STUDIA GEOBOTANICA
4: 101-108, 1984

EFFECTS OF SAMPLING INTENSITY AND RANDOM NOISE ON DETECTION OF SPECIES GROUPS BY INTERSECTION ANALYSIS

Enrico FEOLI and Mario LAGONEGRO

Keywords: intersection, noise, sampling, simulation, vegetation

Abstract. The efficiency of Intersection Analysis in producing species groups at different noise and sampling intensity levels has been tested on the basis of a simulated coenocline. The results suggest that Intersection Analysis is a robust method for detecting species groups and that it could be used in field surveys to find out the sampling intensity sufficient to describe the vegetation under study.

Introduction

Intersection analysis has been introduced by Feoli (1977) and Feoli & Lagonegro (1979) to analyse phytosciological data. The method defines species groups according to the criterion of maximal intersection between sets of relevés defined by the presence of single species. The intersection between the sets of relevés is computed by the well known Jaccard's function, which in terms of set theory is the ratio between the intersection and the union of two sets. If one considers a relevé as a point in the multidimensional ecological space, the intersection between the sets computed by the Jaccard's function can be considered as an estimation of the relative intersection between niches according to the Huchinson's definition (see Hurlbert, 1981 for other definitions and for a deep discussion). The intersection should be considered relative for two reasons: firstly, a sample set of relevés could not include all the hypervolume of a niche (Petraitis, 1979; 1982; Smith, 1982 among others); secondly the function relativizes the intersection of the union of the sets. In other terms, the intersection estimated by the Jaccard function, within a sample, is an estimation of the probability to find two species toghether in the universe represented by the sample, given independently one of the two species.

If the response of the species is considered, i.e. the quantities of the species in the relevés, the Jaccard function can be formulated by the following expression: $I(X, Y) = Sxy/(Sx^{\circ\circ}2 + Sy^{\circ\circ}2 - Sxy)$, in which x means a single score of species X, y a single score of species Y, and S means the sum over the number of relevés. This

function estimates the similarity of the species responses in the ecological space rather than the intersection between niches. However if the resource use is considered proportional to the species response, it could be considerd as an indirect estimator of the niche overlap in the sense of McArthur and Levins (1967), for which some statistical inference methods have been described by Maurer (1982).

The use of species groups (sociological groups, when defined on the basis of phytosociological tables, following Doing, 1969) for the description of plant communities allows to reduce the dimensionality of the sample space (Orlòci, 1978) and in some cases the classifications based on average scores of species groups proved to be more predictive than those based on species (Feoli, Lagonegro & Biondani, 1981; Feoli & lagonegro, 1982). The explanation is to be found in the fact that the use of species groups reduces the effect of many low- or non-predictive species when the predictive species are few, and in the fact that the use of average values of species groups smothers the effect of random variation of the single species. The utility to find species groups is also related to the problem to make objective extrapolations of ecological indicator values (Ellenberg, 1974; Landolt, 1977).

The definition of species groups within a vegetation system depends on the intensity of sampling and on the noise level. In the present paper we test the effects of these two factors in the definition of species groups by intersection analysis and we suggets to use intersection analysis to find out the sampling intensity sufficient to describe the vegetation under study. Noise is here considered as a consequence of random fluctations or dispersion, and of errors in identification or evaluation of species response.

Methods

The effects of sample size and noise level have been tested on data sets generated by a coenocline simulator (program SPAGHET: Lagonegro, 1984). Simulated coenoclines have been already used to test the performances of ordination and classification methods (see Gauch, 1982a, for references). Gauch (1982b) uses simulated coenoclines and coenoplanes to test the effect of noise upon eigenvector ordinations.

SPAGHET allows to generate coenoclines with response curves of Gaussian profile, bimodal profile (resulting from merging two gaussian responses curves) and Poisson profile, with maximum on the right or on the left side. The coenocline my be completely defined by the user throught the position and the dispersion parameters, or generated completely at random, by giving only the number of species and the relative lenght of the x-axis. This is chosen by the user according to the wanted average density of modes for each arbitrary lenght unit. Such a unit is 1/n of the arbitrary lenght n. If for example we want an average density of 2 for 60 species the lenght must be 30.

The level of noise can be set by an option which adds or subtracts a number from the scores given by the response functions. The higher the noise level the wider the range of possible values of such a number. This is a random percentage of the

response value itself, up to a maximum equal to half the noise level chosen by the user. However, the program avoids to reach values of response greather than the maximum value of the single species response (established by user's scale). The user can chose between several scales, binary included. It must be clear that binary data ignore the species response and offer information only about species tolerance. If a user decides to normalize the data, the information of the position of species optima is added to the information about the tolerance, however the difference in quantities between species are neglected.

Coenoclines my be generated in infinite ways for testing the effects of different reasonable situations on methods of data analysis.

In the present paper we rely on a random coenocline. The reason is that if phytosociological data are used, it is impossible to forecast the niche breath and the response function for all the species. Results based on field data (e.g. Feoli, Biondani & Lagonegro, 1982; Feoli-Chiapella, 1983) suggest that a random simulated coenocline with also bimodal curves, and with maximum responses on the right or on the left side of the curve may simulate closely the ecological responses of the species as in a real field data. Furthermore, the results of the papers by Austin (1976a,b: 1980) and Austin & Austin (1980) suggest that the Gaussian response is rarely met even under experimental conditions.

We decided to simulate a random coenocline with 60 species and density 2. 60 is a rough approximation of the average number of relevant species found in phytosociological tables of European grassland or woodland vegetation. Density 2 has been chosen in order to avoid the possibility to have disjoint data matrices.

Since in a real situation we cannot know the position of the relevés in the ecological multidimensional space, the relevés have been selected at random along the simulated coenocline. Mohler (1983) tried to evaluate the effect of sampling pattern on estimation of species distribution along gradients by concentrating relevés in different ways along the simulated coenocline. This is a good exercise, but appliable only under limited circumstances in direct gradient analysis when at least the effects of some measurable factors can be a priori presumed.

In order to test the effects of sampling intensity and random noise on detection of species groups, the species groups defined at different level of noise and sampling intensity have been compared by the theoretical species groups. These have been defined by selecting from the coenocline sets of relevés (150, 500, 1500 and 3000 respectively) at regular intervals. Intersection analysis has been applied to the relevés and the species groups have been compared. The species group defined from 500 relevés are perfectly identical to those defined from 1500 and 3000 relevés and different kind of data (binary, cover, normalized data). This fact is important in that the sampling intensity, when high may compensate for differences in the type of data.

The data to test the effects of noise level and sampling intensity are given in tables of relevés, randomly selected from the same coenocline at three noise level, 0, 20, 40%, and 4 sampling intensity 30, 110, 200 and 300 (corresponding to four levels of sampling density 1, 4, 7, 10). The resulting species groups have been

compared through contingency tables, by the program CLACOMP (Feoli, Lagonetro & Orlóci, 1984), which decomposes the mutual information of the tables into components and finds the corresponding chi-square probabilities.

Results and conclusions

The chi-square probabilities of the comparisons between the species groups obtained on the basis of different data (binary, normalized, cover) and the theoretical species groups are presented in Table 1. Only in the case of the comparisons between the species groups obtained from 30 relevés and the theoretical species groups the probability is regurarly decreasing as the noise increases. In the other cases there is not such a clear trend. From the matrices in Table 1 we can see that in some cases the probability is increasing as the noise increases. In the cases of binary and normalized data the average values of probabilities are decreasing as the noise increases (Table 2). In the case of cover dta the probability is more or less constant. If the standard deviation of probability is considered, it is always increasing as the noise increases. It is increasing more intensively in the case of binary data and lesser in the case of cover data (Table 3).

Table 1 — Chi square probability of the comparisons between species groups obtained at different level of noise and sampling density (1, 4, 7, 10 and *T*. *T* = theoretical). *A*: Binary data; *B*: normalized data; *C*: cover data. *M* = average probability in the tables, *s* = standard deviation of m, *MT* = average probability in the comparisons with the "theoretic species groups", *sT* = standart deviation of *MT*.

Binary data:
Noise 0%:

	4	7	10	T
1	99.2	99.8	100.0	94.7
4		100.0	100.0	90.9
7			100.0	96.9
10				99.4

M = 98.09 s = 3.07 MT = 95.48 sT = 3.6

Noise 20%:

	4	7	10	T
1	99.6	94.2	98.7	92.5
4		99.8	100.0	93.5
7			99.8	66.1
10				88.9

M = 93.31 s = 10.32 MT = 82.25 sT = 12.92

Noise 40%

	4	7	10	T
1	99.4	64.8	81.6	42.5
4		100.0	100.0	99.4
7			99.8	69.5
10				90.5

M = 85.45 s = 19.94 MT = 75.48 sT = 25.31

Normalized data
Noise 0%

	4	7	10	T
1	99.9	100.0	100.0	99.6
4		100.0	100.0	99.9
7			100.0	100.0
10				99.9

M = 99.93 s = .12 MT = 99.85 sT = .17

Noise 20%

	4	7	10	T
1	100.0	100.0	100.0	99.1
4		100.0	100.0	100.0
7			100.0	100.0
10				100.0

M = 99.91 s = .28 MT = 99.78 sT = .45

Noise 40%

	4	7	10	T
1	100.0	100.0	99.9	97.3
4		100.0	100.0	100.0
7			100.0	99.9
10				99.9

M = 99.70 s = .85 MT = 99.28 sT = 1.32

Cover data:

Noise 0%

	4	7	10	T
1	94.4	99.2	99.0	93.3
4		99.5	99.2	97.8
7			100.0	96.6
10				94.7

M = 97.37 s = 2.45 MT = 95.60 sT = 1.99

Noise 20%

	4	7	10	T
1	99.8	99.7	99.8	91.8
4		100.0	99.7	92.6
7			100.0	98.2
10				97.3

M = 97.89 s = 3.13 MT = 94.98 sT = 3.24

Noise 40%

	4	7	10	T
1	100.0	96.7	99.5	86.6
4		100.0	100.0	99.8
7			100.0	95.9
10				98.7

M = 97.71 s = 4.19 MT = 95.25 sT = 5.99

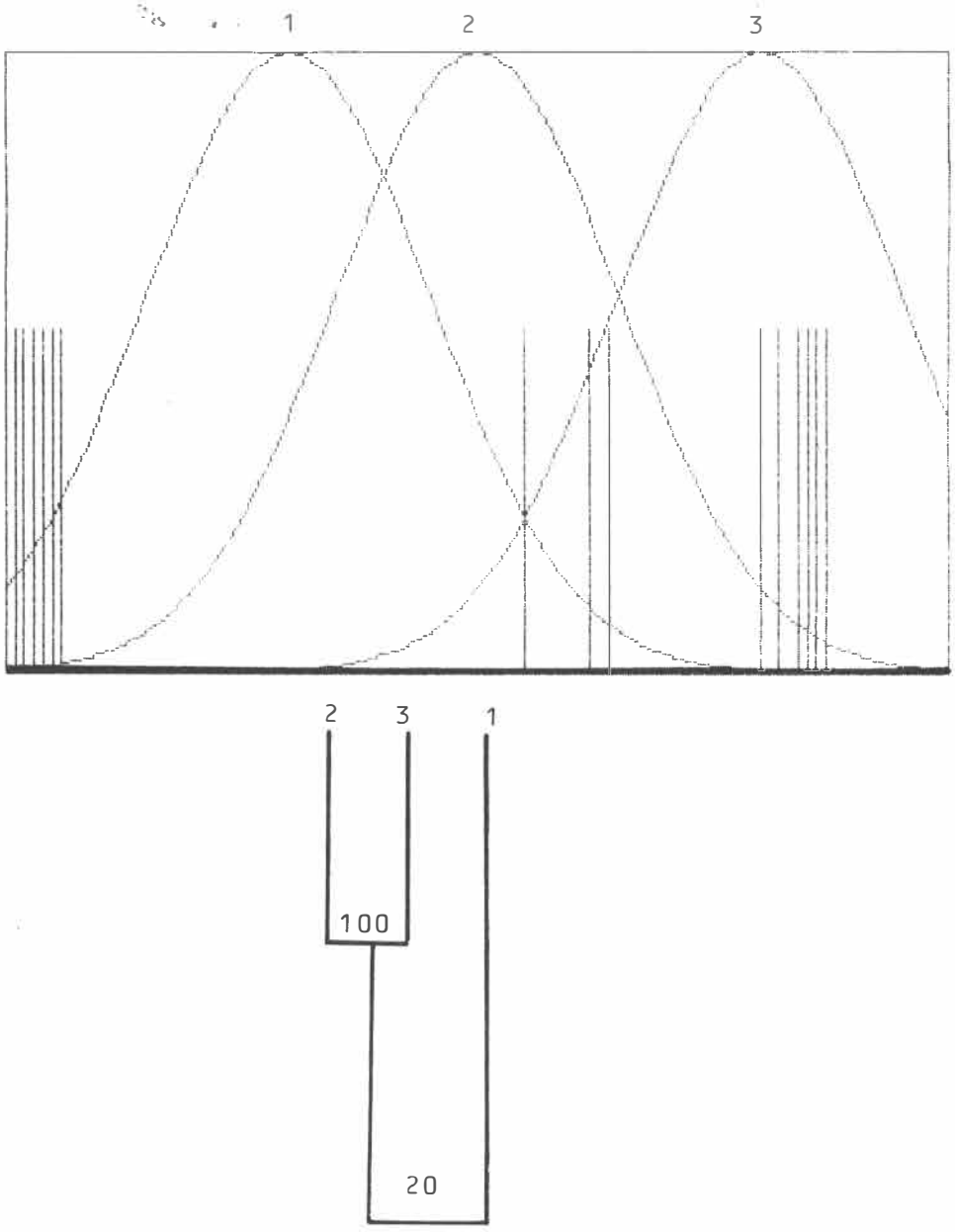

Fig. 1 — Simulated coenocline with 3 species and 15 releves irregularly sampled. On the basis of presence absence data, species 2 and 3 result identical, while it is evident from the figure that the intersection between species 1 and 2 is higher than the intersection between species 2 and 3. The dendrogram below the simulated coenocline indicates the fusion level of the three species based on the relevés marked by the vertical lines. The Jaccard coefficient is multiplied by 100. The fusion criterion is single linkage.

Table 2 — Linear regression functions between averages chi square probabilities (M and MT) and noise, and between standard deviations of M and MT and noise (n).

Binary data

$M = 98.603 - 0.316n$ $\quad s = 2.675 + 0.422n$
$MT = 95.400 - 0.500n$ $\quad sT = 3.092 + 0.542n$

Normalized data

$M = 99.962 - 0.006n$ $\quad s = 0.058 + 0.018n$
$MT = 99.920 - 0.014n$ $\quad sT = 0.075 + 0.029n$

Cover data

$M = 97.487 + 0.009n$ $\quad s = 2.390 + 0.043n$
$MT = 95.450 - 0.009n$ $\quad sT = 1.744 + 0.100n$

The species groups obtained on the basis of 30 relevés, binary and cover data are never so similar to the theoretical species groups as those based on normalized data (Table 1). From these results we can conclude that binary data are the most sensible to the noise effects. In this case the noise is mainly due to errors in species identification therefore when binary data are used a careful identification of the species is necessary. The cover data are less affected by noise and sampling intensity than binary data, however the results most close to the theoretical ones are those obtained with normalized data. If the probability of the comparisons within the same noise level and between different sampling intensity is considered, the most stable results are always obtained by the normalized data. The fact that there is not a clear improvement of the similarity with the theoretical species groups by increasing intensity of sampling is due to the fact that the relevés are not sampled on regular intervals. A possible effect of non regurarly sampling is presented in Fig. 1. This figure suggests that if we are interested in find species groups (sociological) the suggestions of Mohler (1983) are dangerous, notwistanding their usefulness in his particular circumstance.

Intersection analysis looks to be a robust technique, expecially if applied to normalized data. It can be used to detect sociological species groups and also to define the sampling intensity sufficient to give a stable description of the vegetation under study. A user can apply intersection analysis iteratively to an increasing number of relevés untill he starts to get stable species groups. By using intersection analysis the user can work with a method very little sensitive to noise and sampling density.

Riassunto. Il simulatore di cenoclini SPAGHET è stato impiegato per valutare l'efficienza della Analisi dell'Intersezione nel definire gruppi di specie a diversi livelli di "noise" e di intensità di campionamento. L'Analisi dell'intersezione si è rivelato un metodo molto affidabile, avendo dato risultati molto simili nelle diverse situazioni simulate.

Acknowledgements. The work has been supported by CNR, "Gruppo Biologia Naturalistica" and M.P.I. 40%. We are grateful to Prof. L. Orlóci for reading and correcting the text.

References

Austin M.P. (1976a) - *On non-linear species response models in ordination.* Vegetatio 33:33-41.

Austin M.P. (1976b) - *Performance of four ordination techniques assuming different non-linear species response models.* Vegetatio 33:43-49.

Austin M.P. & Austin B.O. (1980) - *Behaviour of experimental plant communities along a nutrient gradient.* J. Ecol. 68:891-918.

Doing H. (1969) - *Sociological species groups.* Acta Bot. Neerl. 18:398-400.

Ellenberg E. (1974) - *Zeigewerte der Gefaesspflanzen Mitteleuropas.* Verlag Erich Goltze KG. Gottingen.

Feoli Chiapella L. (1983) - *Prodrono numerico della vegetazione dei brecciai appenninici.* CNR, AQ/5/40. pp. 99, Roma.

Feoli E. (1977) - *A criterion for monothetic classification of phytosociological entities on the basis of species ordination.* Vegetatio 33:147-152.

Feoli E. & Lagonegro M. (1979) - *Intersection analysis in phytosociology: computer program and application.* Vegetatio 40:55-59.

Feoli E. & Lagonegro M. (1982) - *Syntaxonomical analysis of beech woods in the Appennines (Italy) using the program package IAHOPA.* Vegetatio 50:129-173.

Feoli E. & Lagonegro M. (1982) - *Syntaxonomical analysis of beech woods in the Appennines (Italy) using the program package IAHOPA.* Vegetatio 50:129-173.

Feoli E., Biondani F. & Lagonegro M. (1981) - *Individuazine di cenoclini nell'analisi indiretta di gradienti.* In A. Moroni, O. Ravera e A. Anelli (eds.) "Ecologia. Atti del primo Congresso Nazionale della Società Italiana di Ecologia", pp. 207-211, Edizioni Zara.

Feoli E., Lagonegro M. & Biondani F. (1981) - *Strategies in syntaxonomy: a discussion on two classifications of grasslands of Friuli (Italy).* In H. Dierschke ed. "Syntaxonomie", pp. 95-107. Cramer, Vaduz.

Feoli E., Lagonegro M. & Orlóci L. (1981) - *Information analysis of vegetation data.* H. Lieth & H. Mooney eds. Tasks for Vegetation Science 10. pp. 143. Junk. The Hague, Boston.

Gauch H.G. (1982a) - *Multivariate analysis in community ecology.* Cambridge University Press, Cambridge.

Gauch H.G. - *Noise reduction by eigenvector ordinations.* Ecology 63:1643-1649.

Lagonegro M. (1984) - *SPAGHET: a coenocline simulator to calibrate software tools.* Studia Geobotanica.

Hurlbert S.H. (1981) - *A gentle depilation of the nice: Dicean resource sets in resource hyperspace.* In May, R.M. (ed.) Evolutionary theory, 5: 177-184. The University of Chicago.

Landolt E. (1977) - *Okologische Zeigewerte zur Schweizer Flora.* Ber. Geobot. Inst. ETH. 64:64-207.

Mac Arthur R.H. & Levins R. (1967) - *The limiting similarity, convergence and divergence of coexisting species.* Am. Naturalist 101:377-385.

Maurer B.A. (1982) - *Statistical inference for Mac Arthur-Levins niche overlap.* Ecology 63:1712-1719.

Mohler C.L. (1983) - *Effect of sampling pattern on estimation of species distributions along gradients.* Vegetatio 54:97-102.

Orlóci L. (1978) - *Multivariate analysis in vegetatio research.* 2nd ed., Junk, The Hague, Boston.

Petraitis P.S. (1979) - *Likelihood measures of niche breadth and overlap.* Ecology 60:703-710.

Petraitis P.S.(1981) - *Algebraic and graphical relationships among niche breadth measures.* Ecology 62:545-548.

Smith E.P. (1982) - *Niche breadth, resource availability and inference.* Ecology 63:1675-1681.

Enrico Feoli & Mario Lagonegro
Dipartimento di Biologia
Università degli Studi di Trieste
34100 Trieste, Italia.