Acta Botanica Academiae Scientiarum Hungaricae, Tomus 25 (1—2), pp. 125—130 (1979)

ASSOCIATION-ANALYSIS BASED ON THE USE OF MUTUAL INFORMATION

By

J. PODANI

RESEARCH INSTITUTE FOR BOTANY, HUNGARIAN ACADEMY OF SCIENCES, VÁCRÁTÓT

(Received January 10, 1979)

A new version of association-analysis is presented. The division parameter is defined by means of information theory, the divisive species is that which has the maximum sum of mutual information with all the other species. Application of the method is illustrated by an example. The resulting hierarchy is compared with a previously published one obtained with an agglomerative method and with the result of principal component analysis. Joint application of association-analysis with several different numerical methods is proposed.

I. Introduction

I.1. *Association-analysis* (*AA*) has been one of the most widely used classification methods of plant ecology in the last two decades. The first version of *AA* was the technique of GOODALL (1953) modified by WILLIAMS and LAMBERT (1959, 1960), whose method may be regarded as the most typical monothetic divisive strategy. The set of vegetation samples is subdivided in each cycle of the analysis according to the presence and absence of a single species (key- or divisive species). Main problems of the different *AA* algorithms are the criterion for selection of the divisive species and the definition of a measure for stopping further subdivisions ("stopping rule"). The division parameter of WILLIAMS and LAMBERT was defined by

$$\max \left\{ \sum_{j \neq k} \chi^2_{jk} \right\} \qquad \text{(I.1.1.)}$$

i.e. the divisive species was that which had the maximum sum of χ^2 values calculated with all the other species. A subset of samples were considered homogeneous and its division was abandoned if the maximum value of χ^2 between any two species was less than a fixed value. Other but similar definitions for division parameter and stopping rule are listed by LANCE and WILLIAMS (1965).

I.2. Since the use of χ^2 has many disadvantages, LANCE and WILLIAMS (1968) modified their procedure on the basis of information theory. Following SHANNON (1948), they defined the "information content" of the set E of m entities (e.g. vegetation samples) described by n binary attributes (e.g. species) as

$$\hat{I}(E) = nm \log m - \sum_i (k_i \log k_i + (m - k_i) \log (m - k_i)) \qquad \text{(I.2.1.)}$$

where k_i is the number of entities possessing the ith attribute. The set of entities is divided into two subsets E_1 and E_2 according to the presence and absence of each attribute, the reduction of information content ("information fall") is calculated in each case using the relation

$$\Delta\hat{I}(E_1, E_2, E) = \hat{I}(E) - \hat{I}(E_1) - \hat{I}(E_2) \qquad \text{(I.2.2.)}$$

and that division is performed for which $\Delta\hat{I}$ is maximum. The authors suggested that $2\Delta\hat{I}$ could be used as stopping rule instead of χ^2, but this definition may be strongly criticized (cf. CORMACK 1971).

I.3. More recently numerous authors cast doubt on the ecological reliability of groupings obtained with *AA*. The major criticism is that the resulting hierarchy of vegetation samples is not necessarily an optimal classification because of the monothetic and divisive properties of the algorithm (COETZEE and WERGER 1975). It is easy to see, however, that there is no perfect numerical method which supersede all the others in every respect. Consequently, we do not know a priori that method whose result can be regarded as an "ecologically optimal" classification. Thus absolute disregard of *AA* would be an unwarranted and thoughtless step in spite of its obvious disadvantages. Joint application of *AA* with several different agglomerative methods and ordination procedures seems to be a more reasonable way in vegetation survey.

I.4. In the present paper I give a version of *AA* based on the mutual information measures between attributes. A new definition of division parameter is presented but the question of stopping rule will remain opened. The method, with respect to its result, is probably closely related to that of LANCE and WILLIAMS (1968) and comparison of them requires further investigations.

II. A new version of association-analysis

II.1. As mentioned above the central question of *AA* is the way of selecting the divisive species. If information theoretical definitions are used we obtain a self-explanatory variant of division parameter. In each cycle of the analysis we obviously have to search for that species which contains the highest information about the remaining species. This one is that which has the maximum sum of mutual information values calculated with the others. The mutual information between species s_j and s_k can be calculated using the parameters of the well-known 2×2 contingency table as

$$\begin{aligned} m\hat{I}(s_j; s_k) = {} & m \log m + a \log a + b \log b + c \log c + d \log d - \\ & - (a + c) \log (a + c) - (b + d) \log (b + d) - \\ & - (a + b) \log (a + b) - (c + d) \log (c + d) \end{aligned} \qquad \text{(II.1.1.)}$$

where $m = a + b + c + d$ (cf. KULLBACK 1959). Then the division parameter is

$$\max \left\{ \sum_{j \neq k} m\hat{I}(s_j; s_k) \right\} \qquad \text{(II.1.2.)}$$

The set or subset of samples is subdivided according to the selected species. The analysis, with a few exceptions, can be carried out as far as that every cluster contains only one sample. If the maximum sum of mutual information values appears at more than one species we have to examine their distributions. If these are same or symmetric we can regard every species in question as divisive one. Supposing that the distributions are different and asymmetric the subset can not be subdivided further (automatic stopping rule). It is advisable, however, to stop the analysis at an arbitrary group size since the relationships at lower hierarchic levels are usually out of interest.

Since the measure II.1.2. decreases monotonically, reversals are entirely absent from the dendrogram.

II.2. Computer program ASSINF has been written in FORTRAN IV language for the analysis of vegetation samples. The program accepts binary data, prints out the results of each subdivision and gives the dendrogram on the printer. This program is available from the author on request.

III. Application of *AA* and its comparison with other numerical methods

III.1. The new version of *AA* was applied to the study of a rocky grassland community named *Seslerietum sadlerianae* from Hungary. 33 sampling units of 4×4 m size from 7 stands were involved in the analysis

stand	quadrat No.
I. Sashegy, Buda Hills	1—5.
II. Hunyadorom, Buda Hills	6—9.
III. Tündérszikla, Buda Hills	10—14.
IV. Hármashatárhegy, Buda Hills	15—18.
V. Nagykevély, Pilis Hills	19—23.
VI. Pilistető, Pilis Hills	24—28.
VII. Naszály Hill	29—33.

Matrix of the presence — absence data and a map illustrating the geographic localization of the stands can be found in a previous publication (PODANI, 1978). Dendrogram of the quadrats is shown in Fig. 1.

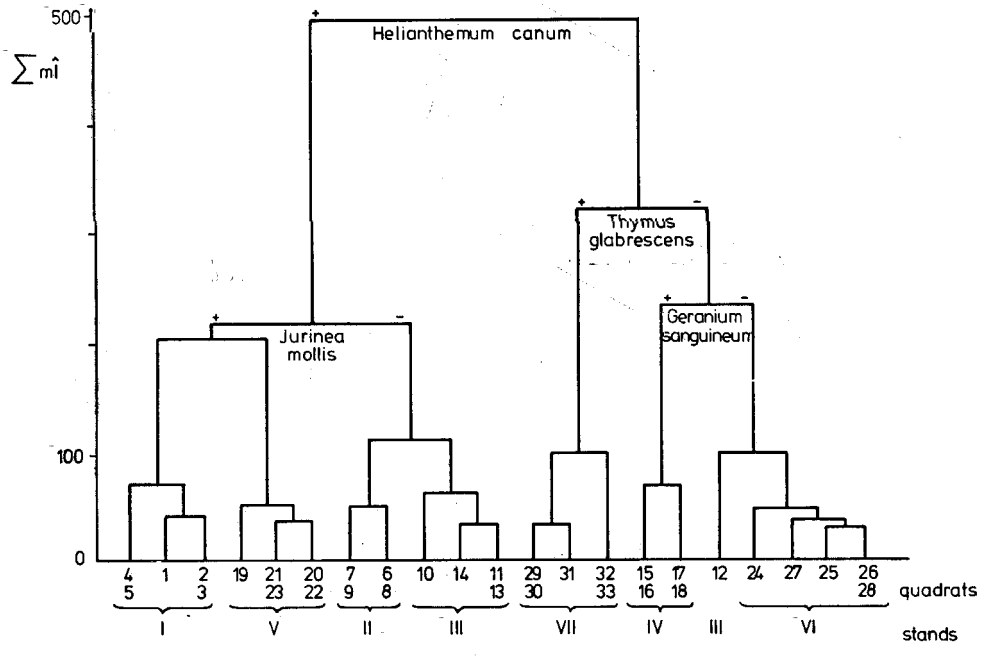

Fig. 1. Hierarchy for quadrats obtained with the new version of association-analysis

The quadrats listed above were formerly subjected to an agglomerative classification (centroid sorting algorithm using the weighted dissimilarity index, see PODANI 1978 for detailed description of the method and for the resulting dendrogram).

PEARSON product-moment correlations between quadrats were also computed. The correlation matrix was analyzed by principal component analysis (see DIXON 1964). For our purposes it is sufficient to examine only the first two components which account for a total of 46% of the variance in the data. The component scores are plotted on the scatter diagram of Fig. 2.

III.2. We can see at a glance on Fig. 1. that the quadrats are clustered according to the stands with the only exception of quadrat 12. First it seems to be a typical case of "misclassification" caused by the absence of Helianthemum canum from quadrat 12 by chance. This quadrat, however, has got close to stands IV and VI in the hierarchy obtained with the centroid sorting method as well so that the possibility of misclassification is surely out of question. This is strengthened by the fact that according to the first two components quadrat 12 is nearest to these stands among the quadrats of the rest.

There are some other remarkable agreements between the results. For example, stands II—III and I—V, respectively, are closely related in every case. Good agreement between *AA* and *PCA* can be illustrated by indicating the first four divisions on the scatter diagram (Fig. 2).

The only essential difference between the classifications appears at the highest hierarchic level. *AA* produced two main groups of almost equal size according to the presence and absence

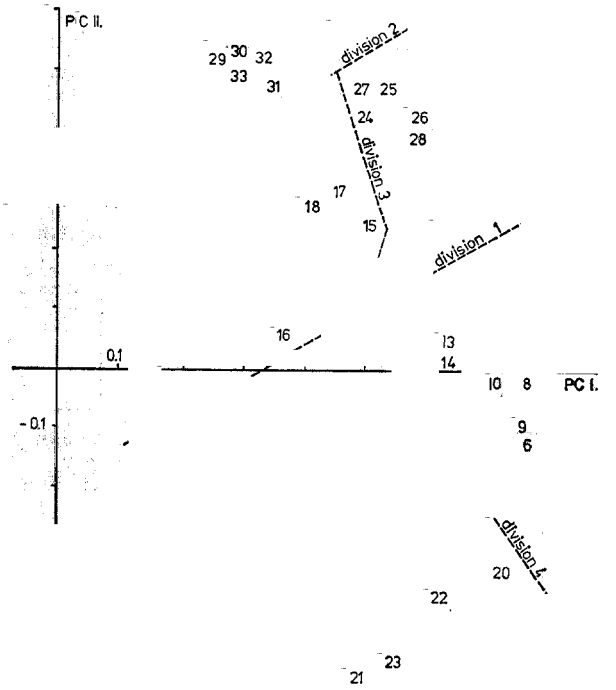

Fig. 2. Principal components ordination of quadrats. Dotted lines indicate the subsequent divisions by association-analysis

of *Helianthemum canum*. In the hierarchy obtained with centroid sorting method stands IV and VI (with quadrat 12) are close to each other as well but they are more similar to the group of stands I, II, III and V than to the stand of Naszály.

III.3. Finally, we can draw the following conclusions from the results:

(i) There are two extreme types within the community studied. These are represented by the stand of Naszály and, on the other hand, by the stands of Sashegy, Hunyadorom, Tündérszikla and Nagykevély. The remaining stands can be considered as transitions between the extreme types.

(ii) According to the classifications and ordination, we have the following order of the stands:

I and V—II—III—IV—VI—VII

(iii) Results can be explained satisfactorily on ecological, floristical and geographical basis. Stand VII is the only one which is grown on limestone, all the others are on dolomite rocks. This stand is nearest to the Northern Range of medium height therefore it contains 17 species which are absent from the other stands. Separation of IV and VI from the rest may be explained by the absence of numerous frequent species (e.g. *Festuca pallens*, *Seseli leucospermum*, *Helianthemum canum*, *Silene otites*, *Linum tenuifolium*, *Fumana procumbens*, *Scorzonera austriaca*, etc.) characterizing the vegetation of dolomite rocks in general.

(iv) Different numerical methods gave rather similar and, for this reason, well-interpretable results. Shortcomings of each method can be reduced by the joint application of them and the comparative evaluation of the results. Cases of misclassification by *AA*, for example, should be treated with caution.

ACKNOWLEDGEMENTS

I am grateful to P. Juhász-Nagy (Eötvös Loránd University, Budapest) and to Z. Szőcs (Res. Inst. for Botany, Hungarian Acad. of Sci., Vácrátót) for their helpful suggestions.

REFERENCES

Coetzee, B. J.—Werger, M. J. A. (1975): On association-analysis and the classification of plant communities. Vegetatio **30**, 201—206.

Cormack, R. M. (1971): A review of classification. J. Roy. Stat. Soc. series A. **134**, 321—367.

Dixon, W. J. (1964): BMD computer programs manual. University of California at Los Angeles.

Goodall, D. W. (1953): Objective methods for the classification of vegetation I. The use of positive interspecific correlation. Aust. J. Bot. **1**, 39—63.

Kullback, S. (1959): Information Theory and Statistics. Wiley, New York.

LANCE, G. N.—WILLIAMS, W. T. (1965): Computer programs for monothetic classification ("association-analysis"). Computer J. **8**, 246—249.

LANCE, G. N.—WILLIAMS, W. T. (1968): Note on a new information-statistic classificatory program. Computer J. **11**, 195.

PODANI, J. (1978): A method for clustering of binary (floristical) data in vegetation research. Acta Bot. Acad. Sci. Hung. **24**, 121—137.

SHANNON, C. E. (1948): A mathematical theory of communication. Bell System Tech. J. **27**, 379—423, 623—656.

WILLIAMS, W. T.—LAMBERT, J. M. (1959): Multivariate methods in plant ecology I. Association-analysis in plant communities. J. Ecol. **47**, 83—101.

WILLIAMS, W. T.—LAMBERT, J. M. (1960): Multivariate methods in plant ecology II. The use of an electronic digital computer for association-analysis. J. Ecol. **48**, 689—710.