ANALYSING ASSOCIATIONS AMONG MORE THAN TWO SPECIES

Z. BOTTA-DUKÁT

Institute of Ecology and Botany, Hungarian Academy of Sciences
Alkotmány u. 2-4, Vácrátót, H-2163 Hungary
(phone: +36-28-360-122; fax: +36-28-360-110)

e-mail: bdz@botanika.hu

(Received 10th Sep 2005, accepted 10th Oct 2006)

Abstract: Although the existence of higher order associations has been proved, interspecific association is generally treated as a pair-wise phenomenon. Its possible reason is that although pair-wise association is only an imperfect description of the relationships among species, its methods are simple and well known. Unfortunately, the complexity of vegetation can not be described by such simplex methods. This paper shows two methods which enable detailed analysis of higher-order associations: Juhász-Nagy's information theory functions and the log-linear contingency table analysis. From mathematical point of view, the two methods are closely related (both methods measure the non-randomness in the multi-way contingency tables). On the other hand, their theoretical backgrounds are different. The log-linear contingency table analysis was developed by statisticians to solve general statistical problems, while Juhász-Nagy's approach was developed by a biologist to solve biological problems.
The aim of this paper is to show how these two approaches decompose the total association, to point at the similarities and differences between the two approaches, and by this way to facilitate the analysis of higher order associations.
Keywords*: associatum, diversity of species combinations, pattern analysis, 3rd-order association*

Introduction

Positive and negative associations among species are very common phenomena in plant communities. Association measures the departure of frequency of species combinations from the random expectation [32]. The association among species may arise from biological interactions or different responses of species to abiotic factors [10], and on the other hand, it influences the possible interactions among species [8, 9], i.e. species can interact only where they co-occur.

In case studies, interspecific association is generally treated as a pair-wise phenomenon. Showing the structure of a community as a plexus graph or table of the significant pair-wise associations has a long tradition [1, 14, 35, 49, 53]. In addition, the matrix of pair-wise associations is often used in multivariate methods (e.g. PCA and related ordination methods [36, 43]).

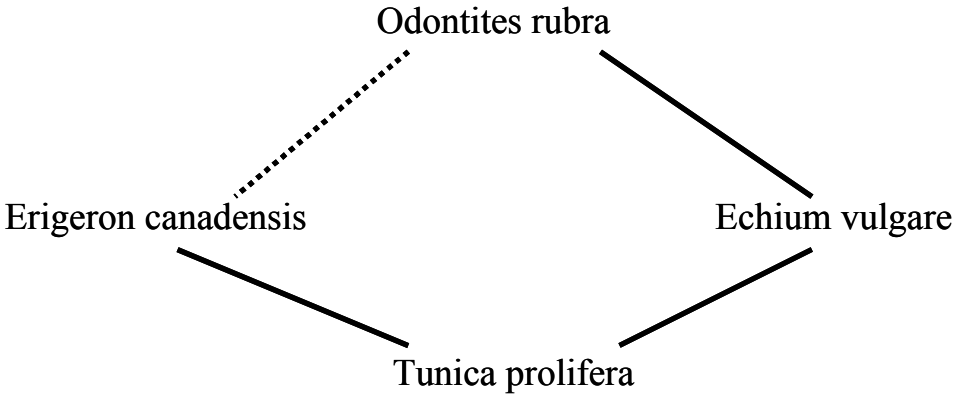
However, there may be higher-order associations among species, i.e. the presence/absence of other species can affect association between a particular pair of species [11]. The following example with an artificial data set was developed by M. Kertész (personal communication):

	plots							
species	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	0	0
B	1	1	0	0	1	1	0	0
C	0	0	1	1	1	1	0	0

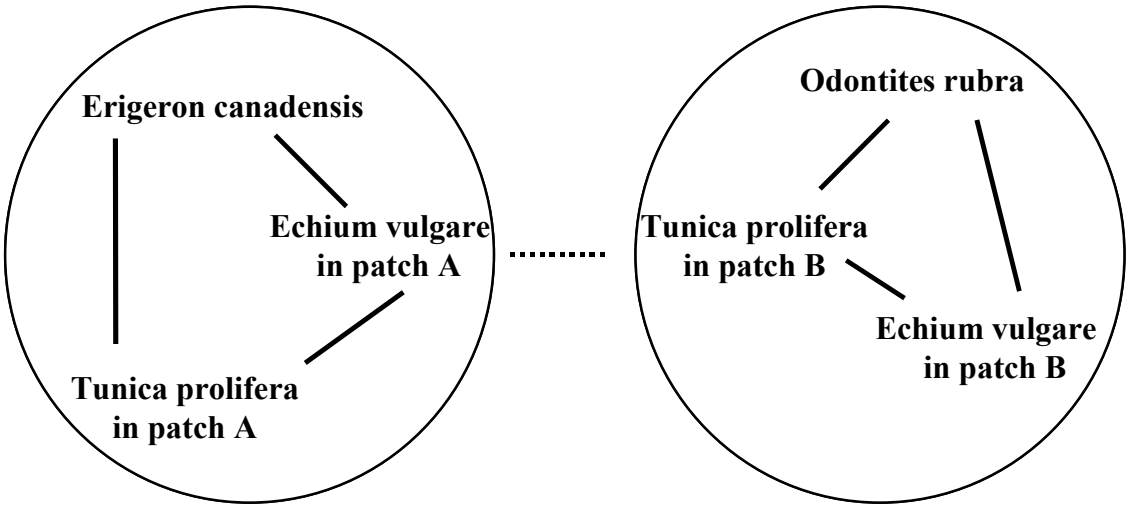
<1>

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

In this data set all pair-wise associations are zero, but any species pair determines the presence/absence of the third species, i.e. the third species occurs only in the plots where one of the other two species occurs, thus the number of species can be only 2 or 0. Disregarding higher order associations may yield underestimation of association among species or a confused network of pair-wise associations (*Fig. 1*).



a)



b)

Figure 1. *Illogical graf structure in the 19-year-old stage of primary succession on dumps of a strip coal-mine (after [2]). Pairwise associations were analysed by the standard chi-squrare procedure at 5.8 m^2 (at the area of maximum associatum). Positive associations were ploted by solid lines, negative association by dashed line. (a.) Erigeron canadensis and Odontites rubra associated negatively directly, while through Tunica prolifera and Echium vulgare they are associated positively. It seems to be paradoxical. The explanation of this paradox (Bartha S. pers. comm.), that there are three patch type (A, B, C) in the area. Erigeron lives only in A, Odontites lives only in B, Tunica and Echium live both in A and B, neither of them lives in C. It means that there are negative associations between two triplets and positive within them (b)*

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

In spite of its importance, only few attempts have been made to measure these higher-order associations. This paper considers two approaches which seem to be suitable to describe the associations among species (incl. higher order associations) in detail.

The log-linear contingency table analysis was developed for describing and testing relations among categorical variables at 1960's [5, 6, 12, 33]. It was showed in detail in many books alone [19] or as a special case of GLM [15, 34]. Statistical handbooks [48, 54] also deal with it and it is implemented by the generally used statistical programs. Fienberg [18] presented the theory and some possible biological application, and it is used by other field of biology [21]. Its application to two-dimensional contingency tables in the vegetation science (incl. describing spatial pattern of plant communities) was discussed by Feoli et collegues [17] and Orlóci [37] in detail. In spite of this fact, I found only one example [11] to use it for describing associations among more than two species, and even there only one (i.e. test of total independence hypothesis) of the numerous possibilities of this method was used. Its reason may be that biological meaning of these possibilities has not been clear yet. By other words, statistical terms are not connected to the biological terms [28, 29, 30]. Without this connection, neither mathematical method can be used in biology.

The generally used systems of biological terms are not complex enough to describe the many possible relations among species (e.g. in the community with 10 species, there are 1024 possible species combinations and the number of possible relations are even higher, because there may be associations among species combinations). Juhász-Nagy recognised this problem in the 1960's, and developed a model family on the „coexistentional structures in coenology" [23, 24, 25, 26, 27, 31]. In this model family he considered the possible types of associations and the relationships among them from biological point of view. Beyond this theoretical foundation, he proposed information theory functions to measure the different types of associations. Due to the theoretical framework, his functions are more than ad hoc indices; they are theoretically well established, their biological meaning and the relationships among them are clear. In spite of its benefits, the approach did not become widespread.

Previously I showed that from mathematical point of view these two approaches are closely related [7]. In that paper I concentrated on the three main functions of Juhász-Nagy's family (i.e. local distinctiveness, diversity of species combinations and associatum), and did not consider the partitioning the overall association of community. This paper concentrates this topic and it aims to compare how the overall association (associatum) is partitioned in these two approaches and to show similarities and differences between them.

I supposed that the readers are familiar with basics of information theory and log-linear contingency table analysis. Therefore, their terminology is used here without any detailed explanation. However, the most important information was summarised and further literatures were listed in Appendices.

Partitioning of the associatum

In the previous paper [7] I showed that the measure of associatum (the overall association in the community) in Juhász-Nagy's approach equals the half of G(A,B,...,S) statistic. In this section I compare how the associatum can be partitioned in the two

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

approaches. I do not follow the logic of either approach, I rather try to list the possible questions/problems and consider how the two approaches answer these questions.

For simplicity, I shall restrict my consideration to a community of three species (A, B, C). Let us introduce some notations. Let f_{ijk} (i = 0, 1; j = 0, 1; k = 0, 1) denote cell of three-way contingency table, e.g. f_{110} is the frequency of plots where species A and B are present and species C is absent. Let be $f_{\bullet jk} = \sum_i f_{ijk}$, $f_{i\bullet k} = \sum_j f_{ijk}$, $f_{ij\bullet} = \sum_k f_{ijk}$, $f_{i\bullet\bullet} = \sum_j \sum_k f_{ijk}$, $f_{\bullet j\bullet} = \sum_i \sum_k f_{ijk}$, $f_{\bullet\bullet k} = \sum_i \sum_j f_{ijk}$, and $f_{\bullet\bullet\bullet} = \sum_i \sum_j \sum_k f_{ijk}$.

Pair-wise association between species

The association between species A and species B is measured by G(A,B)-G(AB) in the log-linear contingency table analysis, and by nI(A,B) in Juhász-Nagy's approach. It can be showed that in the two-way contingency table G(A,B)=2nI(A,B) [7]. However, it is only the special case of the general relationship: G(A,B)-G(AB)=2nI(A,B) (i.e. in the two-way contingency table G(AB)=0).

Homogeneity of pair-wise associations

Let us consider again the example in <1>. It was mentioned above that in this community the pair-wise association between species A and B is zero. When we calculate pair-wise association, we suppose that the association between species is homogeneous, i.e. if the community is divided into parts based on the occurence other species, it is the same in the two parts. Let us divide this community into two parts based on the occurrence of species C:

	plots							
species	1	2	7	8	5	6	3	4
A	1	1	0	0	0	0	1	1
B	1	1	0	0	1	1	0	0
C	0	0	0	0	1	1	1	1

<2>

and calculate association between A and B in the two parts of the table separately. In Juhász-Nagy's approach, the two associations are $f_{\bullet\bullet 1} I(\mathrm{A,B|C}=0) = 4$ bit* and $f_{\bullet\bullet 0} I(\mathrm{A,B|C}=1) = 4$ bit, respectively. $nI(\mathrm{A,B|C}) = f_{\bullet\bullet 0} I(\mathrm{A,B|C}=0) + f_{\bullet\bullet 1} I(\mathrm{A,B|C}=1)$ is called association between A and B conditional to C or shortly, conditional association between A and B. (It should be mentioned that Juhász-Nagy [23] used partial association instead of conditional association. Because the partial association is used with different meaning in the log-linear contingency table analysis, I propose using conditional association to avoid confusion.) Conditional association between two species is higher than the association between them if the association is not homogeneous; i.e. it is not the same in the two parts of the community.

* Calculation of entropy and information in binary tables is easier with binary logarithm than natural logarithm. The values calculated by different logarithms differ only in units (viz. bit and nit, respectively).

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

On the other hand, conditional association may be lower than pair-wise association if both species A and B are strongly associated to C. For example, in the following table:

	plots							
species	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	0	0
B	1	1	1	1	0	0	0	0
C	1	1	1	1	0	0	0	0

<3>

$f_{\bullet\bullet 0} I(\mathrm{A},\mathrm{B}|\mathrm{C}=0) = f_{\bullet\bullet 1} I(\mathrm{A},\mathrm{B}|\mathrm{C}=1) = nI(\mathrm{A},\mathrm{B}|\mathrm{C}) = 0$ bit, while $nI(\mathrm{A},\mathrm{B}) = 8$ bit.

Association of species to the community

The association of species C to the community is called complete association of C [27]. Its complementary is the subassociatum of the species C, which is the associatum of the community if species C is disregarded. The sum of complete association and subassociatum of species C equals the associatum.

The complete association of species C can be calculated by another way, by multiple association between C and [A,B]. Multiple association is a generalisation of pair-wise association. The three-way contingency table can be reduced to two-way table by joining two species. For example, we join species A and B and create a new four-state variable [A,B]. Its states are 11 (both species are present), 10 (species A is present, species B is absent), 01 (species A is absent, species B is present) and 00 (both species are absent):

		[A,B]				
		11	10	01	00	Σ
C	1	f_{111}	f_{101}	f_{011}	f_{001}	$f_{..1}$
	0	f_{110}	f_{100}	f_{010}	f_{000}	$f_{..0}$
	Σ	$f_{11.}$	$f_{10.}$	$f_{01.}$	$f_{00.}$	$f_{...}$

<4>

The multiple association between C and [A,B] (the complete association of C) is measured by the information content of this new two-way table [22]:

$$nI(\mathrm{C},[\mathrm{A},\mathrm{B}]) = nH(\mathrm{C}) + nH([\mathrm{A},\mathrm{B}]) - nH(\mathrm{C},[\mathrm{A},\mathrm{B}]) \quad \text{(Eq. 1)}$$

where:

$$nH(\mathrm{A}) = n\ln n - \sum_k f_{\bullet\bullet k} \ln f_{\bullet\bullet k} \quad \text{(Eq. 2)}$$

$$nH([\mathrm{A},\mathrm{B}]) = nH(\mathrm{A},\mathrm{B}) = n\ln n - \sum_i \sum_j f_{ij\bullet} \ln f_{ij\bullet} \quad \text{(Eq. 3)}$$

$$nH(\mathrm{C},[\mathrm{A},\mathrm{B}]) = nH(\mathrm{A},\mathrm{B},\mathrm{C}) = n\ln n - \sum_i \sum_j \sum_k f_{ijk} \ln f_{ijk} \quad \text{(Eq. 4)}$$

Based on (1), (2), (3), and (4):

$$nI(\mathrm{C},[\mathrm{A},\mathrm{B}]) = n\ln n + \sum_i \sum_j \sum_k f_{ijk} \ln \frac{f_{ijk}}{f_{ij\bullet} f_{\bullet\bullet k}} = \sum_i \sum_j \sum_k f_{ijk} \ln \frac{f_{ijk}}{\left(f_{ij\bullet} f_{\bullet\bullet k}\right)/n} \quad \text{(Eq. 5)}$$

Since $(f_{ij.} f_{..k})/n$ is the expected value of the model (AB, C),

$$2nI(\mathrm{C},[\mathrm{A},\mathrm{B}]) = G(\mathrm{AB},\mathrm{C}) \quad \text{(Eq. 6)}$$

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

Third-order interaction

First, the meaning of third-order interaction has to be clarified. Let us consider the following three examples:

	plots							
species	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	0	0
B	1	1	1	1	0	0	0	0
C	1	1	1	1	0	0	0	0

<5>

	plots							
species	1	2	3	4	5	6	7	8
A	1	1	1	0	0	0	0	0
B	0	0	0	1	1	1	0	0
C	1	1	1	1	1	1	0	0

<6>

	plots							
species	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	0	0
B	1	1	0	0	1	1	0	0
C	0	0	1	1	1	1	0	0

<7>

In table <5>, all pair-wise associations are positive, and the table can be described completely by any two of them, i.e. expected values in all of the model AB,AC, the model AB,BC, or the model AC,BC equal the observed values. Thus, the third pair-wise association does not have additional information; it can be regarded as redundant.

In table <6>, the association between species A and B is negative, while species C is associated positively to both other species. Thus the third pair-wise association have additional information, it is not redundant. The association between A and C conditional to B is higher than their pair-wise association, because the occurrence of species B prevents the manifestation of the positive association between A and C, while in the absence of species B, the positive association is stronger than in the whole community. In spite of this effect, the table can be described completely by the three pair-wise associations, i.e. expected values of model AB,AC,BC equal to the observed values.

In table <7>, all the pair-wise associations are zero, but any two species determine the occurrence of the third. The table can not be described completely by the three pair-wise associations, i.e. expected values of model AB,AC,BC considerably differ from the observed values.

In the terminology of log-linear contingency table analysis, there is third-order interaction only in the third case. Its measure is G(AB,AC,BC)–G(ABC) (in the case of three species G(ABC)=0).

In Juhász-Nagy's approach, third-order interaction means that the sum of pair-wise associations differs from the associatum. Its measure is the third-order interassociatum: the difference between the sum of pair-wise associations and the associatum (Juhász-Nagy 1967a):

$$nI([\mathrm{A};\mathrm{B};\mathrm{C}]) = nI(\mathrm{A},\mathrm{B}) + nI(\mathrm{A},\mathrm{C}) + nI(\mathrm{B},\mathrm{C}) - nI([\mathrm{A},\mathrm{B},\mathrm{C}]) \qquad \text{(Eq. 7)}$$

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

Generally, the k-order interassociatum is obtained by the subtraction of overall association (k-2) times from the sum of (k-1)-way associations [26, 27]. It should be noted that -2nI([A;B;C]) equals 'RCD-interaction' in Kullback [33], and in the case of three species:

(Eq. 8)

$$nI([A;B;C]) = nI(A,B|C) - nI(A,B) = nI(A,C|B) - nI(A,C) = nI(B,C|A) - nI(B,C)$$

In table <5>, the interassociatum is positive; it indicates the redundancy among the pair-wise associations, i.e. we try to describe the fact that the three species form coalition by three pair-wise associations. On the other hand, in table <6> and <7> the interassociatum is negative. It indicates that the pair-wise associations are not homogeneous.

This different definition of third order interaction in the two approaches has historical reasons. Expected values in the model AB,AC,BC can not be calculated without iterative procedure. Although this iterative procedure was developed in 1940 [13, cit. 18], it became widely used much later: for example, Kullback [33] did not mention this procedure. Instead, he applied an inappropriate method to calculate expected values [18]. Later, this iterative procedure became widely used, and RCD-interaction was not applied by statisticians [19].

Juhász-Nagy began developing his models in the early 1960's and he applied the information statistical methods of that time (e.g. he used Kullback's [33] book). Later, he concentrated on the biological meaning of the models, rather than their refinement from mathematical point of view.

Kullback's [33] RCD interaction (it is analogous to the interassociatum) were strongly criticised [18], because it can lead to negative values for G-statistic. From mathematical point of view, the negative value is a mistake. However, in the analysis of associations among species it has meaning. Another possible drawback of interassociatum is that in a species-rich community there may be positive interassociatum among some species, while negative among others and zero in the whole community. This problem can be avoided if interassociatum is calculated not only in the whole community, but also in the sub-communities.

Comparison of the two approaches

Above I considered five topics. Two of them (pair-wise association, complete association of species) can be treated in both approaches and the result is the same (except the multiplication by two). Homogeneity of associations and the corresponding conditional association are very important in ecology. Although it is not part of the standard methodology, conditional associations can be calculated from the G-statistics, due to the additive relations between the information theory functions. The third-order interaction is defined and measured differently in the two approaches. The third-order interaction of log-linear contingency table analysis is a subset of third-order interaction of Juhász-Nagy's approach. I think *G*(AB,AC,BC) /or it analogues in the case of more species/ should be incorporated into Juhász-Nagy's methodological framework, because the distinction of two types of third-order interactions may be useful.

In this paper I concentrate on the associations among species, therefore many functions from Juhász-Nagy's model (e.g. dissociatum) are not considered here. However, it should be mentioned that in Juhász-Nagy's approach not only the

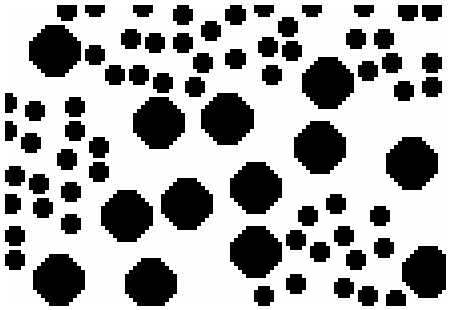
http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

dependencies among species (i.e. association), but also their independence can be measured. In the log-linear contingency table analysis, this aspect is neglected.

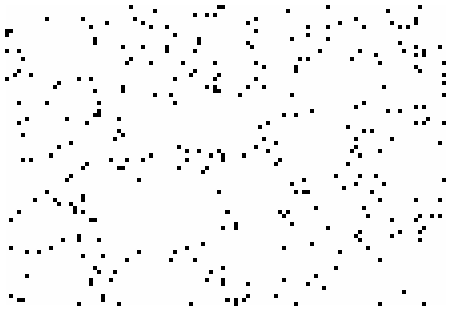
An example for analysing higher-order associations

Above, the meaning or properties of the functions are illustrated only with sporadic examples. Here I will analyse associations in an artificial community in detail. The community consists of three species: species A is a tree, species B is a tussock forming grass, and species C is an annual herb. There are only two assembly rules in this community: (1) Species C lives in the gaps among the tussocks of species B. (2) Both species B and C occur in the shadow of species A, however the tussocks of species B and the gaps among the tussocks become smaller here. In this simple case, the structure of the community seems to be clear by looking at the maps of species (*Fig. 2*). It is expected that there are significant negative association between species A and B, and the other associations are non-significant. However, it will be showed that the situation is more complex.

a) Species A



b) Species B



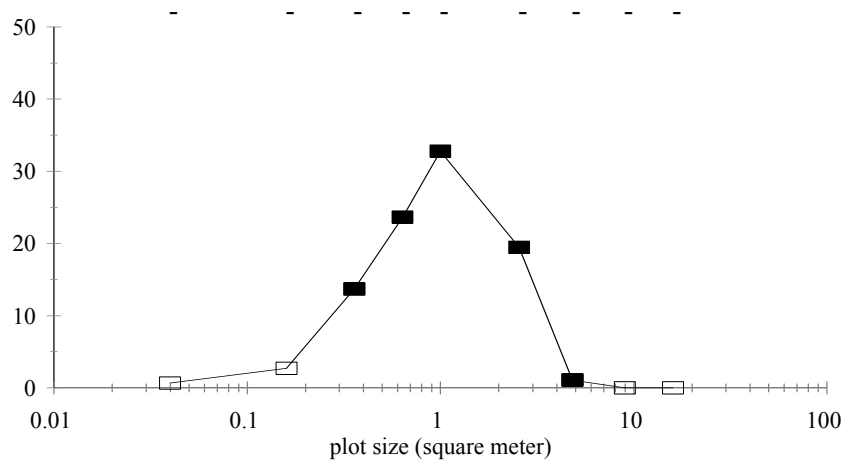
c) Species C

Figure 2. *Distribution map of the species in the artificial dataset*

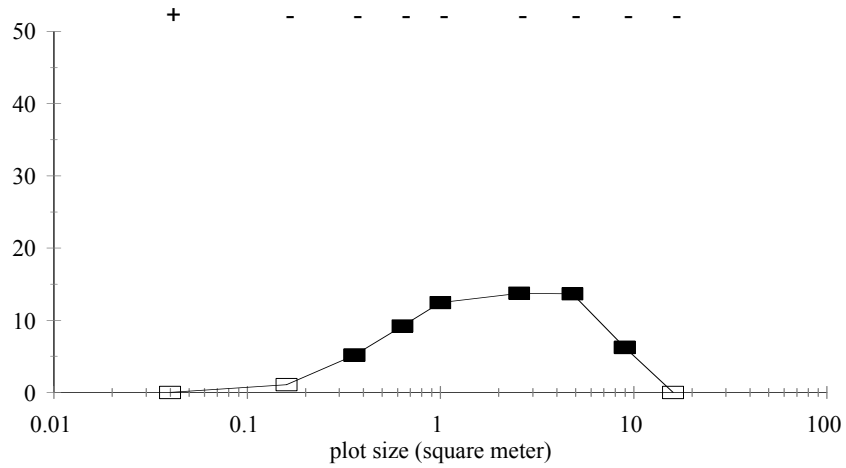
http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

The size of distribution map is 22x15 m; its resolution is 20x20 cm. Nine different plot sizes were used between 0.04 and 16 m^2. From Juhász-Nagy's functions the associatum, the pair-wise and conditional associations, the complete association of species, and the interassociatum were calculated. Beyond these functions, G(AB,AC,BC) statistic was used because it can not be calculated from Juhász-Nagy's functions. To determine their significance, observed functions were compared to functions calculated from random references produced by random shift method [4, 38]. One thousand random references were used. The sign of pair-wise associations were determined according to Bartha and Kertész [4], i.e. the association is positive if $(a+d)>(a_0+d_0)$, while it is negative if $(a+d)<(a_0+d_0)$, where a and d are observed values in the 2x2 contingency table, a_0 and d_0 are their average in the random cases.

As it is expected, there is a significant negative association between species B and C at a wide range of scales (*Fig. 3c.*).

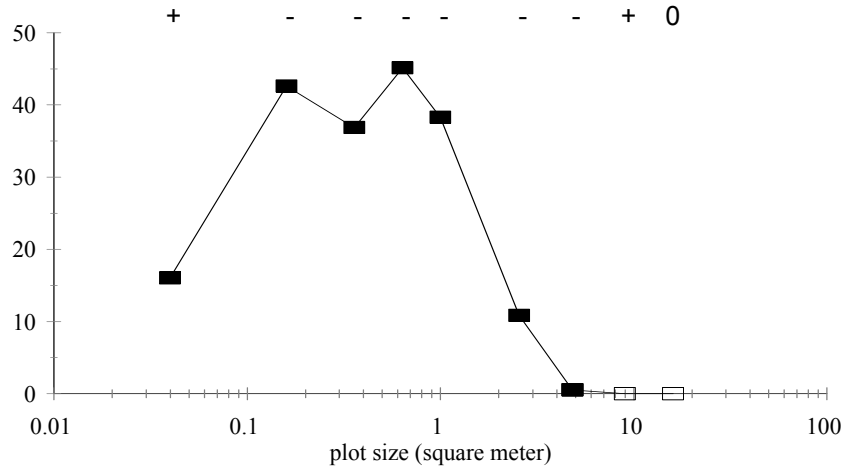


a) association between species A and B



b) association between species A and C

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623



c) association between species B and C

Figure 3. *Pairwise associations as a function of scale. Filled squares mark field data significantly different from random, empty squares field data not different from random. Direction of association are indicated at the top of figure by + or -.*

However, this association is significantly positive at 0.04 m^2. This pattern of signs is inconsistent with the pattern expected by Greig-Smith [20], and Kershaw and Looney [32], i.e. association is negative at fine scale that changes to positive at coarse scale. On the other hand, it is in line with results of Bartha and Kertész [4] who suggested that at fine scale the positive association is the consequence of high proportion of empty cells. Indeed, the detailed analysis of signs (*Fig. 4*) shows that at the smallest plot size the sign of the association is determined by the difference between the observed and expected number of empty cells (d-d_0). Thus, this positive association does not mean that at this scale the two species coexist more frequently than expected, but it means that the frequency of empty plots is higher than expected.

The other two pair-wise associations also have significant negative values at a broad range of scales (*Fig. 3a,b*). These relations can not be predicted without detailed analysis, by looking at the distribution map only. Both negative associations are the consequence of the effect of species A to the spatial pattern of species B (i.e. size and density of tussocks). Due to the smaller tussocks and smaller gaps in the presence of species A, the frequency of both species B and C is smaller here than expected.

Because of the relationships between pair-wise associations (Eq. 8), I consider in detail the homogeneity of association between species B and C only. It is selected because it is the highest from the three pair-wise associations at broad range of plot sizes (*Fig. 3*). At intermediate plot sizes, conditional association between species A and B is considerable higher than pair-wise association between them (*Fig. 5*). It suggests that the association between the two species is inhomogeneous. In the case of three species, the difference between the conditional and pair-wise association equals the interassociatum (cf. Eq. 8), which is significantly lower than the random expectation from 0.16 m^2 to 4.84 m^2 plot size (*Fig. 6*). It means that in this range the association between species B and C is inhomogeneous. The association between the two species is considerably lower and the maximum area of the association is smaller in the presence of species A than in its absence (*Fig. 7*). It is the consequence of smaller gaps and tussocks.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

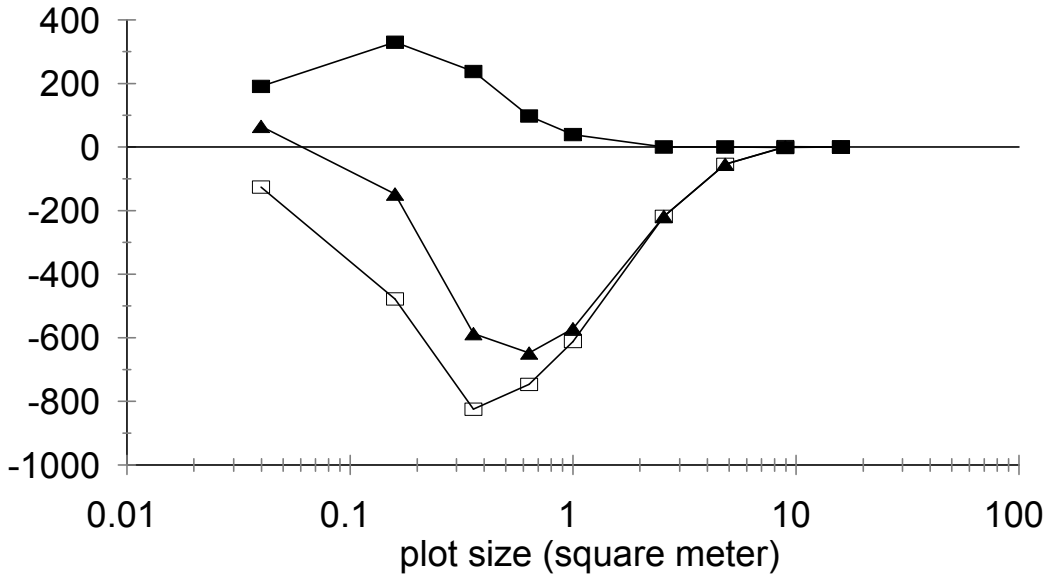


Figure 4. *Factors of sign of assoctiation between species A and B. Empty square: a-a_0, filled square: d-d_0, filled triangle: $(a+d)$-(a_0+d_0).*

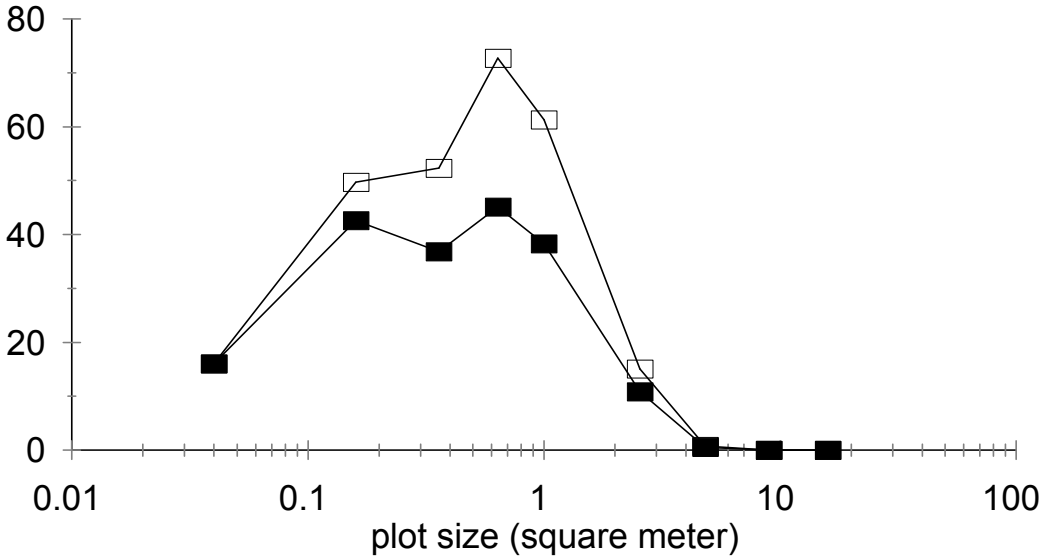


Figure 5. *Pairwise (filled square) and conditional (empty square) association between species B and C.*

G(AB,AC,BC) is significantly higher than expected in random situation between 0.16 and 1 m^2 plot size (*Fig. 8*). It indicates that at this range the structure of the community can not be described sufficiently by pair-wise associations.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

At small and intermediate plot sizes species B has the highest complete association (*Fig. 9*). The explanation may be that on one hand, the pattern of species B is affected by species A, on the other hand, species B affects the pattern of species C. Thus, pattern of this species is strongly related to the pattern of the other two species, while the pattern of species A and C are not related directly, only through species B.

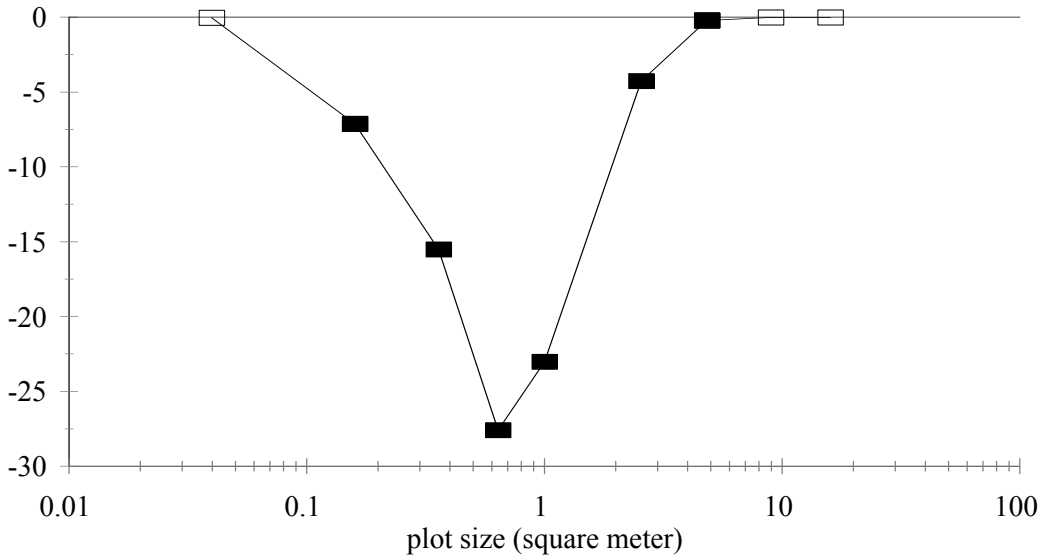


Figure 6. *Interassociatum as a function of scale. Filled squares mark field data significantly different from random, empty squares field data not different from random.*

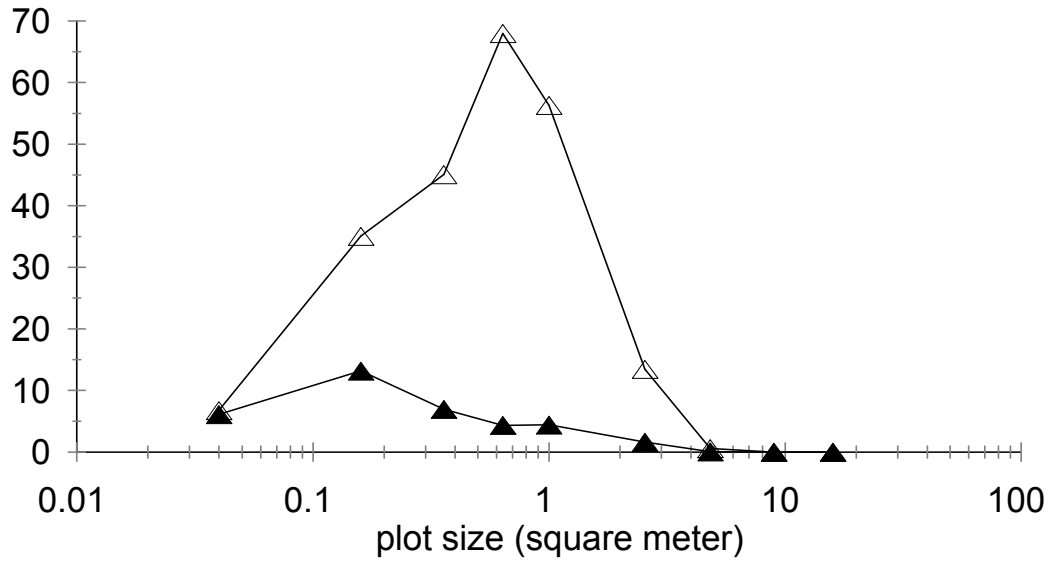


Figure 7. *Association between species B and C in the shadow of species A (filled triangle) and in the light (empty triangle).*

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

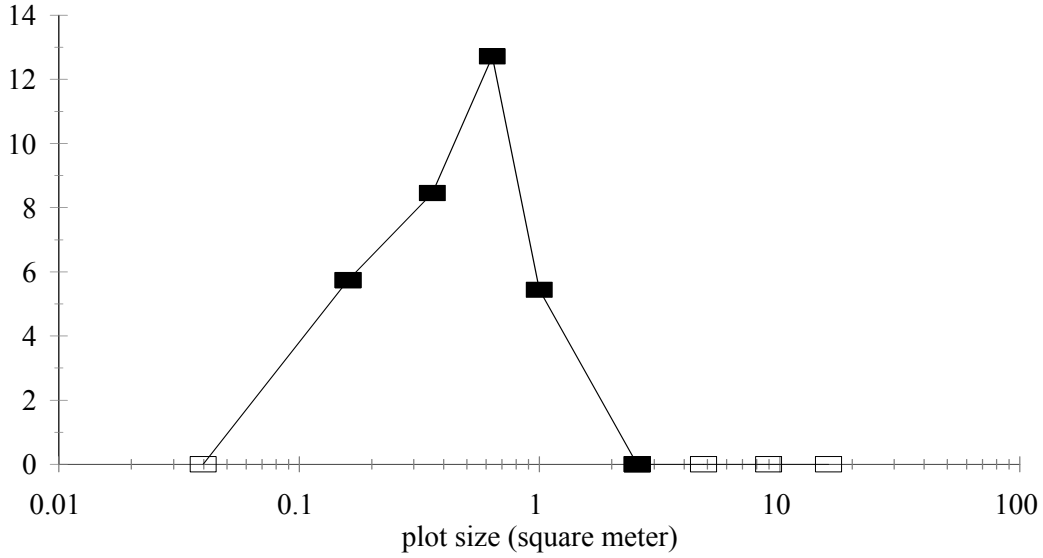


Figure 8. *G(AB, AC, BC)-statistic as a function of scale. Filled squares mark field data significantly different from random, empty squares field data not different from random, filled triangles: upper and lower bound of 95% confidence interval.*

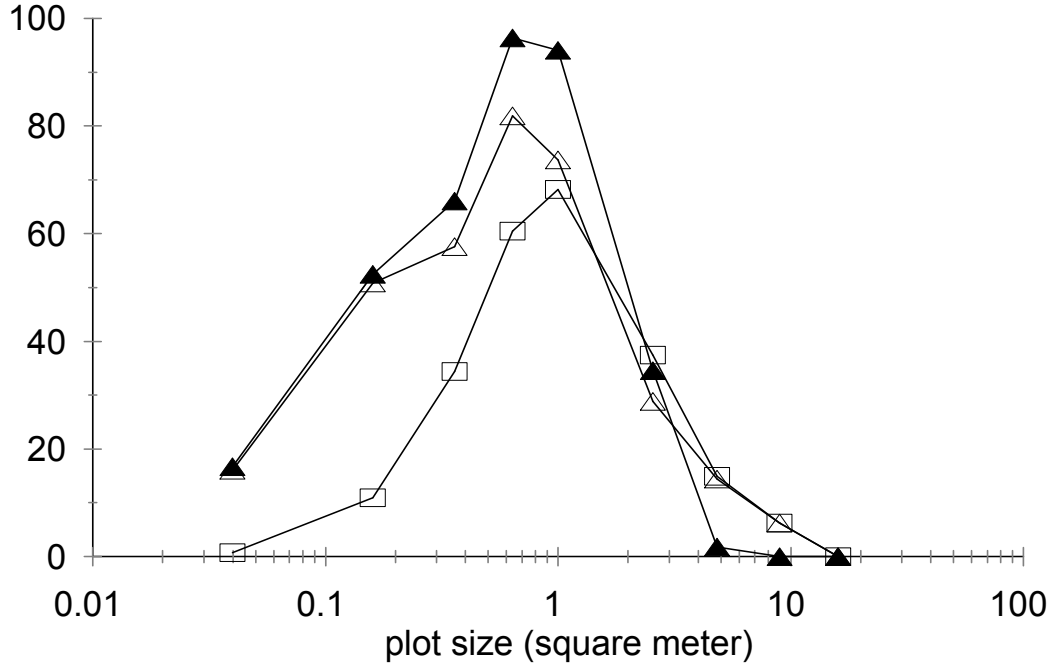


Figure 9. *Complete association of species A (empty square), B (filled triangle), and C (empty triangle).*

Even in this simple example, including only three species and two rules, associations among species can not be understood only by looking at distribution map or by calculating only pair-wise associations. It should be emphasised that in real communities the pattern is much more complex (there are more species and more rules) and they cannot be understood without detailed analyses.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

Conclusion

Although, the existence of higher order associations is theoretically justified [10], only few attempts have been made to reveal them. In case studies, association is generally treated as a pair-wise phenomenon. The possible reason is that although pair-wise association is only an imperfect description of the relationships among species, its methods are simple and well known.

Unfortunately, the complexity of vegetation can not be described by such simple methods. It is well known that vegetation is not in equilibrium [39, 40], homogeneity and stationarity criteria do not hold [9, 52], and several mechanisms (e.g. competition, propagule limitation, etc.) are acting together. However, the consequence of these facts, i.e. the complexity of vegetation, has not been generally acknowledged yet. I think the analysis of more complex relationships than the pair-wise associations should not be delayed longer.

The study of this complex relationship is impossible without solid methodological basis. The aim of both approaches, considered in this paper, are the establishment of such methodological basis. The main difference between them is that the log-linear contingency table analysis was developed by statisticians to solve general statistical problems. It can be used in vegetation science if these general statistical problems are translated into the special problems of vegetation science. This translation has not been done yet. I know only one application in this field by Dale et collegues [11], and they used only one of the numerous possibilities of this method. One aim of this paper is to begin this translation (I considered the possible relations only in the case of three species), and by this way to help to use log-linear contingency table analysis in vegetation science. However, it should be emphasised that this research field is characterised by special features [7], e.g. the importance of diversity, which can not be handled by the standard statistical methods (including log-linear contingency table analysis).

On the other hand, Juhász-Nagy's approach was developed by a biologist to solve biological problems. Therefore, although both approaches use many difficult terms, in Juhász-Nagy's approach, these terms come from the Central-European phytosociological tradition and field experiences and these terms always have biological meaning. By contrast, in the log-linear contingency table analysis terms come from the analogous statistical methods (e.g. regression analysis) and these terms have only statistical meaning. According to this advantage of Juhász-Nagy's framework, I propose to incorporate additional facilities of log-linear contingency table analysis (eg. calculation of G(AB,AC,BC) by iterative algorithm) into this framework, rather than replace this framework by loglinear contingency table analysis.

Acknowledgements. Many thanks to Sándor Bartha for his valuable comments and suggestions.

REFERENCES

[1] Agnew, A.D.Q. (1961): The ecology of Juncus effusus in North Wales. – Journal of Ecology 49: 83-102.

[2] Bartha, S. (1992): Preliminary scaling for multi-species coalitions in primary sucession. – Abstracta Botanica 16: 31-41.

[3] Bartha, S., Czárán, T., Podani, J. (1998): Exploring plant community dynamics in abstract coenostate spaces. – Abstracta Botanica 22: 49-66.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

[4] Bartha, S., Kertész, M. (1998): The importance of neutral-models in detecting interspecific spatial associations from 'trainsect' data. – Tiscia 31: 85-98.
[5] Birch, M. W. (1963): Maximum likelihood in three-way contingency tables. – J. Roy. Statist. Soc. B. 25: 220-233.
[6] Bishop, Y.M.M. (1969): Full contingency tables, logits and split contingency tables. – Biometrics 25: 383-400.
[7] Botta Dukát, Z. (2005): The relationship between Juhász-Nagy's information theory functions and the log-linear contingency table analysis. – Acta Botanica Hungarica 47: 53-73.
[8] Czárán, T. (1998): Spatiotemporal Models of Population and Community Dynamics. – Chapman and Hall, New York.
[9] Czárán, T., Bartha, S. (1992): Spatiotemporal dynamics models of plant populations and communities. – Trends in Ecology and Evolution 7: 38-42.
[10] Dale, M.R.T. (1999): Spatial pattern analysis in plant ecology. – Cambridge University Press, Cambridge.
[11] Dale, M.R.T., Blundon, D.J., MacIsaac, D.A,. Thomas, A.G. (1991): Multiple species effects and spatial autocorrelation in detecting species associations. – Journal of Vegetation Science 2: 635-642.
[12] Darroch, J.N. (1962): Interactions in multi-factor contingency tables. – J. Roy. Statist. Soc. B. 24: 251-263.
[13] Deming, W.E., Stephan, F.F. (1940): On a least square adjustment of a sampled frequency table when the expected marginal totals are known. – Ann. Math. Statist. 11: 427-444.
[14] de Vries, D.M. (1953): Objective combinations of species. – Acta Botanica Neerlandica 1: 497-499.
[15] Dobson, A.J. (2002): An introduction to generalized linear models. – Chapman and Hall/CRC, Boca Raton. 2nd ed.
[16] Erdei, Zs., Tóthmérész, B. (1993): MULTI-PATTERN 1.00. Program package to analyze and simulate community-wide patterns. – Tiscia 27: 45-48.
[17] Feoli, E., Lagonegro, M., Orlóci, L. (1984): Information analysis of vegetation data. – Dr. W. Junk, The Hague.
[18] Fienberg, S.E. (1970): The analysis of multidimensional contingency tables. – Ecology 51: 419-433.
[19] Gokhale, D.V, Kullback, S. (1978): The Information in Contingency Tables. – Marcel Dekker, Inc., New York.
[20] Greig-Smith, P. (1983): Quantitative Plant Ecology. –University of California Press, Berkeley. 3rd ed.
[21] Herben, T., Liska, J. (1988): The use of multi-way contingency tables for the study of epiphytic lichen distribution. – Coenoses 3: 135-139.
[22] Horváth, A. (1998): INFOTHEM program: new possibilities of spatial series analysis based on information theory methods. – Tiscia 31: 71-84.
[23] Juhász-Nagy, P. (1967a): On association among plant populations I. Multiple and partial association: a new approach. – Acta Biologica Debrecina 5: 43-56
[24] Juhász-Nagy, P. (1967b): On some "Characteristic areas" of plant community stands. – In Rényi, A. (ed.): Proceedings of the Colloquium on Information Theory, organized by the Bolyai Mathematical Society in Debrecen, Hungary.
[25] Juhász-Nagy, P. (1976): Spatial dependence of plant population. Part 1. Equivalence analysis (an outline for a new model). – Acta Botanica Academiae Scientiarum Hungaricae 22: 61-78.
[26] Juhász-Nagy, P. (1980): A cönológia koegzisztenciális szerkezeteinek modellezése. – Doctoral Thesis, Budapest.
[27] Juhász-Nagy, P. (1984): Spatial dependence of plant population. Part 2. A family of new models. – Acta Botanica Hungarica 30: 363-402.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

[28] Juhász-Nagy, P. (1985): Some problems of a conceptual system in biology. Part 1. Creative critical comments on a dictionary. – Abstracta Botanica 9: 33-58. (in Hungarian with English summary)
[29] Juhász-Nagy, P. (1986): Some problems of a conceptual system in biology. Part 2. Speculations on missing concepts. – Abstracta Botanica 10: 35-78. (in Hungarian with English summary)
[30] Juhász-Nagy, P. (1987): Some problems of a conceptual system in biology. Part 3. Importance of conceptual constructions. – Abstracta Botanica 11: 97-116. (in Hungarian with English summary)
[31] Juhász-Nagy, P., Podani, J. (1983): Information theory methods for the study of spatial processes and sucession. – Vegetatio 51: 129-140
[32] Kershaw, K.A., Looney, J.H.H. (1985): Quantitative and Dynamic Plant Ecology. – Edward Arnold, Baltimore. 3rd. ed.
[33] Kullback, S. (1959): Information theory and statistics. – John Wiley and Sons, New York.
[34] Lindsey, J.K. (1997): Applying generalized linear models. – Spinger, New York.
[35] Nuñez, C.I., Aizen, M.A., Ezcurra, C. (1999): Species associations and nurse plant effects in patches of high-Andean vegetation. – Journal of Vegetation Science 10: 357-364.
[36] Orlóci, L. (1978): Multivariate Analysis in Vegetation Research. – Junk, The Hague. 2nd ed.
[37] Orlóci, L. (1991): CONAPACK: program for canonical analysis of classification tables. Ecological Computations Series. Vol. 4. – SPB Acedemic Publishing bv, The Hague.
[38] Palmer, M. W., van der Maarel, E. (1995): Variance in species richness, species association, and niche limitation. – Oikos 73: 203-213.
[39] Pickett, S.T.A. (1980): Non-equilibrium coexistence of plants. – Bulletin of the Torrey Botanical Club 107: 238-248.
[40] Pickett, S.T.A., White, P.S. (eds.) (1985): The ecology of natural disturbance and patch dynamics. – Academic Press, New York.
[41] Pielou, E.C. (1969): An introduction to mathematical ecology. – Johm Wiley and Sons, New York.
[42] Podani, J. (1991): SYN-TAX IV. Computer programs for data analysis in ecology and systematics. – In: Feoli, E., Orlóci, L. (eds.): Computer Assisted Vegetation Analysis. Kluwer, The Neatherlands.
[43] Podani, J. (2000): Introduction to the Exploration of Multivariate Biological Data. – Backhuys Publishers, Leiden.
[44] Rényi, A. (1961): On measure of entropy and information. – In Neyman, J. (ed): Proceedings of the 4th Berkeley Symposium on Mathemathical Statistics and Probability. University of California Press, Berkeley.
[45] Rényi, A. (1987): A Diary on Information Theory. – John Wiley and Sons, Chichester.
[46] Shannon, C.E. (1948): A Mathematical Theory of Communication. – Bell System Technical Journal 27: 379-423, 623-656
[47] Shannon, C.E. and Weaver, W. (1999): The Mathematical Theory of Communication. – University of Illinois Press, Urbana.
[48] Sokal, R.R., Rohlf, F.J. (1981): Biometry. The priciples and practice of statistics in biological research. –Freeman, New York. 2nd eds.
[49] Soro, A., Sundberg, S., Rydin, H. (1999): Species diversity, niche metrics and species associations in harvested and undisturbed bogs. – Journal of Vegetation Science 10: 549-560.
[50] Taneja, I.J. (1989): On Generalized Information Measures and Their Applications. – Advances in Electronics and Electron Physics 76: 327-413.
[51] Taneja, I.J. (2001): Generalized Information Measures and Their Applications. – http://www.mtm.ufsc.br/~taneja/book/book.html
[52] Tilman, D. (1994): Competition and biodiversity in spatially structured habitats. – Ecology 75:2-16.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

[53] Welch, J.R. (1960): Observations on deciduous woodland in the eastern province of Tanganyika. – Journal of Ecology 48:557-573.
[54] Zar, J.H. (1999): Biostatistical Analysis. – Prentice Hall International Inc., New Jersey.

Appendix 1.: Basics of log-linear contingency table analysis

Log-linear contingency table analysis was developed to analyse the relationship among categorical variables (or by other words the structure of multidimensional contingency tables). Here I concentrate on the binary variables (species) and the corresponding binary tables only. Frequencies in the cells of such tables depend on the following effects: grand total of the table, frequency of species and associations among species. One aim of the analysis is to choose the important effects. Sets of effects are called models and each model corresponds to an expected contingency table. These expected contingency tables were calculated by an iterative procedure. The number of possible models is restricted by some constructional rules. For example, if the model contains association between species A and B, it has to contain their frequencies.

It can be tested by G-statistic whether the expected contingency table significantly differs from the observed contingency table or not. Let $G(X)$ denote the G-statistic calculated from the expected contingency table corresponding to model X. If value of $G(X)$ does not exceed the critical value (the expected contingency table does not differ significantly from the observed one), the model X contains all important effects.

Another aim of the analysis may be to measure the importance of the factor(s). The difference of G-statistics can be used for this purpose. The advantage of G-statistic over Pearson's chi-square is that the difference between two Pearson's chi-square statistics does not have chi-square distribution, while $G(X_1)$-$G(X_2)$ has approximately chi-square distribution if (and only if) X_1 contains all effects of X_2. Thus, the importance of effects also has approximately chi-square distribution.

More details can be found for example in [18, 19, 48].

There are not consistent notations of models in the literature. Here I use a simple notations, which similar to that used in the information functions. Some possible models, their notations and effects considered by them are listed below for the case of 3 species (A, B, C):

Models	considered effect
0	number of plots,
A	number of plots, frequency of species A
B	number of plots, frequency of species B
C	number of plots, frequency of species C
AB	number of plots, frequency of species A, frequency of species B, association between species A and B
AB,C	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and B
AC	number of plots, frequency of species A, frequency of species C, association between species A and C
AC,B	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and C
BC	number of plots, frequency of species B, frequency of species C, association between species B and C

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

BC, A	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species B and C
AB, AC	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and B, association between species A and C
AB, BC	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and B, association between species B and C
AC, BC	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and C, association between species B and C
AB, AC, BC	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and B, association between species A and C, association between species B and C
ABC	number of plots, frequency of species A, frequency of species B, frequency of species C, association between species A and B, association between species A and C, association between species B and C, third-order interactions among species.

Appendix 2: Entropy and mutual information

Here I consider only the simplest case: two binary variables (species), because it is enough to explain the meaning of terms. Let us assume that we study the spatial association of two species (A and B). During the sampling the presence/absence of species were recorded in n plots, and the results of sampling, i.e. the observed data, were summarised in the well-known 2x2 contingency table:

		A		
		1	0	Σ
B	1	a	b	$a+b$
	0	c	d	$c+d$
	Σ	$a+c$	$b+d$	n

Let us assume that we randomly choose one from the n plot, but before choosing we try to predict the properties of this plot. The entropies measure the uncertainty of our prediction. The entropy of species A (its symbol is H(A)) is the uncertainty of our prediction with respect to occurrence of species A. It is zero if species A is present in or absent from all plots, while it is maximal if species A is present in the half of the plots. The entropy of species B (its symbol is H(B)) can be defined by the same way.

The joint entropy of the two species is the uncertainty of our prediction about the species composition of the chosen plot. In this case, there are four possible species combinations: i.e. both species are present, A is present and B is absent, A is absent and B is present, both species are absent. Our uncertainty (the joint entropy of the two species, (its symbol is H(A,B)) is zero if the species composition of all plots is the same, while it is maximal if the frequency of all possible species combinations are equal.

Let us assume that we try predicting occurrence of species B when we know that species A is present or absent in the plot. The uncertainty of this prediction is the

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

entropy of species B conditional to species A (its symbol is $H(B|A)$). The information about the occurrence of species A can decrease our uncertainty about the occurrence of species B, if the occurrence of two species are not independent. This decrease of the uncertainty is the mutual information of the two species:

$$I(A,B) = H(A) - H(A|B) = H(B) - H(B|A)$$

The mutual information is a symmetric measure. It is zero if the occurrence of two species independent from each other, and positive otherwise. It measures the strength of association between two species. Mutual information can be calculated from the joint entropy and the two entropies:

$$I(A, B) = H(A)+H(B)-H(AB)$$

With some rearrangement of this equation we find that if the occurrence of two species are independent /I(A,B)=0/, their joint entropy equals the sum of their entropies.

Entropies can be estimated by different entropy functions (e.g. Rényi's family of entropy functions, [44]), however Shannon-function is the most commonly used [41].

Further information can be found in the following books and papers: Rényi's [45] book is an excellent introduction to the mathematics of information theory, Shannon's classical work [46, 47] explains well the logic of his entropy function, Kullback [33] treats the statistical aspects of the information theory in detail, Taneja [50, 51] gives a good overview on the properties of Shannon's entropy and the generalized entropy functions, biological applications of information theory were discussed by e.g. Feoli et collegues [17].

Appendix 3: Computer programs

The log-linear contingency table analysis is a standard statistical procedure; therefore, it is part of the widely used statistical programs (e.g. Statistica, SPSS, Statgraphics, etc.). There are some drawbacks of these programs:

- computerised sampling and randomisation have to be done with another program,
- the maximal number of species may be restricted,
- sometimes, the iterative algorithm cannot handle the contingency tables with many empty cells.

These problems should be overcome in the future.

There are many different programs to calculate Juhász-Nagy's functions [3, 16, 22, 42]. They often perform not only the calculations but the computerised sampling and randomisation too. Thus, their use seems to be more convenient than use of general statistical packages. Unfortunately, as I mentioned above, some G-statistics (first of all third-order interactions) can not be calculated from Juhász-Nagy's functions, because they are calculated by the iterative procedure.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623