simon (1962) cited by Gauch (1977), Piaget (1967) and Koestler (1978) provide a philosophical basis for viewing communities hierarchically. All living beings survive and reproduce within a hierarchy of environments like a nest of Chinese boxes packed into each other. To this hierarchy of environments corresponds a hierarchy in communities. Burges (1960) and André (1985a) focused attention on communities which may extend for only a few centimeters or perhaps only a few millimeters. These microassociations are part of larger associations which, in turn, are part of larger communities, and so forth. The whole set forms a hierarchy seemingly open-ended at the top. As emphasized by Koestler (1978), at each level of the hierarchy, communities are sub-wholes in their own and form stable, integrated structures, equipped with self regulating devices and enjoying a considerable degree of autonomy.

There is however another characteristic implicit in most hierarchical procedures in ecology, which has been neglected by ecologists, the dichotomous process. If a divisive procedure is used, this means that the whole set of data is divided into two parts, each of them being in turn divided into two parts, and so on. No serious ecological argument supports such an approach. In addition, such an approach is likely to distort the real structure of data as shown by the simple example of Fig. 1. For convenience, this example involves three series of samples with only two species, one series where species A is dominant (90, 95 and 100% as relative

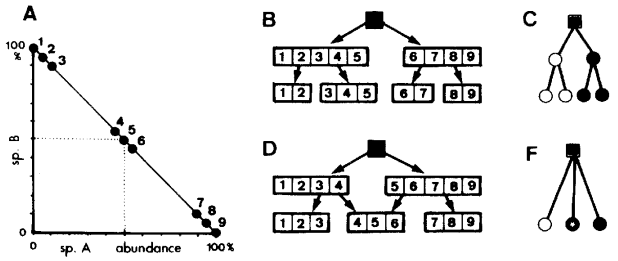


**Fig. 1 (A) Dummy data comprising nine relevés (1 to 9) with two species, A and B. (B) Successive partitions obtained by using divisive and hierarchical methods with the corresponding dichotomous structure imposed on data (C). (D) Successive partitions obtained by using OSCUCL4 with the corresponding trichotomous dendrogram (E).**

abundance), another where species B is dominant (same values) and a third series where species are equally represented. Most hierarchical procedures fail to yield an appropriate classification because they impose a dichotomous structure on data which obviously present a trichotomy. If a hierarchical approach seems to suit the ecological concept of biotic communities, the dichotomous process implemented in most algorithms seems to be unacceptable. This clearly implies that ecologists should use either a nonhierarchical method or a hierarchical approach that respects the possible "polyotomy" of data.

## Variable centered classification

Hereafter, I shall focus attention on variable centered classification algorithms - often called nonhierarchical clustering methods (Anderberg 1973) or partitioning techniques (Everitt 1974). The term "nonhierarchical clustering" is unfortunate, since, as explained in the next section, a hierarchical interpretation may be associated with the results. The variable centered algorithms are likely not to impose a rigid hierarchical structure on data. From a mathematical point of view, they are considered to be the most appropriate partitioning techniques for treating large data sets (Lebart *et al.* 1979). As Everitt (1973) suggested, these algorithms differ in three points: (1) the method for initiating clusters, (2) the method for allocating entities to initial cluster, (3) the method for reallocating points once the initial classification has been completed. Roughly, two major approaches exist to initiate a classification which give rise to two main types of algorithms: multiple partitioning methods and initial cluster methods. From the latter is derived a third type, the dynamic clusters methods.

### *Multiple partitioning methods*

In the first type, multiple partitioning, all the sample points are assigned to one group whose centroid is $C_{1.1}$ (the first subscript refers to the partitioning level and the second to the cluster identification num-

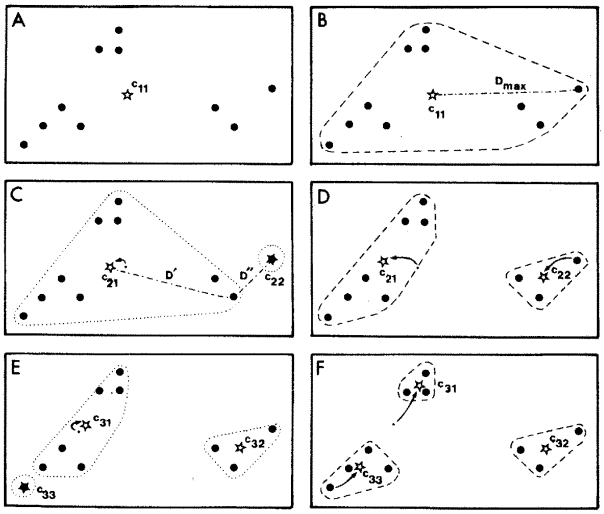


**Fig. 2. Procedure followed by a multiple partitioning method (OSUCL4). (A) The centroid $C_{1.1}$ of all points is estimated. (B) The most distant point from the centroid is identified. (C) The former point serves as a seed point ($C_{2.2}$) for the second cluster while $C_{2.1}$ moves (arrow) following the loss of $C_{2.2}$. (D) Reallocation procedure untill all points are stable. (E) The most distant point from its own centroid serves as a seed point for the third cluster. (F) The final 3-class partition (stars represent centroids, points represent the former position of centroids, arrows indicate the movement followed by centroids; dotted and broken lines respectively oultine intermediate and final clusters at each partitioning level).**

ber) is computed (Fig. 2A). Then this first group is subdivided into two clusters by selecting the point themost distant (dissimilar) from the initial centroid $C_{1.1}$ (Fig. 2B). This point and the centroid are then used as cluster nuclei - also called seed points (Anderberg 1973), center points or "étalons" (Diday 1971) - around which the set of sample points can be grouped. This implies that each sample is compared to the two seed points and placed in the cluster with which it shows the closest link (Fig. 2C). The procedure tends to minimize the distance between points belonging to the same cluster or to maximize the within-group similarity. After the group affiliation of all sample points is determined, a second iteration starts in which again all sample points are compared to the two new centroids $C_{1.1}$ and $C_{1.2}$. Usually, after a few iterations, the two clusters are stable (Fig. 2D). In the next step, the seed point of a third cluster is selected, the sample which is the most distant (dissimilar) from its own cluster centroid (Fig. 2E) and the process described above is repeated untill the user-supplied number of clusters is formed.

It must be emphasized that, at each partitioning level and iteration, points are likely to be subjected to reallocations, *i.e.* transferred from one cluster to another. In actual fact at each partitioning level, clusters may undergo four transformations: they may be maintained (Fig. 3A), merely divided (Fig. 3B) or involved

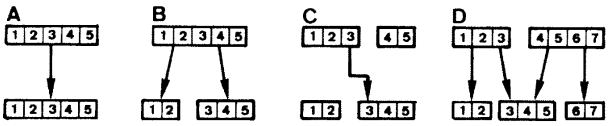


**Fig. 3. The four transformations possibly undergone by clusters between two partitioning levels: statusquo (A), mere divison (B), point transfer (C) and partial fusion (D). The latter two transformations illustrate point reallocations and imply a nonhierarchical relationship between the two partitioning levels.**

in some kind of point reallocation, either through a mere point transfer (Fig. 3C) or through a partial fusion between two clusters (Fig. 3D). It must be emphasized that the occurrence of point reallocations between two partitioning levels negates the existence of a hierarchical relation between these two levels.

The result of the method may thus be presented in the form of a dendrogram or otherwise, depending on whether there were reallocations at all partitioning levels. If the results are expressed in dendrogram form, the term "nonhierarchical method" is inappropriate. Furthermore, the tree is not necessarily dichotomous. When such an algorithm is applied to data as in Fig. 1A, the resulting dendrogram (Fig. 1E) has a trichotomous structure. Indeed, the reallocation of points 4, 5 and 6 at level 3 (Fig. 1D) is interpreted as an indication that the 2-groups partition is inappropriate for describing the real structure of data. Generally, any reallocation occurring between two partitioning levels is a sign that the higher level is "inadequate".

### *Initial clusters methods*

Jancey's (1966) method is a typical case of the second group of algorithms. These methods begin with the establishement of a user-supplied number of initial clusters. The initial clusters are formed by the investigator or, if one wishes to avoid any personal influence, generated automatically by the program. In a variant, k sample points are chosen, at random or preferentially, and serve as nuclei around which other samples are clustered to the nearest center point. The centroid is then computed for each initial cluster; these centroids become the new nuclei around which all other samples are clustered again. This reallocation procedure is repeated untill the clusters are stable. Generally to avoid "useless" or endless calculations, a maximum number of iterations is either defined by the user or provided by the program. Alternately, a stopping rule such as that proposed by Jancey (1974) might be applied.

Such methods are also known in the literature as the k - means methods (MacQueen 1967). Their major drawback, already noted by Orlóci (1967), is that they require assumptions about the number of groups existing within the population before analysis of the data.

*Dynamic clusters methods*

The third group of methods is derived from the second. The procedure is decomposed into two phases. The first one consists of getting a partition of k groups as indicated in the second group of methods. The seed points are chosen at random. The partitioning into k groups is repeated m times, each time with seed points chosen at random. The m partitionings into k groups resulting from the m iterations allow the identification of stable groups or strong forms. These stable groups are formed by sets of samples which are grouped together in all partitionings. These strong forms sensu Diday (1971) constitute the final classes selected by the method. Any other group of points which has been formed for one or several iterations but does not belong to a strong form corresponds to a weak form. Possibly, points which do not belong to strong forms are grouped into a residual class. In fact, these methods have been designed for pattern recognition and pertain to density search or density seeking techniques.

## An experiment with methods

*The programs tested*

Only programs using variable centered algorithms, which had been previously used in community analysis were tested. These programs and their main features are presented in Table 1. The earliest was developed by George Diehr, School of Business Administration, University of Washington (Seattle) and applied to community analysis for the first time by McIntire (1973). Specifically, it attempts to determine the minimum variance partition of a set of n observations in p dimension:

$$\{ X_{1.J} \mid i = 1, 2, \ldots, n ; j = 1, 2, \ldots, p \}$$

into k clusters. The problem is to minimize the sums of squares SS around cluster means given by

$$SS = \sum_{h=1}^{k} \sum_{i \epsilon S_h} \sum_{j=1}^{P} (x_{ij} - x_{hj})^2$$

where $S_h$ is the subset of observations in cluster h, and $x_{hj}$ is the mean in the j th dimension of the observations in cluster h. The original program called MACLUS was a little modified to include different metrics such as the percentage dissimilarity (PD), also called percentage difference by Odum (1950) or percentage distance, is the complement of Bray and Curtis (1957) coefficient. The new version called OSUCL4 (André, 1981) was applied to corticolous arthropod communities by André (1983, 1985a). This program typically belongs to the multiple partitioning methods described in the previous section.

A second program, TABORD, was proposed by Maarel *et al.* (1978) for structuring phytosociological tables. It was used for the same purpose by Jensen and Maarel (1980) and Persson (1981). TABORD is essentially a clustering procedure belonging to the initial clusters methods described above. It incorporates a procedure for obtaining tables with diagonally arranged clusters and it offers 13 similarity coefficients and different devices to homogenize or combine clusters.

The programs PARTS and NUDYCB both belong to the so-called dynamic clusters methods. NUDYCB, presented in Bochi (1973), was used to cluster thecamoebic communities by Bonnet (1976) and collembolan communities by Bonnet *et al.* (1979). The original version of NUDYCB has been replaced by a new one; unfortunately, I have been unable to run it properly even with test data. The other program, PARTS, was derived from Diday's algorithm by Lebart *et al.* (1977) and applied to soil oribatid mites by Wauthy (1982).

The last program listed in Table 1 is COMPCLUS (Gauch 1979, 1980). Although it is related to the variable centered classification, it does not really belong to this group of methods as point reallocations are not al-

**Table 1. Characteristics of variable centered classification programs used in community analys.**

| NAME | TYPE (1) | DIMENSION (2) SP | RE | CL | POINT Sel. (3) | RELOCATION | COEFFICIENTS (4) | OPTIONS (5) | SOLUTION (6) | TYPE (7) |
|---|---|---|---|---|---|---|---|---|---|---|
| OSUCL4 | M | 50 | 400 | 20 | D | yes | ED,PD,1-cos$\theta$ | | U | M |
| TABORD | M | | variable | | R U | yes | 23 coeffic. | T F | M | I |
| PARTS | S | 10 | U | 25 | R | yes | ED | SF | M - U | D |
| NUDYCB | M | 60 | 100 | 30 | R U | yes | ED, $\chi^2$ | SF | M - U | D |
| COMPCLUS | M | 3000 | 5000 | 2500 | R | no | ED, PD | T | M | * |

(1) M = main program; S = subroutine

(2) SP = number of species; RE = number of relevés; CL = number of clusters (strong forms for NUDYCB); U = undetermined.

(3) Selection of initial seed points at random (R), user determined (U) or by selection of the most distant point (D)

(4) ED = Euclidian distance; PD = percentage distance; $\chi^2$ = $\chi^2$ distance

(5) Miscellaneous options: F = fusion of clusters; T = selection of a treshold value for similarity; SF = selection of strong forms sensu Diday (1971)

(6) U = unique; M = multiple

(7) M = multiple partitioning method; I = initial clusters method; D = dynamic clusters method; * = COMPCLUS altough it is related to variable centered methods does not really belong to this group of methods as point reallocations are not allowed.

lowed. Accordingly, COMPCLUS has not been tested. The three programs, OSUCL4, TABORD and PARTS which were tested correspond to the three types of algorithms which I already described.

*The data*

Using simulated data offers the advantages of exactly known properties and ease of varying these properties. Simulated data may be designed to assess the relative merits of different methods and to test their sensitivity to specific data properties (*e.g.* dimensionality, diversity, noise). The experimental approach has been frequently used to compare ordination methods (*e.g.* Austin 1976, Gauch *et al.* 1977, 1981, Gauch and Scrupps 1979, Feoli and Chiapella 1980, Gauch and Whittaker 1981, Kenkel and Orlóci 1986) and classification programs (*e.g.* Robertson 1979).

I do not want to minimize the utility of the analysis of simulated data, but I feel that simulated data supplies theoretical results which may not be convincing to ecologists. Finding clusters, well-defined from a mathematical point of view, is not necessarily tantamount to highlighting ecologically meaningful classes. In other words, the distinction between statistical efficiency and ecological meaning already pointed out by Austin (1968), Austin and Greig-Smith (1968) and Beals (1973) must not be overlooked.

For this reason, I used two sets of field data in the tests. The first set, taken from Goldstein and Grigal (1972) involves 12 plant species and 24 samples in a Tennessee watershed. Densities are expressed as relative abundances, *i.e.* sample totals are relativized to 100. This data set has been extensively studied by several authors using different ordination methods (Goldstein and Grigal 1972, Gauch 1977, Gauch *et al.* 1981) and classification algorithms (Goldstein and Grigal 1972). This sample matrix is easily interpreted and is characterized by a xeric to mesic environmental gradient (Fig. 4). All hierarchical classification programs used by Goldstein and Grigal (1972) produced identical four-group classifications (namely pine, chestnut oak, oak-hickory and yellow popolar community types, cf. Gauch 1977). In most cases, even the hierarchical structure leading to the construction of these four groups was identical (see Fig. 1 in Goldstein and Grigal 1972).

The second set of data is more complex as it involves 50 species and 72 relevés. Beta diversity is high and is estimated to 4.05 HC by using the empirical formula of Gauch and Scruggs (1979). It concerns arthropod communities living in corticolous epiphytes in southern Belgium and is taken from André (1985a). As in the first set, data are expressed as relative abundances. DCA ordination (Fig. 5B) applied to the matrix allowed the recognition of three gradients: the first axis corresponded to a gradient from crustose to foliose lichens; the second was interpreted as a juxta to a discortical environmental gradient while the third axis was related to seasons. The use of OUSCL 4 allowed the recognition of five major classes and twelve facies (Fig. 5A). Two classes are confined to special habitats or sites at certain seasons, viz. a *Pseudochermes fraxini* (Homoptera) community found on *Fraxinus* during summer and a *Vertagopus arborea* (Collembola) community observed in foliose lichens collected from one site during winter. The three other classes were directly related to the epiphytic cover, viz. a *Dometorina plantivaga* (Oribatida) community found in crustose epiphytes, an *Eueremaeus oblongus / Trichoribates trimaculatus* (Oribatida) community sheltered by foliose lichens and an *Entomobrya nivalis* (Collembola) / *Cerobasis guestfalicus* (Psocoptera) community observed in fruticose lichens. This set of data is interesting in that it supported the hypothesis that arthropod communities were related to the three major types of epiphytes. However, a point was misclassified by OSUCL4 in the sense that a sample from foliose lichens was grouped together with the nine samples collected from fruticose lichens.

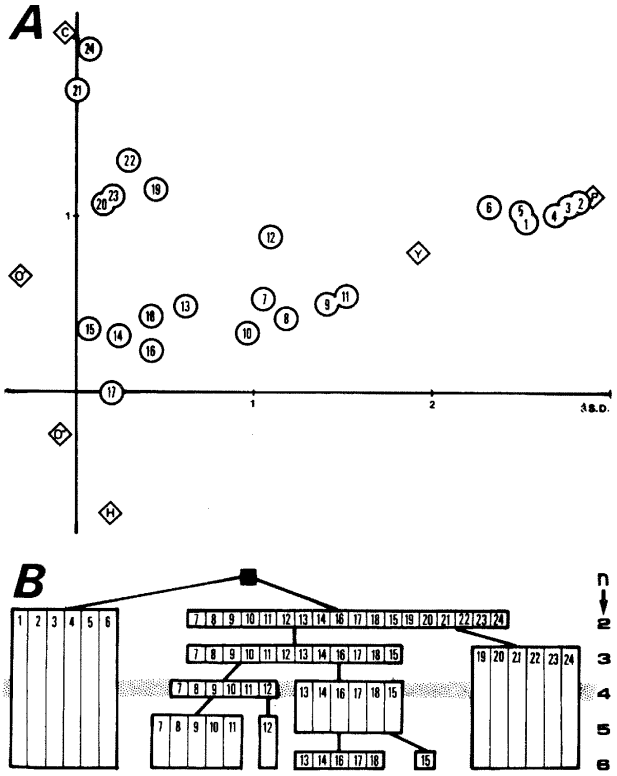


**Fig. 4. Tennessee watershed data. (A) DCA ordination. Circles represent relevés (1 to 24), diamonds represent characteristic species (C: chesnust; H: hickory; O': chesnut oak; O'': white oak; P: shortleaf pine; Y: yellow popolar). (B) Classification obtained with OSCUCL4 using PD from n = 2 to 6 partitioning levels.**

*Running programs*

Programs were run on UCL's IBM 370. Cluster analyses were performed on both data sets and all metrics

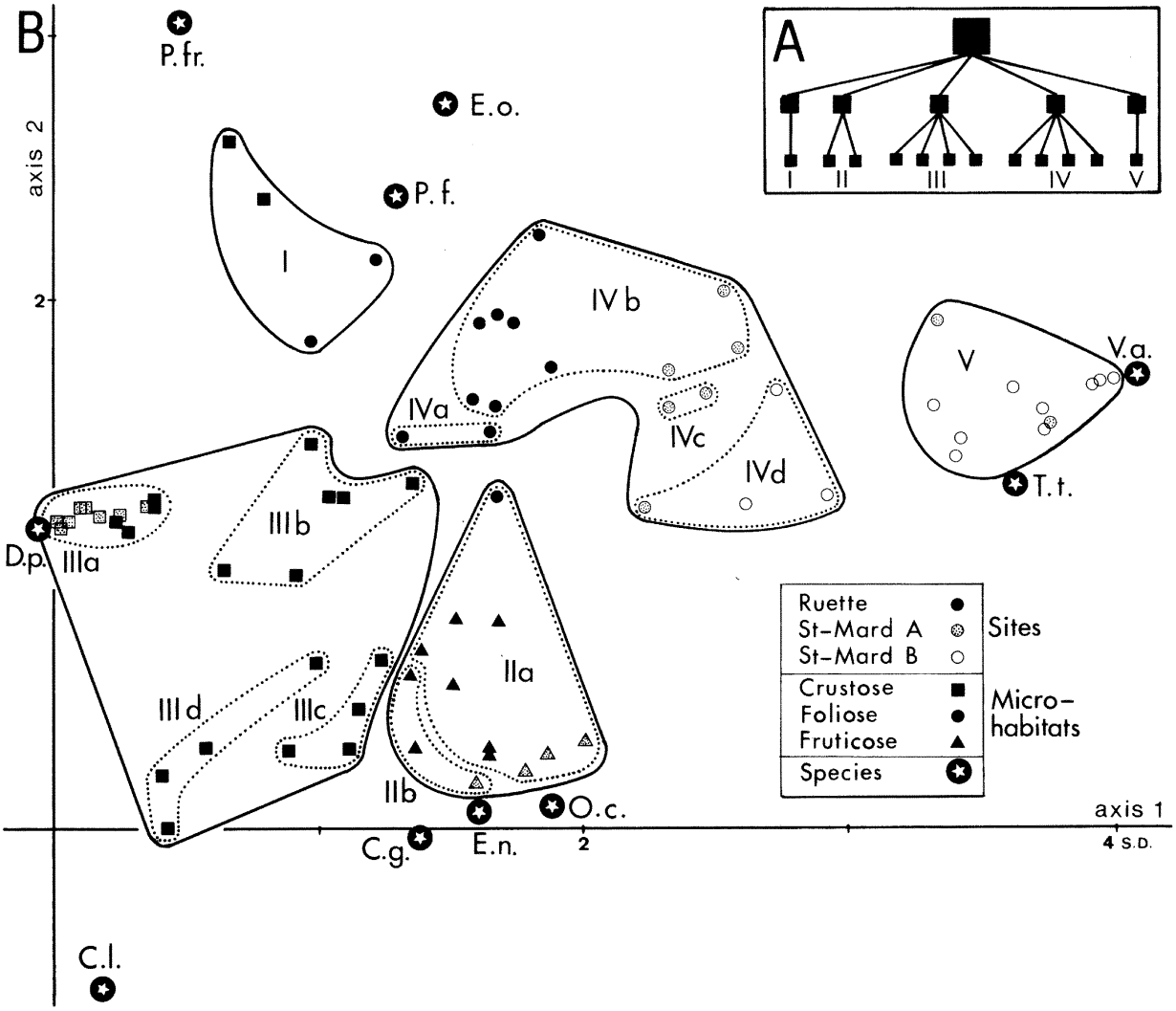


**Fig. 5. Corticolous arthropod communities. (A) Hierarchical classification into 5 classes and 12 facies obtained with OSUCL4 using PD. (B) DCA ordination where classes and facies are respectively outlined with continuous and dotted lines. Species are identified by initials while classes and facies are designated by Roman numerals with a letter. Relevés are indicated by symbols; shape and shading respectively refer to the epiphytic types and sites (see the inset) (from André, 1985a).**

available in the programs were tested. However in order to avoid an excessive number of different partitionings, only default options provided by programs were used and special devices such as the cluster fusion offered by TABORD were not tested.

*Comparing classifications*

Methods of comparing classifications are discussed by Rohlf (1974). The program used in my study is GOODM (Goldstein and Grigal 1972). GOODM is a program for calculating the degree of association between alternative classifications of the same points and relies on a measure proposed by Goodman & Kruskal (1954). The similarity coefficient ranges from zero to one. The larger the coefficient, the more similar are the two classifications being compared. However, the fundamental criterion for evalutation and comparing clustering methods remains the level of understanding gained from classifications.

## Results

*Tennessee watershed data*

Testing all three programs with all similarity coefficients involved 17 runs and gave rise to 6 different classifications (Table 2). The influence of the metrics is obvious since the same algorithm gives different classifications depending on the selected coefficient (4 classifications obtained with TABORD and 2 with OSUCL4). Conversely, the selection of the same metric used with different programs (for instance, the Euclidian distance available in all three programs) does not imply iden-

**Table 2. Triangular similarity matrix between the different classifications of the Tennessee watershed data.**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| (1) | ● | ● | ● | ✪ | ○ | ○ |
| (2) | 0.97 | ● | ● | ● | ○ | ○ |
| (3) | 0.94 | 0.97 | ● | ● | ○ | ○ |
| (4) | 0.88 | 0.91 | 0.94 | ● | ○ | ○ |
| (5) | 0.48 | 0.48 | 0.46 | 0.48 | ● | · |
| (6) | 0.20 | 0.17 | 0.17 | 0.18 | 0.00 | ● |

(1) classification obtained with most coefficients of TABORD and OSUCL4
(2) PARTS
(3) OSUCL4 used with $(1 - \cos \theta)$
(4) TABORD used with the dispersion coefficient and another non metric coefficient
(5) TABORD used with variance
(6) TABORD used with error sum

tical classifications.

OSUCL4 used with PD or ED highlights a hierarchical structure similar to that found by using hierarchical algorithms MINFO, MDISP and CLUSTER (Goldstein and Grigal, 1972). It turns out that it is nonsense to go beyond a four-class partitioning since, from a 5-class partition, classes are peeling off, *i.e.* loss of a point at each partitioning level (relevé 12 at level 5, 15 at level 6, 24 at level 7, etc). Peeling seems to be a good indicator to define a stopping rule.

My last comment concerns PARTS which was unable to identify the four classes defined by phytosociologits. Indeed, PARTS recognized only three groups of 6 points, a 5-point group and isolated relevé 15. Relevé 15 is of course the most dissimilar within its class (see Fig. 4), but this does not justify its rejection as an outlier.

*Corticolous arthropod communities*

Running all programs with the $72 \times 50$ matrix leads to 11 different classifications. A hierarchical structure is highlighted when OSUCL4 is used with PD but the structure is not dichotomous as explained previously. Five major classes related to the epiphytic cover are recognized and subdivided into 12 facies. The use of ED yields different classifications depending on the program. The similarity between classifications by TABORD and OSUCL4 is 0.84. It varies from 0.67 to 0.77 between TABORD and PARTS and from 0.79 to 0.91 between OSUCL4 and PARTS. Once again the use of different metrics leads to the construction of very different classifications between which the similarity may be as low as zero! PARTS gave as many classifications as there were runs. Partitioning into 6 classes (five real classes plus one residual) results in a major class comprising 31, 38 stat. 43 points out of 72 depending on the run and is found to be difficult to interpret and relate to habitats. The similarity between the different classifications obtained with PARTS was found to be as low as 0.87. Similarly, TABORD gives rise to a great number of classifications depending on the metric used and the random selection of initial seed points.

To investigate the ecological significance of results, relevés were also classified into 9 classes related to their origin (epiphyte, phorophyte, site). This a priori (theoretical) classification was compared to those obtained by using OSUCL4 with different coefficients. The most similar classification was obtained with PD (similarity 0.82) followed by (1-SIMI) (0.75), ED applied to standardized densities (chord distance of Orlóci 1967) (0.76) and lastly ED applied to relative abundances (0.64). Low similarity coefficients do not imply that the classifications are bad, but merely means that it is difficult to relate those classifications based on species composition to habitat.

Inspection of the details reveals that some groups of relevés are almost always identified whatever the procedure. This is the case of class I which comprises four samples taken on *Fraxinus* and dominated by *Pseudochermes fraxini*. In contrast, facies IIIc comprising only the series of four samples taken from *Lecanora conizaeoides* on *Betula* was often overlooked. This is easily explained if the composition of the four relevés are compared to the centroid, *i.e.* the average composition of the group. The mean composition is characterized by the dominance of two oribatid species, *Carabodes labyrinthicus* which represent over 40% of arthropods and *Dometorina plantivaga* which represent over 20%. Three relevés have a composition close to this mean composition. However, the winter relevé differs slightly in that the two oribatid species respectively represent only 18 and 11% of arthropods and are dominated by the collembolan *Entomobrya nivalis* which represents 21%. Although the ranges of species are similar, such a modification of the most dominant species is sufficient to induce "misclassifications". This is an illustrative example of the sensitivity of certain algorithms or coefficients to the most abundant species irrespective of the full range of species. In extreme cases, some algorithms give so much weight to dominant species that they tend to be monothetic.

## Conclusions

As stressed by Popper (1959), similarity presupposes the adoption of a point of view or interest. In other words, things are similar in different aspects and any two things which are similar from one point of view may be dissimilar from another. Accordingly, only confusion

will arise if ecologists do not agree about the adoption of such a point of view.

In the introduction, biotic communities were characterized by their species composition, precisely by the full range of their component species and not only by the dominant one. Mathematically speaking, this point of view must be formalized. Relative abudances (relevé totals relativized to 100) are thought to better express the species composition than any other data transformation. Indeed, any field ecologist will immediatly understand what means a statement such as "Species A represents 80% of individuals in community X". Furthermore, test results indicate that relative abundances used for community classification give satisfactory results, *i.e.* ecologically meaningful classes.

However, the performance of a transformation is affected by the choice of the similarity coefficient. These coefficients have been discussed in numerous publications. Recently, criteria governing the choice of a coefficient have been discussed by Legendre *et al.* (1985) while tables displaying decision-making processes have been published by Gower & Legendre (1986). In fact, much depends on our perception of similarity (Orlóci 1978) and the nature of data (Gower and Legendre 1986). Gauch (1973, 1979, 1982) encourages the use of PD, explains that the dissimilarity value of PD conforms well to the mental scaling of dissimilarities originating from ecologist's field observations, and attributes this conformity to the linear weighting of species abundances by PD which emphasizes dominants somewhat but still considers minor species. From a mathematical point of view, Gower and Legendre (1986) also encourage the use of PD for clustering (at least if double zeros, *i.e* double absences of species, are to be excluded) because this coefficient weighs equally a given difference for variables with different ranges of variation and because it is characterized by a high resolution. On the basis of the results presented herein, PD used in combination with relative abundances proves to be a good - if not the best - choice for community classification. That PD give ecologically more easily interpretable results is consistent with the conclusions of Bannister (1968), Beals (1973), and even Williams *et al.* (1966) who selected PD because of "its historical interest".

A last source of confusion is provided by the choice of a classification algorithm likely to conform to the point of view selected by ecologists. While the present analysis must be regarded as exploratory, and in some respects inconclusive, the variable centered classification algorithms, especially the multiple partitioning methods such as OSUCL4, have proved to deserve more attention. This algorithm, although it was designed as a non-hierarchical procedure, is unique in that it is likely to detect - and not to impose - a hierarchical structure in the data. Furthermore, the possible hierarchical classification is not necessarily a dichotomous dendrogram. Finally, even if there are no real stopping rules, a careful monitoring of point reallocations will allow the identification and rejection of inadequate partitions, *i.e.* partitions which do not reflect the real structure of data. In this respect, the algorithm might offer a solution to the problem of finding a correspondance between the levels in a hierarchical numerical classification and the levels of the syntaxa, a solution fundamentally different from that proposed by Feoli and Lausi (1980). Study of point reallocations also allows the identification of outliers. This is especially useful when outliers have to be deleted prior to data analysis trough ordination methods.

For very large data sets requiring preliminary classifications, initial clusters methods based on faster procedures may be preferred. TABORD, or possibly COMPCLUS (dimensioned for a $3000 \times 5000$ matrix) are convenient even if the solution provided is not unique and varies depending on the number and selection of initial seed points.

I am quite aware that the arguments propounded above and the resulting choices are somewhat arbitrary. Unavoidably, any classification will partly reflect community structure and partly reflect the thought pattern of ecologists (Whittaker 1962, Gauch 1982). Unavoidably also, clustering procedures involve a numerical process and any classification depends partly on the data and partly on the process itself (Williams 1971). The whole Art consists of choosing a numerical process that fits the ecologists thinking, and finding the procedure that meets the ecologists' requirements and concepts.

**Acknowledgements.** This paper was presented at a joint meeting of the "Société royale zoologique de Belgique" and A. Quételet Society, the Belgian section of the International Society for Biometrics held in Namur, 24 May 1986. The author wishes to thank Ph. Lebrun, under whom this work was launched and carried out, as well as S. Bochi, H.G. Gauch Jr., L. Lebart, P. Legendre, E. van der Maarel, C.D. McIntire and L. Orlóci for providing him with programs, suggestions and comments on preliminary drafts of the MS.

## REFERENCES

ANDRÉ, H. M. 1981. *OSUCL 4, un programme de groupement ("Clustering")*. U.C.L., laboratoire d'Ecologie animale, Bibliothèque de programmes ECAN, 4: 1-52.

ANDRÉ, H.M. 1983. Notes on the ecology of corticolous epiphyte dwellers. 2. Collembola. Pedobiologia 25: 271-178.

ANDRÉ, H.M. 1984. Overlapping recurrent groups: an extension of Fager's concept and algorithm. Biom. Praxim. 24: 49-65.

ANDRÉ, H.M. 1985a. Associations between the corticolous microarthropod communities and epiphytic cover on bark. Holart. Ecol. 8: 113-119.

ANDRÉ, H.M. 1985b. Arthropods overlapping reccurent groups

in corticolous epiphytes: a case study. Acta Oecologica, Oecol. gener. 6: 243-260.

ANDERBERG, M.R. 1973. *Cluster Analysis for Applications*. Academic Press, New York.

AUSTIN, M.P. 1968. An ordination study of a chalk grassland community. J. Ecol. 56: 739-757.

AUSTIN, M.P. 1976. Performance of four ordination techniques assuming three different non-linear species response models. Vegetatio 33: 43-49.

AUSTIN, M.P. and P. GREIG-SMITH. 1968. The application of quantitative methods to vegetation survey. II. Some methodological problems of data from rain forest. J. Ecol. 56: 827-844.

BANNISTER, P. 1968. An evalutation of some procedures used in simple ordinations. J. Ecol. 56: 27-34.

BARKMANN, J.J. 1970. Enige nieuwe aspecten inzake het probleem van synusiae en microgezelschappen. Mede Landbouwhogeschool Wageningen 5: 85-116.

BARKMANN, J.J. 1973. Synusial approaches to classification. *In*: Whittaker, R.H. (ed.), *Ordination and Classification of Communities*, pp. 435-491. Junk, The Hague.

BEALS, E.W. 1973. Ordination: mathematical elegance and ecological naivete. J. Ecol. 61: 23-35.

BOCHI, S. 1973. *Programme NUDYCB*. Notes I.N.R.I.A., Le Chesnay (France).

BONNET, L. 1976. Le peuplement thécamoebien édaphique de la Côte d'Ivoire. Sols de la région de Lamto. Protistologica 12: 539-554.

BONNET, L., P. CASSAGNAU, and L. DEHARVENG. 1979. Recherche d'une méthodologie dans l'analyse de la rupture des équilibres biocénotiques: applications aux Collemboles édaphiques des Pyrénées. Revue Ecol. Biol. Sol 16: 373-401.

BRAY, J.R. and J.T. CURTIS. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27: 325-349.

BURGES, A. 1960. Time and size as factors in ecology. J. Ecol. 48: 273-285.

DIDAY, E. 197. Une nouvelle méthode de classification automatique et reconnaissance des formes. La méthode des nuées dynamiques. Revue Stat. appl. 19: 19-33.

EVERITT, B. 1974. *Cluster Analysis*. Heinemann Educ. Books Ltd, London.

FEOLI, E. and L. FEOLI CHIAPELLA. 1980. Evaluation of ordination methods through simulated coenoclines: some comments. Vegetatio 42: 35-41.

FEOLI, E. and D. LAUSI. 1980. Hierarchical levels in syntaxonomy based on information functions. Vegetatio 42: 113-115.

GAUCH, H.G. 1973. A quantitative evaluation of the Bray-Curtis ordination. Ecology 54: 829-836.

GAUCH, H.G. 1977. *ORDIFLEX, a flexible computer program for four ordination techniques: weighted averages, polar ordination, principal components analysis and reciprocal averaging*. Release B. Cornell University Press, Ithaca.

GAUCH, H.G. 1979. *COMPCLUS, a FORTRAN program for rapid initial clustering of large data sets*. Cornell University Press, Ithaca.

GAUCH, H.G. 1980. Rapid initial clustering of large data sets. Vegetatio 42: 103-111.

GAUCH, H.G. 1982. *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

GAUCH, H.G. and W.M. SCRUGGS. 1979. Variants of polar ordinations. Vegetatio 40: 147-153.

GAUCH, H.G. and R.H. WHITTAKER. 1981. Hierarchical classification of community data. J. Ecol. 69: 537-557.

GAUCH, H.G., R.H. WHITTAKER, and S.B. SINGER. 1981. A comparative study of nonmetric ordinations. J. Ecol. 69: 135-152.

GAUCH, H.G., R.H. WHITTAKER, and T.R. WENTWORTH. 1977. A comparative study of reciprocal averaging and other ordination techniques. J. Ecol. 65: 157-174.

GOLDSTEIN, R.A. and D.F. GRIGAL. 1972. *Computer programs for the ordination and classification of ecosystems*. Oak Ridge National Laboratory, Oak Ridge.

GOODALL, D.W. 1986. Classification and ordination: their nature and role in taxonomy and community analysis. Coenoses 1: 3-9.

GOODMAN, L.A. and W.H. KRUSKAL. 1954. Measures of association for cross classifications. J. Amer. Statist. Ass. 49: 732-764.

GOWER, J.C. and P. LEGENDRE. 1986. Metric and Euclidian properties of dissimilarity coefficients. J. Classification 3: 5-48.

GREIG-SMITH, P. 1983. *Quantitative Plant Ecology*. 3rd Ed. Blackwell, Oxford.

KENKEL, N.C. and L. ORLÓCI. 1985. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67: 919-928.

KOESTLER, A. *Janus: a Summing Up*. Hutchinson & Co. London.

JANCEY, R.C. 1966. Multidimensional group analysis. Aust. J. Bot. 14: 127-130.

JANCEY, R.C. 1974. Algorithm for the detection of discontinuities in data sets. Vegetatio 29: 131-133.

JENSEN, S. and E. van der MAAREL. 1980. Numerical approaches to lake classification with special reference to macrophyte communities. Vegetatio 42: 117-128.

LAMBERT, J.M. and W.T. WILLIAMS. 1966. Multivariate methods in plant ecology. VI. Comparison of information-analysis and association-analysis. J. Ecol. 54: 635-664.

LEBART, L., A. MORINEAU, and N. TABARD. 1977. *Techniques de la description statistique. Méthodes et logiciels pour l'analyse des grands tableaux*. Dunod, Paris.

LEGENDRE, L. and P. LEGENDRE. 1983. *Numerical Ecology*. Elsevier Scient. Publ. Co., Amsterdam.

LEGENDRE, P., S. DALLOT, and L. LEGENDRE. 1985. Succession of species within a community: chronological clustering with applications to marine and freshwater zooplankton. Amer. Natur. 125: 257-288.

MAAREL, E. van der, J.G.M. JANSSEN, and J.M.W. LOUPPEN. 1978. TABORD, a program for structuring phytosociological tables. Vegetatio 38: 143-156.

MCINTIRE, C.D. 1973. Diatom association in Yaquina estuary, Oregon: a multivariate analysis. J. Phycol. 9: 254-259.

MACQUEEN, J.B. 1967. Some methods for classification and analysis of multivariate observations. *In*: M.L. Le Cam and J. Neyman (Eds.), *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability (1)*, pp. 281-297. University of California Press, Berkeley.

ODUM, E.P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology 31: 587-605.

ORLÓCI, L. 1967. An agglomerative method for classification of ploant communities. J. Ecol. 55: 193-205.

ORLÓCI, L. 1978. *Multivariate Analysis in Vegetation Research*.

2nd ed. Yunk, The hague.

ORLÓCI, L. and S.L. STOFELLA. 1986. A taxon-free numerical approach to the study of plant communities. Annals of Arid Zone 25: 111-131.

PERSSON, S. 1981. Ecological indicator values as an aid in the interpretation of ordination diagrams. J. Ecol. 69: 71-84.

PIAGET, J. 1967. *Biologie et connaissance*. Gallimard, Paris.

PIELOU, E.C. 1977. *Mathematical Ecology*. J. Wiley & Sons, New York.

PIELOU, E.C. 1984. *The Interpretation of Ecological data. A Primer on Classification and Ordination*. Wiley & Sons, New York.

POPPER, K.R. 1959. *The Logic of Scientific Discovery*. Hutchinson & Co. London.

ROBERTSON, P.A. 1979. Comparisons among three hierarchical classification techniques using simulated coenoplanes. Vegetatio 40: 175-183.

ROHLF, F.J. 1974. Methods of comparing classifications. Ann. Rev. Ecol. Syst. 5: 101-113.

SHIMWELL, D.W. 1971. *The Description and Classification of Vegetation*. Sidgwick & Jackson, London.

SIMON, H.A. 1962. The architecture of complexity. Proc. Amer. Phil. Soc. 106: 467-482.

WAUTHY, G. 1982. Synecology of forest soil oribatid mites of Belgium. 3. Ecological groups. Acta Oecologica, Oecol. Gener. 3: 469-494.

WESTHOFF, W. and E. van der MAAREL. 1973. The Braun-Blanquet approach. *In*: Whittaker, R.H. (ed.), *Ordination and Classification of Communities*, pp. 617-726. Yunk, The Hague.

WHITTAKER, R.H. 1962. Classification of natural communities. Biol. Rev. 28: 1-239.

WHITTAKER, R.H. 1973. Approaches to classifying vegetation. In: Whittaker R.H. (ed.). *Ordination and Classification of Communities*, pp. 323-354. Junk, The Hague.

WILLIAMS, W.T. 1971. Principles of clustering. Ann. Rev. Ecol. Syst. 2: 303-326.

WILLIAMS, J.T., J.M. LAMBERT, and G.N. LANCE. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. J. Ecol. 54: 427-445.