Acta Bot. Neerl. 35(3), August 1986, p. 137–142

FLEXCLUS, AN INTERACTIVE PROGRAM FOR CLASSIFICATION AND TABULATION OF ECOLOGICAL DATA

O. VAN TONGEREN

Botanisch Laboratorium, afd. Experimentele Plantenoecologie, Katholieke Universiteit,
Toernooiveld, 6525 ED Nijmegen
(present address: Limnologisch Instituut, Vijverhoflaboratorium, Rijksstraatweg 6, 3631 AC
Nieuwersluis.)

SUMMARY

A new computer program for the clustering of ecological data is presented. It is different from the programs which are commonly used in two aspects: Firstly, the program is interactive to a large extent, which implies that a priori decisions are limited. Secondly, the results of the clustering can be evaluated in terms of environmental data for the sites. The program provides several options to construct a species by site table. It is concluded that interactive clustering is preferable when data properties in terms of the (dis)similarity measure or clustering algorithm are less familiar to the user of the program.

1. INTRODUCTION

Classification procedures are widely used in ecology. By classifying sites they are assigned to classes or groups of sites, which are called clusters. The central purpose of community classification is to summarize large community data sets (GAUCH 1982). The results of a classification can be presented in many different ways:

1. as an arranged species by site table;
2. as a synoptical table in which the date are summarized for the classes;
3. as a dendrogram in which the hierarchical structure of the data is elucidated.

From the beginning of community classification the species by site table has played an important role for the classification process and for the interpretation of the results of a classification. For instance in the Zürich-Montpellier school of vegetation science classifying vegetation by tabular comparison still is one of the most widely used methods (cf. MUELLER-DOMBOIS & ELLENBERG 1974; VAN DER MAAREL & WESTHOFF 1978).

During the last decades many computer programs and packages have been developed to ease the task of classification and to make the techniques more formal and objective. Programs and packages are different in several aspects. CLUSTAN (WISHART 1969) and TABORD (VAN DER MAAREL et al. 1978) use a complete (full format) data matrix. Others like the Cornell Ecology programs COMPCLUS (GAUCH 1979) and TWINSPAN (HILL 1979) use a so-called condensed format, which saves a lot of computer memory and computing time, since ecological data matrices tend to be sparse. Packages like CLUSTAN (WISHART 1969) and BIOPAT (HOGEWEG & HESPER 1972) use several methods and indices of similarity, while special purpose programs like COMPCLUS (GAUCH 1979), TWINSPAN (HILL 1979) and CLUSLA (LOUPPEN & VAN DER MAAREL 1979) are very much restricted in their options. The FLEXCLUS (flexible clustering) program which is described in this paper is intermediate by the combination of a small choice of methods with relatively few indices of similarity and dissimilarity. The program is different from other programs because it is possible to influence the results by interaction with the computer (thereby overruling the chosen method) and by the fact that during execution of the program several options can be changed, including the clustering strategy.

2. STRUCTURE OF THE PROGRAM

Main sections of the program, which are performed sequentially are:

2.1. Data editing

The species by site data are read from a 'Cornell condensed' file (e.g. HILL 1979). Data editing includes sample selection, data transformations, weighing of sites and species. For minor changes in the data matrix this avoids the use of a separate program.

2.2. Initial clustering

Optionally a site by site (dis)similarity matrix is computed. An initial clustering based on this site by site matrix is constructed following the algorithm of SØRENSEN (1948). Alternatively a composite clustering (cf. GAUCH 1979) is performed or an initial classification is read from a disk file. These methods do not require a full site by site (dis)similarity matrix and therefore are computationally quicker.

2.3. Final clustering

Using the results of the initial clustering this can be refined iteratively by:

a. Fusion of two clusters following the centroid clustering algorithm.
b. Fusion of a TWINSPAN (HILL 1979) dichotomy if a TWINSPAN result is used as the initial classification.
c. Fusion of any two clusters, indicated by the user.
d. Division of the most heterogeneous cluster either by a 'pseudodivisive' method (within the cluster the sites are agglomeratively fused according to the SØRENSEN (1948) algorithm or by a random division.
e. Division of any cluster, as specified by the user.
f. Automatic subsequential fusions following the centroid clustering algorithm with optional reallocation after each fusion (cf. TABORD; VAN DER MAAREL et al. 1978).
g. Automatic subsequential divisions of the most heterogeneous cluster.
h. Removal of outliers from clusters by the specification of a maximum cluster radius.

Reallocation is optional after each step.
A summary of 'cluster statistics' is displayed at the terminal screen after each step. This summary includes:

1. Average similarity of the members of a cluster to its centroid;
2. Similarity of the centroid to the centroid of the nearest cluster;
3. The quotient of these two (a kind of measure for the optimality of the cluster).

The summary is optionally followed by a species by cluster table to help the user to make decisions for the next step.

2.4. Construction of a species by sites table.

The final clusters are arranged along the first axis of a reciprocal averaging ordination (HILL 1973) of the cluster centroids. Small clusters, which tend to be composed of outlying samples, are excluded from this ordination and moved towards the right hand side of the table. If a TWINSPAN result was used as the initial clustering and only TWINSPAN dichotomies have been fused the TWINSPAN order of the clusters can be maintained. A third possibility is to rearrange the clusters by hand. Species are ordered according to their preference for clusters. A block of constant species is followed by two diagonal structures. The first diagonal is for species with a narrow ecological amplitude, the second one for species with a wider ecological amplitude (in reference to the first RA axis). Rare species are moved towards the tail of the table.

2.5. Interpretation of the clusters in the light of environmental variables.

After completing the final table the program checks whether more input data follow the species by site data. Environmental data must be preceded by a code, indicating the procedure that should be followed. Three options are available:

1. Computation of mean and standard deviation for continuous environmental variables
2. Counting the number of occurrences of values within specified classes
3. Listing of the 'value' of nominal variables for all members of each cluster.

Several sets of environmental data can be chained in the input file and will be analyzed sequentially till no more data are available.

3. EXAMPLE

The data for the example are taken from a report of BATTERINK & WIJFFELS (1983). After an initial clustering (method SØRENSEN, 1948) eight clusters were formed. These eight clusters were fused in three steps to five clusters using centroid clustering with reallocation till stability after each fusion. One of these clusters, consisting of site 1 only, was considered to be a species poor representative of the second cluster in the table (inspection of the species by cluster table on the terminal screen). This site was therefore added to the second cluster, overruling the centroid clustering. Centroid clustering would have resulted in one very large cluster (the first plus the second cluster, without site 1). The result

sites:	1	1		1		1	1								1	1	1	1	1	2
	7	9	6	1	7	0	8	5	1	2	3	4	8	9	2	3	4	5	6	0
species:																				
Leo aut	2	6	3	5	3	3	5	3		5	2	2	3	2	4	5	4	4	7	5
Bra rut		3	9	4	2	2	6	2			2	2	2	2	4			4	4	4
Tri rep		2	5	3	2	6	2	2		5	2	1	2	3	3	2	6	1		
Agr sto											4	8	4	3	4	5	4	4	7	5
Ach mil	2		2		2	4		2	1	3										
Ant odo	4	4	3		2	4		4												
Pla lan	2		5	3	5	3	3	5												
Poa pra	1		3	4	4	4	3	2	4	4	5	4	4		2					
Lol per			6	7	6	6	2	2	7	5	6	4	2							
Bel per						2	2	2		3	2	2								
Ely rep								4	4	4	4	4		6						
Alo gen										2	7	2	5	3	8	5			4	
Poa tri			4		5	4		6	2	7	6	5	4	5	4	9			2	
Sag pro		3		2								5	2	2	4	2				
Jun buf					2									4	4	3				
Jun art													4	4				3	3	4
Cal cus																	4		3	3
Ele pal													4				4	5	8	4
Ran fla													2			2	2	2	2	4
Air pra	2	3																		
Bro hor					2	4		2		4		3								
Hyp rad	2	5		2																
Pot pal																	2	2		
Rum ace			6		3			5						2	2					
Sal rep		3					3													5
Tri pra			5		2			2												
Vic lat				2		1	1													
Che alb																1				
Cir arv												2								
Emp nig		2																		

environmental parameters

depth A1				
mean	4.0	3.8	5.9	7.5
SD	1.1	0.6	0.1	3.6
% owned				
by SBB	38	0	0	75
moisture				
class 1	*****	**		
class 2	**	**		
class 3				
class 4		*	*	
class 5	*	*	*	****
manure				
class 0				***
class 1	*****	*		
class 2	**	*		
class 3	*	*	*	*
class 4		***		

Table 1.
An example of the results of FLEXCLUS. The steps involved are explained in the text. Species names are abbreviated to the first three characters of the genus and the first three characters of the epitheton specificum. Classes for environmental variables are arbitrarily chosen.

is presented in *table 1*, together with the analysis of the environmental data. The analysis reveals a relation between the clustering and moisture class, represented by the shift in the peak of the histogram for moisture class from low to high. The first and second cluster are different in respect to manure class, the second one consisting of more heavily fertilized sites than the first one.

4. DISCUSSION

Clustering can be used for several purposes and most computer programs are constructed to achieve an optimal result in respect to one of the possible aims of cluster analysis. FLEXCLUS, by combining several strategies gives a result which is optimized in respect to several of these aims in one run.

The initial, non-hierarchical clustering is a means for handling noise and redundancy by combining several samples into groups. Outliers which are expected to be in small clusters can be identified in this stage. Refinement of the initial clustering by reallocation and fusions is achieved in the final clustering step. This final step can also be used to express the hierarchical relations between clusters. However, these relations can only be expressed up to the hierarchical level at which the table is constructed. Unless the construction of a species by site table is abandoned, a second run of the program is needed for the construction of the higher hierarchy. The hierarchical relations between the clusters are different from those obtained when composite samples are used for a hierarchical clustering, because the composite samples formed by the initial clustering are implicitly weighted according to the sizes of the clusters.

By different combinations of options the results of several other programs can be simulated: A composite clustering (cf. COMPCLUS, GAUCH 1979) can be constructed in the initial clustering step. The results are slightly different from those obtained with COMPCLUS, because each sample is allocated to the nearest node instead of being allocated to the first node which becomes available. The CLUSLA (VAN DER MAAREL & LOUPPEN 1979) result can be simulated (with restrictions to the number of samples) by combining the composite clustering with reallocations till stability in the final clustering step. The result is less dependent on the sequence of the input data. A table comparable to the result of TABORD (VAN DER MAAREL et al. 1978) is obtained by reading an initial clustering followed by automatic fusions, with or without reallocations, till a specified number of clusters or a specified between cluster similarity is reached. Also in this case the results are different. During reallocation of sites to clusters the centroids are newly computed after each single reallocation, instead of being computed after one complete iteration cycle. The ordering of clusters is obtained by reciprocal averaging (HILL 1973) instead of Principal Component Analysis (PCA) or the algorithm of Janssen (VAN DER MAAREL et al. 1978), and the species order is defined in a different way. The slower, but better reallocation procedure is compensated for by the huge saving of computing time by the quicker algorithms for sparse matrices.

In other programs which are frequently used in ecology the options and stopping rules have to be specified on beforehand. This implies that one not only should know his data, but also the properties of his data in respect to the specified similarity or dissimilarity index and the cluster algorithm used. In general this means, that several runs of these programs have to be made before the result is satisfactory. The inspection of the clusters in respect to their species composition and in terms of 'cluster statistics', followed by the decision for the next step makes it possible to achieve a satisfactory result in one run, even for less experienced users of the program.

Automatic evaluation of the clusters in the light of environmental variables eases the task of choosing those environmental variables which should be evaluated by means of statistical tests.

ACKNOWLEDGEMENTS

The technical help of M. Manders and P. Groen of the University Computer Centre in Nijmegen is acknowledged as well as the initial stimulus I received from Prof. Dr. E. van der Maarel. Most of all I am grateful to Prof. Dr. H. F. Linskens who stimulated my work in Nijmegen, where this program was developed.

REFERENCES

BATTERINK, M. & G. WIJFFELS (1983): *Een vergelijkend vegetatiekundig onderzoek naar de typologie en invloeden van het beheer van 1973 tot 1982 in de duinweilanden op Terschelling.* Report V.P.O., Agricultural University, Wageningen.

GAUCH, H. G. (1979): *COMPCLUS. A FORTRAN Program for Rapid Initial Clustering of Large Data Sets.* Ecology and Systematics, Cornell University, Ithaca, New York.

— (1982): *Multivariate Analysis in Community Ecology.* Cambridge University Press, Cambridge.

HILL, M. O. (1973): Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* **61**: 237–249

— (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.

HOGEWEG, P. & B. HESPER (1972): *BIOPAT, program system for biological pattern analysis.* Theor. Biol. Group. University of Utrecht.

LOUPPEN, J. M. W. & E. VAN DER MAAREL (1979): CLUSLA: A computer program for the clustering of large phytosociological data sets. *Vegetatio* **40**: 107–114.

MAAREL, E. VAN DER & V. WESTHOFF (1973): The Braun-Blanquet approach. In: R. H. WHITTAKER (ed.) (1980): *Classification of plant communities,* 2nd edn., Junk publishers, the Hague: 289–399.

—, J. G. M. JANSSEN & J. M. W. LOUPPEN (1978): TABORD, a program for structuring phytosociological tables. *Vegetatio* **38**: 143–156.

MUELLER-DOMBOIS, D. & H. ELLENBERG (1974): *Aims and methods of vegetation ecology.* Wiley, London.

SØRENSEN, T. (1984): A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Det. Kong. Danske Vidensk. Selsk. Biol. Skr* (Copenhagen) **5**(4): 1–34.

WISHART, D. (1969): *CLUSTAN IA. User Manual.* St. Andrews Computing Centre.