Acta Botanica Hungarica 42 (1–4), pp. 55–63, 1999/2000

USING CONCENTRATION ANALYSIS FOR OPERATING WITH INDICATOR VALUES: EFFECT OF GROUPING SPECIES

Z. BOTTA-DUKÁT[1] and E. RUPRECHT[2]

[1]*Institute of Ecology and Botany, Hungarian Academy of Sciences*
H-2163 Vácrátót, Alkotmány u. 2–4, Hungary; E-mail: bdz@botanika.hu
[2]*R-3900 Satu Mare, str. Martirilor Deportati 7, Romania*

(Received: 10 June, 1999)

Précsényi (1995) suggested the use of concentration analysis for studying the pattern of indicator values. During the analysis the researcher has to make decisions in some questions and these judgements may influence the final results. The effect of two decisions will be discussed in this paper: 1. How many species-groups to be used? 2. Which indicator values should belong to the same group? The different partitions of species, which seem to be biologically meaningful in theory, may yield different results. We concluded that the use of fewer groups may be more favourable, because the preliminary conditions of χ^2 test are fulfilled more frequently in this case. Concerning grouping, the distribution of F values in any relevé-groups should be unimodal, but in non-equilibrium cases the bimodality may be biologically meaningful.

Key words: indicator values, concentration analysis, grouping, random models

Introduction

The use of indicator values in vegetation research became widespread following Ellenberg's works (Ellenberg 1950, Ellenberg et al. 1991). Précsényi called attention several times (Zólyomi and Précsényi 1964, Précsényi 1996) to the fact that the indicator numbers are just symbols and are not real values. The equal, smaller, bigger relations on the ordinal scale could be applied, but we may not do any other mathematical operation (for example: averaging them) with these "numbers". Zólyomi and Précsényi (1964), and Précsényi (1995, 1996) gave information about the right statistical analysis with the indicator values. Some examples for the correct application of the indicator values were given in Zólyomi et al. (1988), Précsényi (1995), Borhidi and Dénes (1997), Morschhauser and Salamon-Albert (1997).

Generally, the aim of the researchers is describing the changes of indicator value patterns with as few variates as possible. The use of average indicator values may be strongly criticized from a mathematical point of

0236-6495/99/00/$ 5.00 © 2000, Akadémiai Kiadó, Budapest

view (see above). To solve this problem Précsényi (1995) suggested the use of concentration analysis. This method was developed by Feoli and Orlóci (1979) for analysing the relationship between the groups of species and the groups of relevés. Précsényi (1995) pointed out that if species are divided into groups based on their indicator values, the pattern of indicator values can be analysed by concentration analysis.

During the analysis the researcher has to make decisions which may influence the final results. The effect of the following two decisions will be discussed in this paper:

– How many species-groups to be used?

– Which indicator values should belong to the same group?

The data set that consists of relevés which were made in fen associations at the Malom Valley (near Cluj-Napoca, Romania) in different years (Ruprecht and Botta-Dukát 2000) are used to answer these questions. The groups of relevés applied in these analyses are listed in Table 1. The indicator values for moisture (W) and acidity (R) were developed by Borhidi (1995). In case of species which did not occur in Hungary the work of Sanda et al. (1983) was used.

Table 1

The data set used in this paper. All relevés were made at the Malom Valley (near Cluj-Napoca, Romania)

Groups of relevés	Association	Year	No of relevés	Source
I	CE	1940–44	11	Soó 1949
II	CE	1956	2	Pop et al. 1962
III	CE	1961	5	Pop et al. 1962
IV	CE	1998	14	Ruprecht and Botta-Dukát 2000
V	JS	1956	5	Pop et al. 1962
VI	JS	1961	2	Pop et al. 1962
VII	JS	1998	8	Ruprecht and Botta-Dukát 2000
VIII	C	1961	2	Pop et al. 1962
IX	C	1998	4	Ruprecht and Botta-Dukát 2000
X	CEcp	1956	3	Pop et al. 1962

Description of the area and associations see Ruprecht and Botta-Dukát (2000).
Abbrevations: CE = *Carici flavae-Eriophoretum latifolii*, JS = *Junco obtusiflori-Schoenetum nigricantis*, C = *Cladietum marisci*, Cecp = *Carici flavae-Eriophoretum latifolii caricosum paniceae*

Concentration analysis

The main point of the method is the following: let f_{jk} be the total number of occurrences of species belonging to species-group j in relevés belonging to relevé-group k. The value of f_{jk} depends on the sizes of species-group j and relevé-group k. Eliminating this effect the corrected values (F_{jk}) have to be used in the analysis (Orlóci and Kenkel 1985):

$$F_{jk} = \frac{f_{jk}}{p_j q_k} \cdot \frac{\sum_{g=1}^{n} \sum_{h=1}^{m} f_{gh}}{\sum_{g=1}^{n} \sum_{h=1}^{m} \frac{f_{gh}}{p_g q_h}}$$

where: n = number of species-groups, m = number of relevé-groups, p_j = number of species belonging to group j, q_k = number of relevés belonging to group k.

In the analysis we regard **F** as a contingency table although this matrix may contain fractions. First the independence of species-grouping and the grouping of relevés is statistically tested. If they are independent from each other **F** will not differ from $\mathbf{F}^0$ significantly:

$$F_{jk}^{0} = \frac{\sum_{j=1}^{n} F_{jk} \sum_{k=1}^{m} F_{jk}}{\sum_{j=1}^{n} \sum_{k=1}^{m} F_{jk}}$$

This hypothesis can be tested by χ^2 (Feoli and Orlóci 1979) or G^2 test (Précsényi 1995). If **F** significantly differs from $\mathbf{F}^0$ the matrix **F** will be analysed by correspondence analysis. The scores of species- and relevé-groups in the same min{m, n}–1 dimensional ordination space and the canonical correlation coefficients between scores are obtained this way. The sum of canonical correlation coefficients are connected with the χ^2 value computed earlier:

$$\chi^2 = F_{..} R_1^2 + F_{..} R_2^2 + \ldots + F_{..} R_S^2$$

where: S = min{m, n}–1, R_1 = canonical correlation between first scores of species-groups and first scores of relevé-groups, $F_{..}$ = the grand total of F.

The χ^2 value can be taken to components. The values of canonical correlation coefficients show the importance of the axes. If the value of $F..R_j^2$ is smaller than the appropriate critical value of χ^2 distribution with $(m-1)(n-1)$ degree of freedom, the jth axis may be left out of consideration.

The number of species-groups

The indicator values for moisture (W) are used here to demonstrate the effect of number of species-groups (on the analysis). In the case of indicator values for soil reaction (R) we got similar results.

Two different numbers of groups were compared. In the first case only the extremely small groups (WB1 and WB2) are amalgamated. So the number of species-groups was nine. In the second case the species are divided into three groups in the following way:

- xerofrequent group (WB1, WB2, WB3, WB4)
- mesofrequent group (WB5, WB6, WB7)
- hygrofrequent group (WB8, WB9, WB10)

We were showing above that the concentration analysis strongly corresponded with the χ^2 test. The preliminary condition of the test is that the empirical distribution of any F_{jk} value must not differ from the normal distribution substantially. F_{jk} has a binomial distribution with two parameters distribution p and $F..$. If $F..$ is large enough and p is not too small the binomial distribution will not differ from the normal distribution substantially. This fact is the theoretical basis for using χ^2 test (Yule and Kendall 1957). This preliminary condition of the test is likely not to be fulfilled entirely.

$F..$ was about 1000 in the data set used here. Our preliminary studies showed that if $F.. = 1000$ and $p < 0.005$ the asymmetry of binomial distribution would not be negligible. The exact value of p was not known, but $p^* = F_{jk}/F..$ is an undistorted estimation of p. When the species were divided into nine groups there were 17 cells where p* was less than 0.005. However, when only three species-groups formed the preliminary conditions of the test were fulfilled entirely in all cases.

The effect of the lack of preliminary conditions can be examined by null-models. In our case two different null-models could be used. In both cases the group memberships of species were randomized. In the first case the size of groups did not change (fixed group size method), in the other case only the number of groups was fixed (random group size method). In both cases 10,000 permutations of species-groups were made.

First the appropriate null-model had to be selected from the two possibilities discussed above. For this reason the significance levels based on hypothetical χ^2 distribution and based on randomization were compared in the case of three species-groups. Since the preliminary conditions were fulfilled entirely here, the results should not differ. The significance levels based on χ^2 distribution and randomization with random group sizes were similar. The randomization with fixed group sizes proved to be more rigorous because here an insufficient condition (size of groups) was used which diminished the number of possible different groupings. Therefore, only the random group method was used further.

When the number of groups is nine, the difference between significance levels of G^2 statistics based on randomization and hypothetical distribution were negligible despite the fact that the preliminary conditions of the test are not fulfilled entirely as seen above. But there are several differences between significance levels of canonical correlation coefficients which could be mentioned. Only the first canonical correlations were higher than the critical value based on the hypothetical distribution. Whereas the first four canonical correlations proved to be significant based on the randomization test. This means that four axes should be used, which cannot be regarded as an effective information compressing. On the other hand if only three species-groups were applied there was only one significant canonical correlation coefficient, therefore it was sufficient to interpret only the first axis. Moreover, the number of species-groups had little effect on the values of first canonical correlation coefficients. Its value was 0.4008 in the case of nine species-groups and 0.3705 in the case of three species-groups.

Shortly summarizing the main results of this section we can say that the appropriate randomization test is the random group sizes method. The fact that preliminary conditions did not fulfil entirely had only little effect on the significance level of G^2 statistics, but the robustness of the statistical test of canonical correlation coefficients was significantly smaller. The effectiveness of the method may be increased by the decrease of number of species-groups.

Which indicator values should belong to the same group?

Sometimes the answer to this question is not so trivial. For example we wanted to study R indicator values of the data set used in the previous sec-

Table 2

The F matrix of the first case in the analysis of R indicator values

	I	II	III	IV	V	VI	VII	VIII	IX	X
1	26.7	58.2	47.8	29.8	31.1	40.4	25.8	35.5	17.7	63.6
2	17.6	42.1	38.8	33.2	24.5	29.1	34.7	45.3	24.2	36.6
3	24.6	43.9	42.4	44.4	33.6	47.6	38.4	65.9	34.7	51.2

The species-groups are the following: 1. acidophilous and acidofrequent species (R3, R4, R5), 2. plants living mostly neutral soils (R6), 3. basiphilous and basifrequent species (R7, R8, R9). It can be seen that in six groups of species (I, II, III, V, VI and X) the importance values of both acidofrequent (R1) and basifrequent (R3) higher than the groups of neutral reaction indicators

tion. Based on the results of the previous section we tried to establish three species-groups. There were two solutions which were acceptable from a biological point of view. The first: 1. acidophilous and acidofrequent species (R3, R4, R5), 2. plants living mostly on neutral soils (R6), 3. basiphilous and basifrequent species (R7, R8, R9); the second: 1. extremely acidophilous species (R3, R4), 2. plants living mostly on neutral soils, including slightly acidophilous and slightly basiphilous species (R5, R6, R7), 3. extremely basiphilous species (R8, R9).

The two possibilities were compared. First it can be stated that in the first case the preliminary conditions of χ^2 test are entirely fulfilled, but in the other case there are cells whose values are smaller than the critical value. That is why the significance levels were computed by randomization test. In first case $G^2 = 27.229$, which does not prove to be significant. When the second partition of species was regarded, the G^2 value was sig-

Table 3

The F matrix of the second case in the analysis of R indicator values

	I	II	III	IV	V	VI	VII	VIII	IX	X
1	2.930	80.57	77.35	4.604	25.78	48.34	0.00	64.46	16.11	85.94
2	23.24	46.88	39.20	30.29	26.85	33.02	28.50	34.09	18.64	48.30
3	22.82	40.71	39.35	41.19	31.21	44.10	35.62	61.06	32.23	47.49

The species-groups are the following: 1. extremely acidophilous species (R3, R4), 2. plants living mostly on neutral soils including slightly acidophilous and slightly basiphilous species (R5, R6, R7), 3. extremely basiphilous species (R8, R9). In this case there are only two relevé-groups (VI, VIII), where the importance values of both extremely acidophilous (1) and basiphilous species (3) are higher than the groups of neutral reaction indicators (2)

Table 4

An artificial data matrix with 3 species-groups (a, b, c) and 3 relevé-groups (I, II, III)

	I			II			III		
	1	1	1	1	1	1	0	0	0
a	1	1	1	1	1	1	0	0	0
	1	1	1	1	1	1	0	0	0
	1	1	1	0	0	0	1	1	1
b	1	1	1	0	0	0	1	1	1
	1	1	1	0	0	0	1	1	1
	0	0	0	1	1	1	1	1	1
c	0	0	0	1	1	1	1	1	1
	0	0	0	1	1	1	1	1	1

The relevés are in the columns, the species are in the rows. The groups are separated by lines. The data matrix is well structured ($G^2 = 43.79$, $p < 0.1\%$). The bimodality in the second group yields that the two canonical correlation coefficients are equal ($R_1 = R_2 = 0.5$)

nificantly higher ($G^2 = 169{,}541$, $p < 0.1\%$). The cause of differences could be understood if the two F matrixes were examined in detail. In the case of the first partitioning of species there were six relevé-groups where both the first (acidophilous) and the third (basiphilous) groups have high values (Table 2). It can be explained first of all by the fact that sligthly acidofrequent and slightly basifrequent species can co-occur on soils with nearly neutral pH (see Table 3). Therefore, it is proper that in the second partition they are regarded as indicators of neutral pH. This example shows, that the partitioning of species, which seems biologically meaningful in theory, may prove to be inappropriate in practice.

In general, if there are many relevé-groups, where distribution of F values is bimodal (e.g. in the case of three groups both first and third F values are higher than the second) the partitioning of species is probably inappropriate. In spite of bimodality the contingency table may be well structured (e.g. Table 4). In such cases the second canonical correlation is not significantly smaller than the first, so it is not enough to use the first axis. It is connected with using correspondence analysis, which method is applicable to analyse unimodal distributions (ter Braak and Prentice 1988). In non-equilibrium cases, of course, the bimodality of distribution may be biologically meaningful, when the species of two successional phases temporarily co-occur. The two types of bimodality (biologically meaningful or consequence of wrong partitioning of the species) cannot be distinguished by any statistical test, only by the personal judgement of the researcher.

Conclusion

The partitioning of species strongly influences the final results. We recommend that if the F value is small, the number of species-groups must be decreased. Due to the decreasing number of groups the probability of fulfilling the preliminary conditions of χ^2 test increase.

If the conditions of statistical test do not fulfil the significance levels can be established by random models. We compared two possible random models and concluded that the "random group size" method is more appropriate.

If there are more than one possible partition of species with the same number of groups, the one with rare data bimodality should be chosen. In non-equilibrium communities the bimodality of distribution may be biologically meaningful. Of course, this type of bimodality should not be eliminated from the analysis.

Acknowledgements

We wish to thank Sándor Bartha, Ferenc Horváth, Miklós Kertész, Attila Borhidi and Zsolt Molnár for critically reading the manuscript. Attila Borhidi and Gabriella Pászty did much in improving the English version of the text. We are grateful to Attila Borhidi and Zsolt Molnár for making the second author's work possible in the Institute of Ecology and Botany Hungarian Academy of Sciences.

References

Borhidi, A. (1995): Social behaviour types, the naturalness and relative ecological indicator values of the higher plants in the Hungarian Flora. – *Acta Bot. Hung.* 39: 97–181.

Borhidi, A. and Dénes, A. (1997): A Mecsek és a Villányi-hegység sziklagyeptársulásai. (The rock sward of the Mecsek and Villány Mts. in South Hungary.) – In: Borhidi, A. and Szabó, L. Gy. (eds): Dissertationes in honorem jubilantis Adolf Olivér Horvát doctor academiae in anniversario nonagesimo nativitatis 1907–1997. *Studia Phytologica Jubilaria.* JPTE, Pécs.

Ellenberg, H. (1950): *Landwirtschaftliche Pflanzensociologie I: Unkrautgemeinschaften als Zeiger für Klima und Boden.* – Ulmer, Stuttgart.

Ellenberg, H., Weber, H. E., Düll, R., Wirth, V., Werner, W. and Paulissen, D. (1991): Zeigerwerte von Pflanzen in Mitteleuropas. – *Scripta Geobotanica* 18. Goltze Verlag, Göttingen.

Feoli, E. and Orlóci, L. (1979): Analysis of concentration and detection of underlying factors in structured tables. – *Vegetatio* 40: 49–54.

Morschhauser, T. and Salamon-Albert, É. (1997): Changes in composition of the acidophilous forest on the Mecsek Hills of Pécs. – In: Borhidi, A. and Szabó, L. Gy. (eds): Dissertationes in honorem jubilantis Adolf Olivér Horvát doctor academiae in anniversario nonagesimo nativitatis 1907–1997. *Studia Phytologica Jubilaria*. JPTE, Pécs.

Orlóci, L. and Kenkel, N. C. (1985): *Introduction to data analysis*. – International Cooperative Publ. House, Burtonsville, Maryland.

Pop, I., Csűrös-Káptalan, M., Ratiu, O. and Hodisan, I. (1962): Vegetatia din Valea Morii, Cluj, conservatoare de relicte glaciare. – *Contrib. Bot., Cluj*, pp. 183–204.

Précsényi, I. (1995): A homoki szukcesszió sorozat tagjai és a W indikátor számok közötti kapcsolat. (Relationship between the stages of succession series and the water indicator values.) – *Bot. Közlem.* 82: 59–66.

Précsényi, I. (1996): Az ökológiai értékszámok statisztikai feldolgozása. (Statistical analysis at ecological indicator figures.) – *Bot. Közlem.* 83: 155–157.

Ruprecht, E. and Botta-Dukát, Z. (2000): Long-term vegetation textural changes of three fen communities near Cluj-Napoca (Romania). – *Acta Bot. Hung.* 42: 265–284.

Sanda, V., Popescu, A., Doltu, M. I. and Donita, N. (1983): Caracterizarea ecologica si fitocenologica a speciilor spontane din flora României. – *Studii şi Comunicari* 25, Supliment.

Soó, R. (1949): Les associations végétales de la Moyenne, Transylvanie. – *Acta Geobot. Hung.* 6(2).

ter Braak, C. J. F. and Prentice, I. C. (1988): A theory of gradient analysis. – *Adv. Ecol. Res.* 18: 271–317.

Yule, G. U. and Kendall, M. G. (1957): *An introduction to the theory of statistics*. 14th ed. – Charles Griffin and Company Limited, London.

Zólyomi, B. and Précsényi, I. (1964): Methode zur ökologischen Charakterisierung der Vegetationseinheiten und zum Vergleich der Standorte. – *Acta Bot. Acad. Sci. Hung.* 10: 377–416.

Zólyomi, B., Précsényi, I., Bodnár, T. and Vadkerti, E. (1988): Az ökológiai indikátorszámok mintázatának változása szukcesszió alatt. (Changes in pattern of ecological indicator values during succession.) – *Bot. Közlem.* 74–75: 101–109.