COENOSES 10(1):11-16, 1995

ROBUSTNESS OF NUMERICAL METHODS FOR VEGETATION CLASSIFICATION

P. S. Torres, I. M. Barberis & J. P. Lewis

Universidad Nacional de Rosario, Facultad de Ciencias Agrarias, Casilla de Correo 14, 2123 Zavalla, Argentina.

Keywords: Classification, Clustering, Robustness, Similarity measures, Vegetation.

Abstract: The robustness of MULVA and PC-ORD clustering techniques were measured with three different new indices. A method is robust when the results for a given data set are stable despite minor perturbations of the data. In order to test the robustness of these methods 22 plots were classified according to the abundance of trees or trees and shrubs and then 10 new plots were added and reclassified. Of the tested methods, Minimum Variance with Cross Product matrix of non-centred data was the most robust and Complete Linkage with Chord Distance was the least robust.

Introduction

Numerical methods are increasingly used in vegetation analysis and continuously new algorithms and more sophisticated techniques are being developed. The two main types of numerical methods used in vegetation analysis are ordination and classification.

Their most noted advantage over classical phytosociological methods is that they are more objective. However, the choice between two different methods is not always as objective as it should be. Frequently, different clustering algorithms produce dissimilar classifications when applied to the same set of data (Podani 1989a) and different ordination techniques may produce different results under the same circumstances.

Mazzoleni et al. (1991) identify three main groups of papers in the ecological literature comparing different methods. In the first group, classical phytosociological classification has been compared with different numerical methods; in the second, different techniques of numerical clustering were compared to observe the effects of changing resemblance measures or agglomeration criteria and the third group examined the results of applying various ordination techniques to different data sets, both real and simulated, to compare their relative performance.

Clustering methods should produce good classifications, the heterogeneity between groups should be larger than the heterogeneity within groups (Orlóci 1978) and hierarchical relationships between groups should be shown by dendrograms. Dendrograms may be compared with several descriptors such as the cophenetic correlation coefficient (Sokal & Rohlf 1962), the topological distance (Phipps 1971), cluster membership divergence, subtree membership divergence, and partition membership divergence (Podani & Dickinson 1984, Podani 1989b)

Good numerical methods should be robust, i.e. effective results are produced in most applications of the technique (Gauch 1982). Robustness has two components: a) the results for a given data set are stable despite minor perturbations of the data and b) effective results are produced for a wide variety of data sets.

The usefulness of several ordination techniques in the survey of the centre of North England mires was analysed by Clymo (1980). An evaluation of the robustness of different ordination techniques was done by Minchin (1987). When different classification techniques are compared, usually only the results over the same set of data are considered.

Two sets of programs are widely used in vegetation analysis, MULVA (Wildi & Orlóci 1990) and PC-ORD (McCune 1991). The objective of this paper is to evaluate the robustness of their programs when new plots are added to an original set of plots and when different sets of species are used to classify the different plots. If a method is robust, when a new set of plots is added it shoud be expected that they will form new groups altogether if they are completely different from the original ones, or if they are members of the same universe, they will be classified into the original groups. The original plots should remain in the original groups. We test the techniques under this proposition.

Material and methods

Data are from a *Schinopsis balansae* forest of the Gran Chaco analysed by Lewis (1991). Along two parallel transects, 22 and 10 plots of 10 x 10 m were laid, and the number of trees and shrubs were counted. These data were analysed with the different combinations of methods and similarity measures provided by the MULVA and PC-ORD program packages. The results of the analysis of 22 plots in one of the transects were compared with the results of the analysis of all the 32 plots from both transects. First the analysis was done

Table 1. Robustness of different methods tested with three indices when tree and tree and shrub data were analyzed. Confidence limits for p=0.01.

	RESEMBLANCE MATRIX	AGGLOMERATION CRITERIA	N°	TREE DATA ALONE I_{R1}	I_{R2}	I_{R3}	TREE + SHRUB DATA I_{R1}	I_{R2}	I_{R3}
MULVA	Cross product of centred data	Complete linkage	1	0.67	0.80	0.73	0.86 *	0.80	0.76
	Covariance	Complete linkage	2	0.67	0.80	0.73	0.86 *	0.80	0.76
	Correlation coefficient	Complete linkage	3	0.69	0.87	0.77	0.80	0.85	0.82
	Euclidean distance	Complete linkage	4	0.72	0.98 *	0.92 *	0.83	0.95 *	0.91 *
	Chord distance	Complete linkage	5	0.14	0.58	0.55	0.27	0.61	0.61
	Ochiai's coefficient	Complete linkage	6	0.72	0.87	0.82	0.81	0.83	0.79
	van der Maarel's coefficient	Complete linkage	7	1.00 *	1.00 *	0.95 *	0.83	0.96 *	0.96 *
	Cross product of non-centred data	Complete linkage	8	0.52	0.79	0.75	0.83	0.96 *	0.94 *
	Cross product of centred data	Minimum variance	9	0.92 *	0.93 *	0.90 *	0.67	0.83	0.77
	Covariance	Minimum variance	10	0.92 *	0.93 *	0.90 *	0.67	0.83	0.77
	Correlation coefficient	Minimum variance	11	0.71	0.86	0.78	0.90 *	0.95 *	0.86
	Euclidean distance	Minimum variance	12	0.89 *	0.94 *	0.94 *	0.60	0.92	0.90 *
	Chord distance	Minimum variance	13	0.83	0.91	0.87	0.92 *	0.87	0.79
	Ochiai's coefficient	Minimum variance	14	0.83	0.91	0.87	0.90 *	0.75	0.66
	van der Maarel's coefficient	Minimum variance	15	1.00 *	1.00 *	0.85	0.89 *	0.94 *	0.94 *
	Cross product of non-centred data	Minimum variance	16	1.00 *	0.96 *	0.96 *	1.00 *	1.00 *	0.96 *
PC-ORD	Squared euclidean distance	Complete linkage	17	0.81	0.98 *	0.98 *	0.83	0.96 *	0.96 *
	Squared relative euclidean distance	Complete linkage	18	0.85 *	0.90	0.90 *	0.78	0.95 *	0.90 *
	Sørensen's coefficient	Complete linkage	19	0.76	0.86	0.78	0.92 *	1.00 *	0.92 *
	Squared euclidean distance	Median	20	0.60	0.58	0.55	0.61	0.90	0.89
	Squared relative euclidean distance	Median	21	0.80	0.84	0.84	0.68	0.98 *	0.85
	Sørensen's coefficient	Median	22	0.70	0.76	0.73	0.60	0.92	0.90 *
	Squared euclidean distance	Group average	23	0.95 *	0.80	0.78	1.00 *	1.00 *	0.96 *
	Squared relative euclidean distance	Group average	24	0.86 *	0.95 *	0.92 *	0.45	0.74	0.74
	Sørensen's coefficient	Group average	25	0.66	0.90	0.84	0.85 *	0.97 *	0.95 *
	Squared euclidean distance	Centroid	26	0.80	0.70	0.68	1.00 *	1.00 *	0.95 *
	Squared relative euclidean distance	Centroid	27	0.70	0.90	0.90 *	0.70	0.96 *	0.87
	Sørensen's coefficient	Centroid	28	0.95 *	0.83	0.80	0.67	0.95 *	0.94 *
	Squared euclidean distance	Ward's method	29	0.80	0.94 *	0.92 *	0.60	0.92	0.92 *
	Squared relative euclidean distance	Ward's method	30	0.78	0.89	0.86	0.83	0.94 *	0.88
	Sørensen's coefficient	Ward's method	31	0.80	0.92 *	0.89 *	0.54	0.83	0.63
	Squared euclidean distance	Mc Quitty's method	32	0.69	0.89	0.89 *	0.54	0.77	0.76
	Squared relative euclidean distance	Mc Quitty's method	33	0.87 *	0.94 *	0.90 *	1.00 *	1.00 *	0.90 *
	Sørensen's coefficient	Mc Quitty's method	34	0.83	0.91	0.91 *	0.53	0.92	0.81
		INDEX AVERAGE		0.77	0.87	0.83	0.76	0.90	0.86
		INDEX STANDARD DEVIATION		0.1667	0.1034	0.1062	0.1707	0.0936	0.1010
		UPPER VALUE OF THE CONFIDENCE LIMITS		0.85	0.92	0.88	0.84	0.94	0.90

with only tree data and then with the tree and shrub data together.

The robustness of methods was measured with the following three indices:

1) Fixed weighted index, I_{R1}:

$$I_{R1} = \frac{1}{5}\left(\sum_{i=1}^{k} p_i \frac{n_i}{n}\right) \quad (1)$$

where k is the number of groups, p_i is the weight of group i, n_i is the size of group i and n is the size of the initial set of data.

The size of groups is the number of plots from the initial set of data. Their weights are obtained in the following way: Groups from the original set of data are divided in three classes (A, B and C) according to the percentage of plots that remain united when the analysis is done with all data; A if 100% remain united, B from 70% to 99% and C less than 70%. Then each class is divided in two (o and n) if the order of plot fusion is kept in more or less than 70% of the cases. The weights will be Ao = 5, An = 4, Bo = 3, Bn = 2, Co = 1 and Cn = 0. This index varies from 0 to 1 (maximum robustness).

2) Plot permanence index, I_{R2}:

$$I_{R2} = \frac{1}{k}\left(\sum_{i=1}^{k} \frac{m_i}{n_i}\right) \quad (2)$$

where m_i is the number of plots that remain in group i of the original set when data from more plots is added, and k and n_i are the same as in (1). This index also varies from 0 to 1.

3) Plot order index, I_{R3}:

$$I_{R3} = \frac{1}{k}\left(\sum_{i=1}^{k} \frac{o_i}{n_i}\right) \quad (3)$$

where o_i is the number of plots that remain in the same order

as in the original analysis when the analysis is done with added data. k and n_i are defined as before. This index also varies from 0 to 1.

A method is robust if the index value approaches 1. To test relative robustness measured with any particular index confidence limits were estimated for p = 0.01; if robustness value was higher than the upper limit, the method is significantly robust (Sokal & Rohlf 1979).

In order to test if robust methods produce ecologically meaningful results they were compared with the outcome of the classical phytosociological method (Werger & Sprangers 1982, Mazzoleni et al 1991).

Results and discussion

Single Linkage (Nearest Neighbor) agglomeration methods from both MULVA and PC-ORD were discounted because these methods are prone to produce chaining, that is, the sequential addition of single plots to one large group. Therefore, no defined groups are formed and as a whole they can be considered a second order method of ordination (Nimis, pers. comm.). Flexible beta also was discarded because its results depend on the stated value of β which varies from 1 to -1; if it approaches 1, chaining increases so results will be like those of Single Linkage and if it approaches -1, the results will be similar to Complete Linkage (Mc Cune 1991). All other 34 possible combinations of methods and similarity measures, even if they were incompatible, were tested. In a compatible method, similarity measures calculated later in the analysis are of the same kind as the initial interstand similarity measures (Greig-Smith 1983), which is not the case when Sorensens coefficient is used with the Median, Ward's or Mc Quitty's agglomerative strategies of PC-ORD. Ward's method in PC-ORD is equivalent to the Minimum Variance method in MULVA.

Complete Linkage using van der Maarel's coefficient (7 in Table 1), Minimum Variance using van der Maarel's coefficient (15) and Cross Product of non-centred data resemblance matrix (16), all from MULVA and Mc Quitty's method with the square of relative Euclidean Distance (squared Chord Distance, 33) from PC-ORD were the most robust methods measured by any of the robustness indices whether tree alone (Fig. 1) or tree plus shrubs (Fig. 2) data were analysed (Table 1). From these, MULVA's Minimum Variance with cross product matrix of non-centred data (16) and PC-ORD's Mc Quitty's method with squared Chord Distance (33) appear to be the most robust methods. Minimum variance with van der Maarel's coefficient (15) or cross products of non-centred data (16) and Complete Linkage with van der Maarel's coefficient (7) produced the same groups (clusters) and were almost equally robust.

MULVA's Minimum Variance with Cross Product matrix of centred data (9) and Covariance matrix (10) as well as PC-ORD's Group Average with the square of relative Euclidean Distance (squared Chord Distance) (24) were very robust when only tree (14 species) data were analysed, but when tree and shrubs (27 species) data were analysed their robustness was lower, especially the latter (Group Average). It may be that higher diversity (floristic richness) affects the robustness of these methods. However, this is not the general case, as several methods, especially PC-ORD's Complete

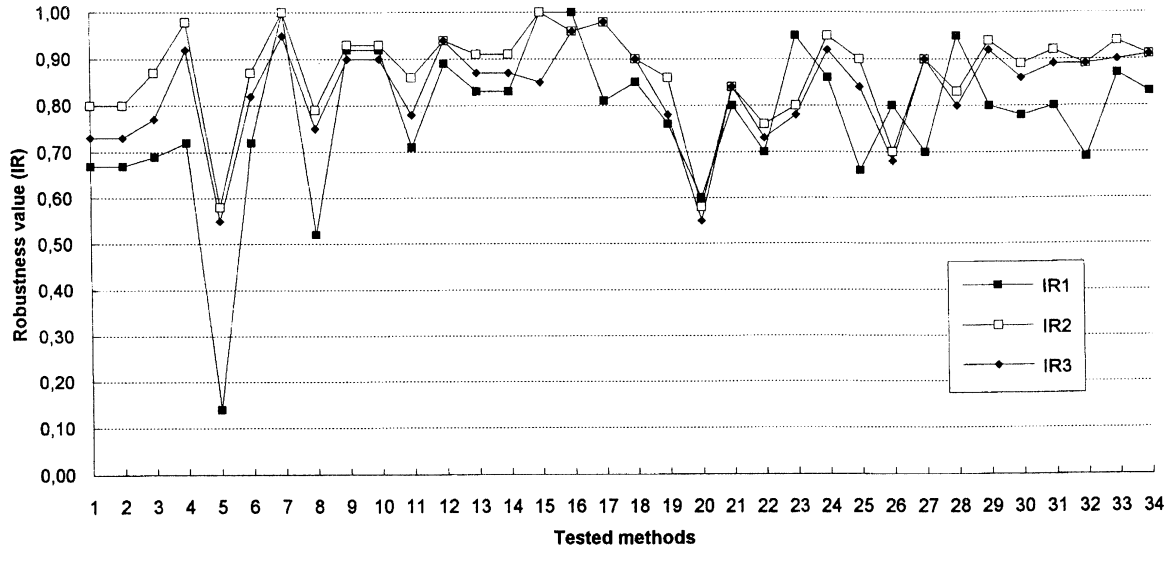

Figure 1. Robustness of all tested methods estimated with the three indices when only tree data are analysed. Methods numbered as in Table 1.

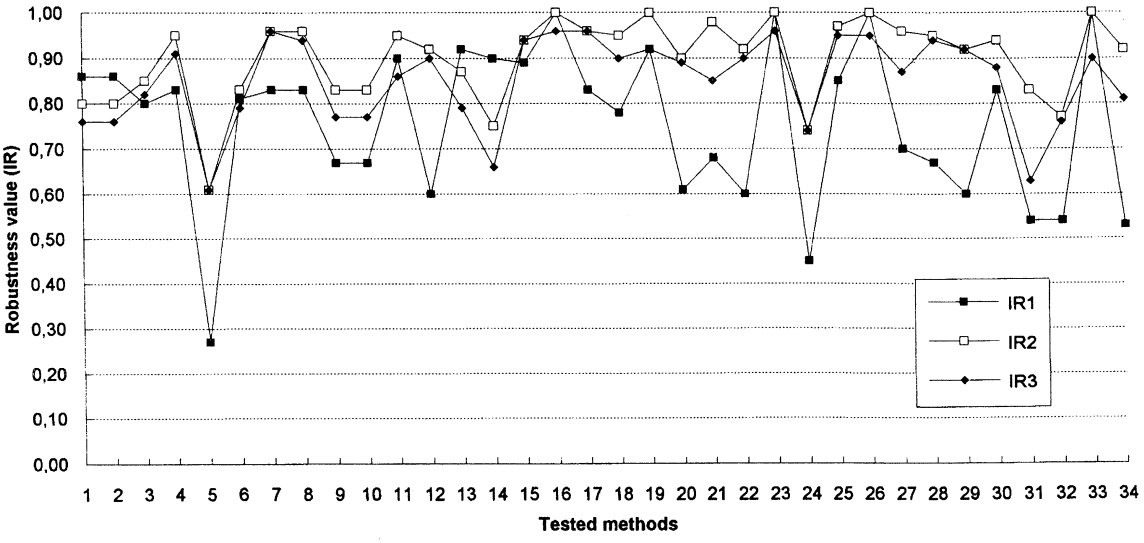

Figure 2. Robustness of all tested methods estimated with the three indices when tree and shrub data are analysed. Methods numbered as in Table l.

Minimum variance with Cross product of non-centred data (N° 16)	Braun-Blanquet's method	Complete linkage with Squared euclidean distance (N° 17)
20 ●	13 ●	1 ★
12 ●	12 ●	6 ○
13 ●	20 ●	3 ○
8 ●	8 ●	16 ✧
9 ●	9 ●	21 ✧
22 ○	7 ●	18 ■
7 ●		14 ★
	22 ○	19 ○
15 □	19 ○	4 ■
2 □	6 ○	17 ■
5 □	3 ○	
10 □		2 □
	17 ■	5 □
17 ■	11 ■	15 □
11 ■	18 ■	10 □
4 ■	4 ■	
18 ■		7 ●
	15 □	9 ●
21 ✧	2 □	11 ■
16 ✧	5 □	22 ○
	10 □	
19 ○		8 ●
14 ★	14 ★	
6 ○	1 ★	12 ●
1 ★		13 ●
3 ○	21 ✧	20 ●
	16 ✧	

Figure 3. Minimum variance with cross product matrix of non-centred data and Complete Linkage with D^2 clusters compared with the results of Braun-Blanquet's method when only tree data are used.

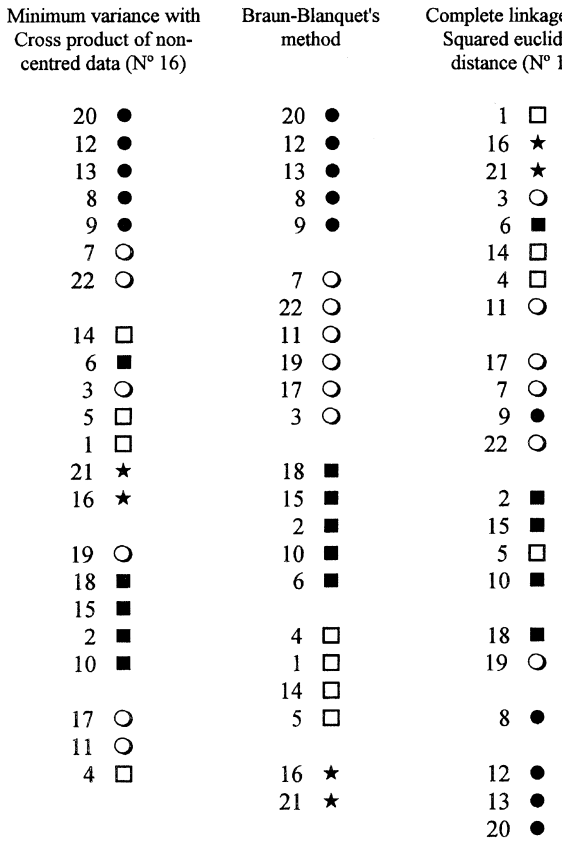

Figure 4. Minimum variance with cross product matrix of non-centred data and Complete Linkage with D^2 clusters compared with the results of Braun-Blanquet's method when tree and shrub data are used.

Linkage and Group Average with Sorensen coefficient (19 and 25) as well as Centroid with the square of Euclidean Distance (D^2, 26) were more robust when the larger set of data is analyzed (Table 1).

When the agglomeration method is not compatible with the resemblance measure, robustness tends to be low. However, other combinations are even less robust as in the case of Complete Linkage with Chord Distance (5) which is apparently the worst of all the methods. With these data, absolute abundance, Chord distance as similarity measure has very poor performance, however, with other types of data it may be not the case.

Minimum variance with van der Maarel's coefficient (15) for tree data and Minimum Variance with Correlation Coefficient (11); Median, Centroid and Ward's method with squared relative Euclidean Distance (21, 27 and 30) for tree and shrub data were robust when they were measured with the plot permanence index (I_{R2}). However, this was not the case when they were measured with the plot order index (I_{R3}). These results show that some methods could be robust in keeping the same groups but not their internal stability.

Four methods for tree data (18, 27, 32 and 34) and three methods for tree and shrub data (12, 22 and 29) were significantly robust when they were measured with the third index (I_{R3}) but were not when measured with the second one. This may be due to differences between the upper values on the confidence limits of each index, since I_{R3} values must be less than or equal to those of I_{R2}.

The first index (I_{R1}) is a good measure of robustness, but it has the disadvantage that external information is added to the data in the fixed weighting. The second one (I_{R2}) measures only if groups remain stable while the third (I_{R3}) measures the internal stability of groups as well; the disadvantage of both indices is that stable small groups are overvalued.

The results of the most robust method, MULVA's Minimum Variance and Cross Products of non-centred data, were more similar to those of the classical phytosociological method (Braun Blanquet 1932) than other less robust method like PC-ORD's Complete Linkage with D^2 (Figures 3 and 4).

Acknowledgment. Studentship from CONICET for I. M. Barberis is gratefully acknowledged.

References

Braun-Blanquet, J. 1932. Plant Sociology. Mc Graw-Hill, New York and London.

Clymo, R.S. 1980. Preliminary survey of the peat-bog Hummell Knowe Moss using various numerical methods. Vegetatio 42: 129-148.

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

Greig-Smith, P. 1983. Quantitative Plant Ecology. University of California Press, Berkeley and Los Angeles.

Lewis, J.P. 1991. Three levels of floristical variation in the forests of Chaco, Argentina. J. Veg. Sci. 2: 125-130.

Mazzoleni, S., D.D. French & J. Miles. 1991. A comparative study of classification and ordination methods on successional data. Coenoses 6(2): 91-101.

Mc Cune, B. 1991. Multivariate analysis on the Pc-Ord system. Oregon State University, Corvallis.

Minchin, P.R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69: 89-107.

Orlóci, L. 1978. Multivariate analysis in vegetation research. Dr Junk, The Hague.

Phipps, J.B. 1971. Dendrogram topology. Syst. Zool. 20: 306-308.

Podani, J. 1989a. Comparison of ordinations and classifications of vegetation data. Vegetatio 83: 111-128.

Podani, J. 1989b. A method for generating consensus partitions and its application to community classification. Coenoses 4: 1-10.

Podani, J. & T.A. Dickinson. 1984. Comparison of dendrograms: a multivariate approach. Can. J. Bot. 62: 2765-2778.

Sokal, R.R. & F.J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon 11: 33-40.

Sokal, R.R. & F.J. Rohlf. 1979. Biometrìa. H. Blume Ediciones, Madrid.

Werger, M.J.A. & J.T.C. Sprangers. 1982. Comparison of floristic and structural classification of vegetation. Vegetatio 50: 175-183.

Wildi, O. & L. Orlóci. 1990. Numerical exploration of community patterns. SPB Academic Publishing bv, The Hague.

Manuscript received: December 1995.