DATA ANALYSIS AND INFORMATICS, IV
E. Diday et al. (editors)

INTERPRETING A HIERARCHICAL CLASSIFICATION WITH SIMPLE DISCRIMINANT FUNCTIONS: AN ECOLOGICAL EXAMPLE

Cajo J.F. ter Braak,
TNO Institute of Mathematics, Information Processing and Statistics
P.O.Box 100, 6700 AC Wageningen
The Netherlands

This paper proposes a method for relating a hierarchical classification to external information. The method makes pairwise comparisons between the branches at each node of the hierarchy. These comparisons are likely to show major differences between branches high up in the hierarchy and more subtle differences between adjacent clusters that are lower down in the hierarchy. This idea has been implemented in a FORTRAN-program called DISCRIM. Used in combination with Hill's (1979) cluster program TWINSPAN, DISCRIM forms a simple tool to explore the relationship between a set of response variables and a set of explanatory variables in heterogeneous data sets. In the ecological example provided, bird communities in Dutch heathlands are related to heathland characteristics.

INTRODUCTION

To interpret results of cluster analysis one frequently needs to relate the group structure to external information, for example by calculating mean values of extrinsic variables and using analysis of variance to test the differences between means, or alternatively by reference to the first few axes of a discriminant analysis. Such methods compare all clusters simultaneously and in practice often show only the more obvious differences among clusters. More interesting differences may reside in specific pairwise comparisons among clusters at various levels. It is impractical to make all possible pairwise comparisons. But for datasets in which hierarchical classification is judged appropriate, it seems reasonable to restrict attention to pairwise comparisons of branches at each node of the hierarchy. Major differences are expected to exist between branches high up in the hierarchy and more subtle differences between adjacent clusters that are lower down in the hierarchy.

In this paper these ideas are applied to a commonly occurring problem in community ecology, namely that of relating species composition data to quantitative or qualitative environmental variables. This problem is the ecological version of the general problem of relating response variables to predictor variables and our approach to the problem may find application in other disciplines as well. As an example we shall relate the densities of bird species in Dutch heathlands to geomorphological characteristics of the heathlands. In general terms, a typical data set in community ecology consists of information on the occurrence or abundance of a set of species, and on a set of environmental variables, at a series of 'sites' separated in space or in time. Hereby each site forms a sampling unit, each species a response variable and each environmental variable a predictor variable. Commonly two-way indicator species analysis (TWINSPAN) is then used to obtain a hierarchical classification of the sites on the basis of the species presences or abundances at the sites (Hill et al. 1975; Gauch and

Whittaker 1981). TWINSPAN is a divisive, polythetic cluster method in which successive divisions are obtained by successive correspondence analyses of the sampling units: at each division the first correspondence analysis axis is split at its centroid, then the resulting division is refined by iterative character weighting. The properties of TWINSPAN therefore largely follow from the optimality properties of correspondence analysis, notably, those concerning the discovery of both diagonal and block structure in two-way tables (Benzécri et al. 1974, IIA no. 2 §3.2; Hill 1974). The TWINSPAN program of Hill (1979) has three further attractive features that facilitate the interpretation of the clusters:

(1) Each division is succinctly characterized by a limited set of indicator species, using a simple discriminant function, the coefficients of which take the values -1, 0, +1.

(2) Species are classified in the same way as the sites; however, not on the basis of the original abundance at each site, but on the basis of the average abundance at each node of the site classification.

(3) The site classification and species classification are each converted into an ordering and the original two-way table of species by sites is rearranged according to these orderings.

The TWINSPAN program was modified to allow characterization of a supplied hierarchical classification in terms of external variables (in community ecology, the environmental variables). The new program, DISCRIM (Ter Braak 1982), has proved useful in a variety of ecological applications (e.g. Kalkhoven and Opdam 1984). This paper describes the theory behind the combined use of TWINSPAN and DISCRIM and its application.

SIMPLE DISCRIMINANT FUNCTIONS

A method for comparing the branches of each node of the hierarchy should provide a succinct characterization of the differences between the branches. With quantitative external variables, linear discriminant analysis could be used. For nominal variables, correspondence analysis could be applied to a $2 \times c$ table where the rows correspond to the two branches of the node and c is the total number of categories of the nominal variables (cf. Jambu 1978, p. 83). However, general linear discriminant functions are not easy to assimilate. Moreover, lower down in the hierarchy the branches may contain so few sites that the coefficients of the discriminant function cannot be estimated accurately, if at all. Using presence-absence data, Hill (1977) got round these difficulties by proposing to use simple discriminant functions, the coefficients of which can take only three values: -1 and +1 for attributes that are characteristic for the one and the other branch, respectively, and 0 for non-discriminating attributes. Such functions are easier to interpret. Both quantitative and nominal variables can be accomodated into this scheme after recoding (see next Section).

It is convenient at this point to call the left-hand and right-hand branches of a node by the negative and positive group, respectively (see Fig. 2) and to term an attribute a negative indicator if the attribute is characteristic for the negative group and a positive indicator if it is characteristic for the positive group. The discriminant score of a site is then found by simply adding +1 for each positive indicator and -1 for each negative indicator that it contains. Sites with a score less than or equal to a certain threshold value are assigned to the negative group and sites with scores greater than this value are assigned to the positive group. The threshold value should be chosen so as to make this division agree as far as possible with the original grouping in order to minimize the number of misclassifications.

In TWINSPAN and DISCRIM the simple discriminant functions are constructed in a very simple way (Hill 1979). An attribute is a possible positive indicator if its frequency of occurrence is higher in the positive group than in the negative group. Analogously, possible negative indicators are defined. The n attributes with highest absolute difference in frequency of occurrence are included in the discriminant function, where n is the smallest integer that minimizes the number of misclassifications. (In practice, an upperbound is imposed on n ($n < 7$) and attributes that occur with about the same frequency in both groups cannot be considered as possible indicators.) The great advantage of the discriminant functions in TWINSPAN and DISCRIM is their simplicity; the sign of an attribute is taken with the same sign as that of the frequency difference, and the number of possible sets of indicator attributes is restricted by ordering the attributes on the basis of the absolute frequency difference. These restrictions avoid the need for optimization by integer programming and are likely to facilitate the interpretation of the discriminant functions so constructed.

CODING OF NOMINAL AND QUANTITATIVE VARIABLES

To analyse nominal and quantitative variables by simple discriminant functions, these variables must be recoded. A straightforward way is to define dummy variables, one for each category of each nominal variable. The dummy variable for a category gets the value 1 if the site scores on that category and the value 0 otherwise. After discretization, quantitative variables can be coded in the same way. The program DISCRIM does not generate these dummy variables automatically; hence these must be supplied as data input. Heiser (1981, p. 124) terms this coding scheme 'disjoint coding'.

Hill et al. (1975) proposed a different coding scheme, which is called 'the method of pseudo-species' (see also Hill 1977) or 'conjoint coding' (Heiser 1981, p. 123). This method originated in the context of abundances of species in sites, e.g. areal cover of plant species in quadrats. The information on cover can be represented on a crude scale by binary variables such as 'Is the species present?', 'Is the species present with cover greater than 5%?', 'Is the species present with cover greater than 10%?', etc. These binary variables are termed pseudo-species that are defined in the example by the pseudo-species cut levels 0, 5 and 10. This method is most suited for non-negative quantitative variables that can be absent, i.e. where the value 0 has a special meaning, and also for variables for which the numerical coding cannot be reversed because of asymmetry in the meaning of low values and high values. In DISCRIM this method of coding is available and can be used for variables such as 'Is there clay at the site and if so, what is the surface area of clay?'. The quantitative variable clay is then replaced by the pseudo-attributes 'Is clay present?', 'Is clay present with areal fraction greater than 5%?', 'Is clay present with areal fraction greater than 10%?'. Only one set of cut levels can be supplied to DISCRIM and this set is used for all variables. Therefore some prior transformation of the data may be needed if the units of measurement of the variables differ. A possible transformation is to rank numbers so that the cut levels determine percentiles of the distribution of each variable. To define quartiles four cut levels are needed: the 0-, 25-, 50- and 75-percent point of the ranked data. Quartiles may be sufficient for many applications.

Simple discriminant functions are constructed according to the additional rule that not more than one pseudo-attribute of each quantitative variable may be used in the discriminant function. Selection of a pseudo-attribute as a negative indicator means that values of the variable higher than the corresponding cut level occur more frequently in the negative group than in the positive group.

For discretized quantitative variables fuzzy coding (French: codage flou) can have advantages over disjoint coding. The dummy variable for a category then gets, for example, the value 2 if the site scores on that category, the value 1 if the site scores on one of the adjacent categories, and the value 0 otherwise. Fuzzy coding expresses in this way the similarity of adjacent categories. In TWINSPAN and DISCRIM, fuzzy coding must be used in conjunction with conjoint coding, because these programs act on binary variables.

CLASSIFICATION OF ATTRIBUTES

TWINSPAN first classifies the sites on the basis of species occurrences (attributes). Secondly, species are classified in a hierarchical way on the basis of average values at each node of the site classification. In this way species are clustered that have a 'similar' distribution across the clusters of sites. The actual classification is derived in the same way as the classification of the sites, namely by successive correspondence analyses and iterative character weighting. (The dissimilarity measure that is implicit in this method is thus the chi-square distance applied to node averages.) This method is also useful for relating the site classification to external information and is therefore adopted in DISCRIM. Consequently,

TABLE 1: Species used in TWINSPAN to classify the heathlands

No.	Abbreviation		Latin name	Name
1	ALAU	ARVE	Alauda arvensis	Sky Lark
2	ANTH	TRIV	Anthus trivialis	Tree Pipit
3	EMBE	CITR	Emberiza citrinella	Yellowhammer
4	NUME	ARQU	Numenius arquata	Curlew
5	SYLV	COMM	Silvia communis	Whitethroat
6	CARD	CANN	Carduelis cannabina	Linnet
7	CUCU	CANO	Cuculus canorus	Cuckoo
8	ANTH	PRAT	Anthus pratensis	Meadow Pipit
9	VANE	VANE	Vanellus vanellus	Lapwing
10	ERIT	RUBE	Erithacus rubecula	Robin
11	LULL	ARBO	Lullula arborea	Wood lark
12	ANTH	CAMP	Anthus campestris	Tawny Pipit
13	PICU	VIRI	Picus viridis	Green Woodpecker
14	PHYL	TROC	Phylloscopus trochilus	Willow Warbler
15	LYRU	TETR	Lyrurus tetrix	Black Grouse
16	FALC	TINN	Falco tinnunculus	Kestrel
17	FALC	SUBB	Falco subbuteo	Hobby
18	SAXI	RUBE	Saxicola rubetra	Whinchat
19	PERD	PERD	Perdix perdix	Partridge
20	TRIN	TOTA	Tringa totanus	Redshank
21	GALL	GALL	Gallinago gallinago	Snipe
22	HAEM	OSTR	Haematopus ostralegus	Oystercatcher
23	TADO	TADO	Tadorna tadorna	Shelduck
24	OENA	OENA	Oenanthe oenanthe	Wheatear
25	LOCU	NAEV	Locustella naevia	Grasshopper Warbler
26	LIMO	LIMO	Limosa limosa	Black-tailed Godwit
27	MOTA	FLAV	Motacilla flava flava	Blue-headed Wagtail
28	MOTA	ALBA	Motacilla alba	White Wagtail
29	CYAN	SVEC	Cyanosylvia svecica	Bluethroat
30	CAPR	EURO	Caprimulgus europaeus	Nightjar
31	EMBE	SCHO	Emberiza schoeniclus	Reed Bunting
32	LANI	EXCU	Lanius excubitor	Great Grey Shrike
33	GALE	CRIS	Galerida cristata	Crested Lark
34	CIRC	CYAN	Circus cyanus	Hen Harrier

attributes are considered similar or dissimilar according to whether they occur in the same site groups. The classification of attributes tends to reveal the sets of environmental conditions that prevail in each group of sites.

ECOLOGICAL EXAMPLE

Opdam and Retel Helmrich (1984) described the bird communities of Dutch heathlands, and related them to 24 heathland characteristics that included area, recreational usage, isolation, landscape, geographical position, topography, soil-type and soil-heterogeneity. They sampled 82 heathlands, some of which included patches of woods and agricultural fields. The bird census data gave abundances (number of territories per 100 hectare) of 34 species that supposedly make use of heathland in some way (Table 1).

The first step in the analysis was to use TWINSPAN to produce a hierarchical classification of the heathlands based on the bird data. Fig. 1 shows the resulting two-way table and Fig. 2 the indicator species characterizing the first two levels of division. The indicator species describe the divisions

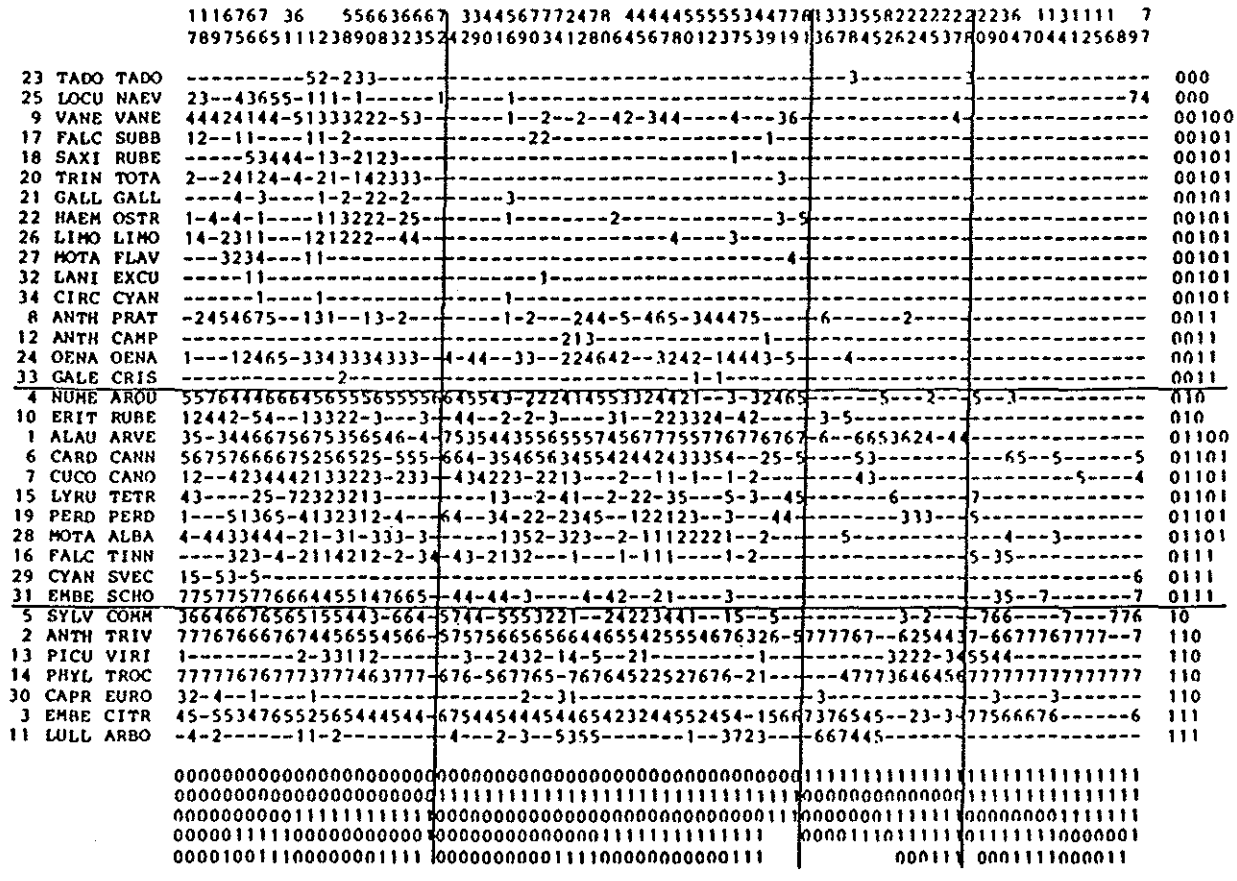

Fig.1: TWINSPAN two-way table of bird species (rows) of Dutch heathlands (columns). Values are logarithmic classes of abundance (number of pairs per 10 hectares). (-: absent; 1: < 0.5; 2: 0.5 - 0.9; 3: 1.0 - 1.9; 4: 2.0 - 3.9; 5: 4.0 - 7.9; 6: 8.0 - 15.9; 7: > 16.0). The top margin gives site identification numbers, printed vertically. The bottom and right-hand margins show the hierarchical classifications of the heathlands and birds, respectively, each with five levels of division. Vertical lines separate groups of sites at level 2; horizontal lines separate the first two species divisions. See Table 1 for the abbreviation of species names.

fairly accurately as can be seen from the number of heathlands that are misclassified by the discriminant function of indicator species (Fig. 2). For example, in the first division only one sample of the negative group is misclassified and none of the positive group.

It can be seen from Fig. 1 that the first division is between species-rich and species-poor heathlands. The species-rich group of 53 heathland had many indicators, but none were found for the species-poor group (Fig. 2). These two groups were further divided, the species-rich group into a species-rich group and a less species-rich group, the species-poor group into heathlands with Sky Lark and heathlands with a high density of Willow Warbler (Fig. 2). In Fig. 1 these groups of heathlands are divided further.

TABLE 2: Heathland characteristics used in DISCRIM to interpret clusters of heathlands.

No.	Abbreviation	Description	Values
AREA			
1	AREA < 20	Area of heath smaller than 20 hectares	0/1
2	AR20 - 100	Area of heath between 20 and 100 hectares	0/1
3	AREA > 100	Area of heath greater than 100 hectares	0/1
RECREATIONAL USAGE			
4	RECR MILI	Index of recreational use of heath, including military usage, based on inquiry	ranked 1-82
24	RECR EATI	Index of recreational use of heath, excluding military usage, based on inquiry	ranked 1-82
ISOLATION			
5	HEAT < 5 KM	Number of other heaths within a radius of 5 km from the border of the heath	ranked 1-82
LANDSCAPE			
6	OPEN SAND	Presence of open sand within the heath	0/1
7	MOOR POOL	Presence of moorland pools within the heath	0/1
8	WET	Presence of wet patches within the heath	0/1
9	SURR FORE	Heath at least partly surrounded by woodland	0/1
10	SURR AGRI	Heath at least partly surrounded by grassland or arable land	0/1
GEOGRAPHICAL POSITION			
11	VELU WE	Heath lies on the VELUWE	0/1
12	BRAB ANT	Heath lies in BRABANT	0/1
13	DREN THE	Heath lies in DRENTHE	0/1
14	GRON INGE	Heath lies in GRONINGEN	0/1
15	GOOI	Heath lies in 'het GOOI'	0/1
16	LIMB URG	Heath lies in LIMBURG	0/1
TOPOGRAPHY, SOIL, AND SOIL HETEROGENEITY (based on soil maps)			
17	UNDU LATI	Heath is undulating	0/1
18	FEN LAND	Presence of fen-land	0/1
19	SAND SOIL	Presence of sandy soil	0/1
20	SAND FEN	Presence of sandy soil in fen-land	0/1
21	1SOI LTYP	Presence of only one soil type	0/1
22	2SOI LTYP	Presence of two soil types	0/1
23	3SOI LTYP	Presence of three or more soil types	0/1

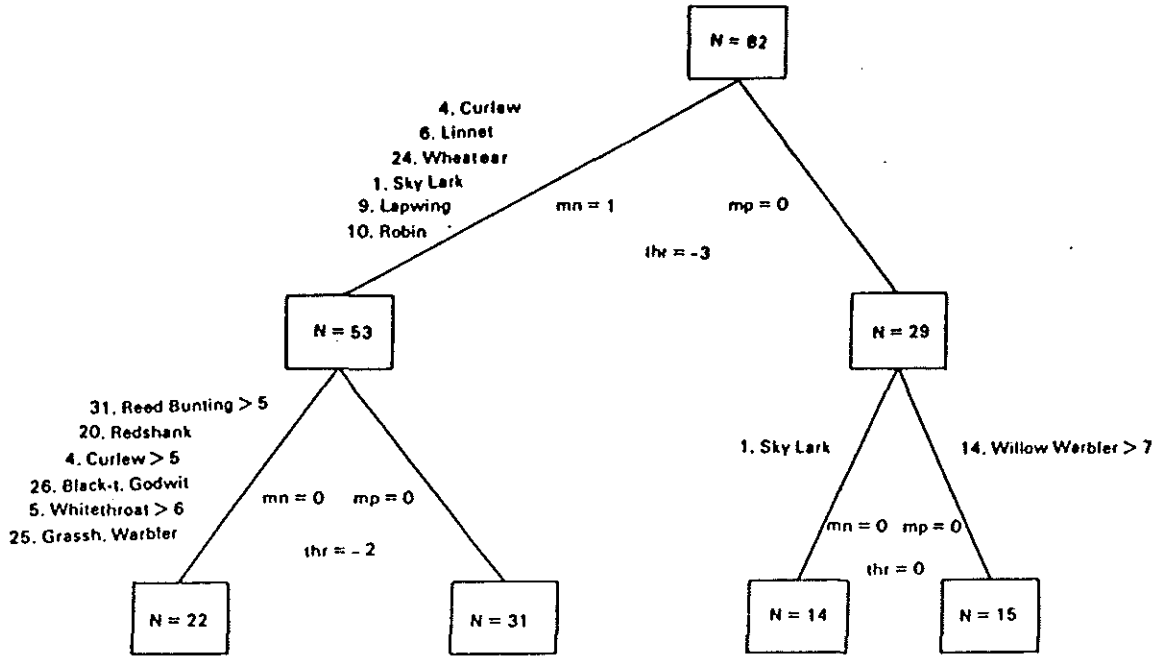

Fig.2: Indicator species for the first two levels of division of TWINSPAN. Some species are only an indicator if they occur with high abundance, e.g. Curlew > 5 means Curlew is an indicator if the abundance reaches abundance class 5 or over. (N: number of heathlands in group; thr: threshold value (maximum discriminant score for negative group); mn: number of misclassified negatives; mp: number of misclassified positives.)

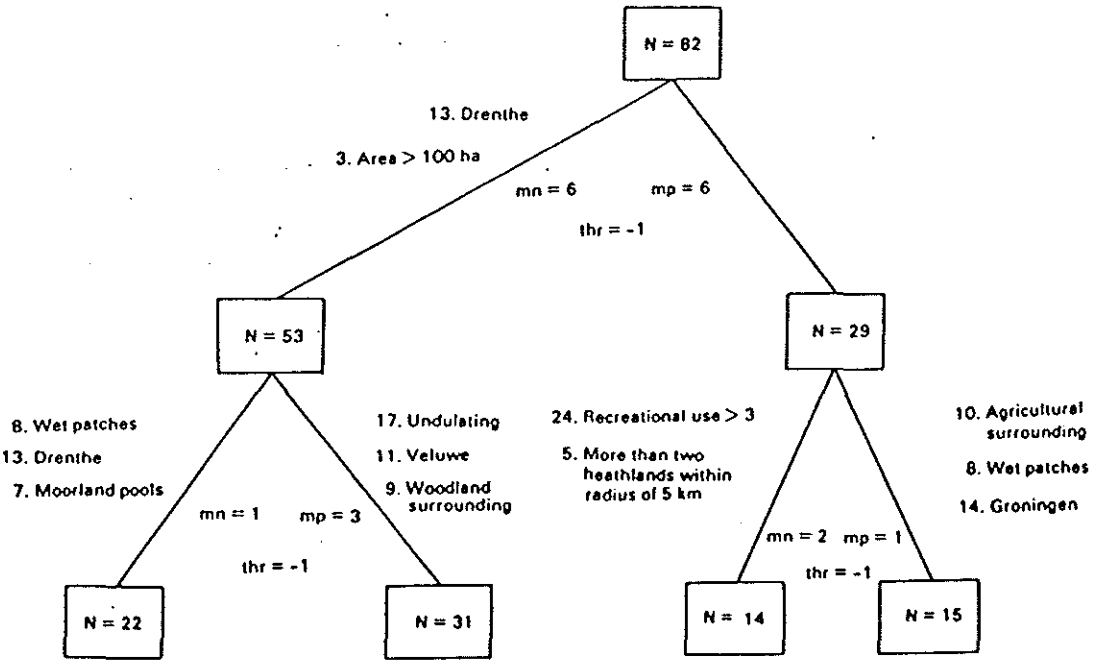

Fig.3: Indicator attributes resulting from DISCRIM that best predict the divisions of TWINSPAN. Legends as in Fig. 2 and Table 2.

One could ask why the sites were not classified on the basis of the environmental variables at the beginning of the analysis and why the hierarchy so obtained was not then interpreted in terms of the species data. However, reversing the procedure would not be asking the right question. What is of ecological interest is which environmental variables effect species composition. If a single environmental variable determines the occurrence of different species, then a cluster analysis based on all the environmental variables can only mask the differences in the causal variable. Species composition does not need to be related to the group structure obtained, if a large number of 'nonsense variables' are added. In contrast, the procedure followed in this paper would show the causal variable as an indicator. Therefore the correct procedure is to first base the classification on the response variables (also termed dependent variables) and then interpret the classification obtained in terms of the explanatory variables.

I strongly believe that the cluster program TWINSPAN also has many potentials outside ecology. When nominal and quantitative variables are recoded as dummy variables, successive divisions will be based on successive multiple correspondence analyses, which also have attractive properties. Although primarily designed for ecologists, TWINSPAN and DISCRIM may therefore also be useful for other research workers to relate a set of dependent variables to a set of explanatory variables in heterogeneous data sets through cluster analysis. The FORTRAN-program TWINSPAN can be obtained by writing to Dr. H.G. Gauch, Ecology and Systematics, Cornell University. Ithaca, New York 14853, and DISCRIM is available from the author at a nominal cost.

ACKNOWLEDGEMENTS

I am grateful to Dr. M.O. Hill for his permission to base DISCRIM directly on the FORTRAN-code of TWINSPAN. I would also like to thank the Research Institute of Nature Management, in particular Dr. P. Opdam, whose projects initiated this research. Dr. J.J. de Gruijter suggested that I used the hierarchical structure in cluster interpretation. Thanks are also due to Dr. I.C. Prentice for stimulating discussions. Mrs. C. Hengeveld corrected the English.

REFERENCES

[1] Benzécri, J.-P. et al. (1973). L'Analyse des Données. Tome 2: L'Analyse des Correspondances. Dunod, Paris.

[2] Braak, C.J.F. ter (1982). DISCRIM - A modification of TWINSPAN (Hill, 1979) to construct simple discriminant functions and to classify attributes, given a hierarchical classification samples. Report C 82 ST 107 56. TNO Institute of Mathematics, Information Processing and Statistics, P.O.Box 100, 6700 AC Wageningen, the Netherlands.

[3] Gauch, H.G. (1982). Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge.

[4] Gauch, H.G. and Whittaker, R.H. (1981). Hierarchical classification of community data. J. Ecol., 69, 537-557.

[5] Greig-Smith, P. (1983). Quantitative Plant Ecology. 3rd Ed. Blackwell, Oxford.

[6] Heiser, W.J. (1981). Unfolding analysis of proximity data. PhD thesis. University of Leiden, Leiden, The Netherlands.

[7] Hill, M.O. (1974). Correspondence analysis: a neglected multivariate method. Appl. Stat., 23, 340-354.

[8] Hill, M.O., Bunce, R.G.H. and Shaw, M.W. (1975). Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. J. Ecol., 63, 597-613.

[9] Hill, M.O. (1977). Use of simple discriminant functions to classify quantitative phytosociological data. In: Diday et al. (eds.), First International Symposium on Data Analysis and Informatics, Versailles, 6-7 September 1977, Vol. 1, pp. 181-199. Institute de Recherche d'Informatique et d'Automatique, Domaine de Voleceau, Rocquencourt, B.P. 105, 78150 Le Chesnay, France.

[10] Hill, M.O. (1979). TWINSPAN. A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Ecology and Systematics. Cornell University, Ithaca, New York, USA.

[11] Jambu, M. (1978). Classification automatique pour l'analyse des données. 1-méthods et algorithmes. Dunod, Paris.

[12] Kalkhoven, J. and Opdam, P. (1984). Classification and ordination of breeding bird data and landscape attributes . In: Brandt, J. and Agger, P. (eds.), Methodology in landscape ecological research and planning. Proceedings of the first international seminar of IALE. Vol. 3, theme 3, pp. 15-26. Roskilde Universitetsforlag GeoRuc, Roskilde, Denmark.

[13] Lefkovitch, L.P. (1976). Hierarchical clustering from principal coordinates: an efficient method for small to very large numbers of objects. Math. Biosc., 31, 157-174.

[14] Lerman, I.C., Hardouin, M. and Chantrel, T. (1980). Analyse de la situation relative entre deux classifications floues. In : Diday et al. (eds.), Data Analysis and Informatics, pp. 523-552. North-Holland, Amsterdam, The Netherlands.

[15] Noy-Meir, I. (1973). Divisive polythetic classification of vegetation data by optimized divisions on ordination components. J.Ecol., 61, 753-760.

[16] Opdam, P. and Retel Helmrich, V. (1984). Vogelgemeenschappen van heide en hoogveen: een typologische beschrijving. Limosa, 57, 47-63.