THE RELATIONSHIP BETWEEN JUHÁSZ-NAGY’S INFORMATION THEORY FUNCTIONS AND THE LOG-LINEAR CONTINGENCY TABLE ANALYSIS

Z. BOTTÁ-DUKÁT

Institute of Ecology and Botany, Hungarian Academy of Sciences
H-2163 Vácrátót, Alkotmány u. 2–4, Hungary; E-mail: bdz@botanika.hu

(Received 22 April, 2004)

Spatial pattern of the vegetation, due to its complexity, cannot be described by ad hoc indices. It requires a solid methodological basis, which can describe different levels of patterns (patterns of populations, coalitions and whole community) in the same framework. There are two methods, which seem to be appropriate for this purpose: Juhász-Nagy’s information theory functions and log-linear contingency table analysis. This paper shows that from mathematical point of view they are close relatives. The main advantage of Juhász-Nagy’s approach that it is developed to describe spatial pattern of the vegetation, therefore biological meaning of the functions is a central part of the approach. Whereas log-linear contingency table analysis is a general statistical method without any special (biological) meaning of terms. On the other hand, it is a well-known statistical method, while most of the biologists are unfamiliar in information theory. The relationship between these two approaches makes it possible to hybridise their advantages.

Key words: association, diversity of species combinations, pattern analysis

INTRODUCTION

It is well known that vegetation is not in equilibrium (e.g. Pickett 1980, Pickett and White 1985), homogeneity and stationarity criteria do not hold (e.g. Czárán and Bartha 1992, Tilman 1994), 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, and often simple, ad hoc indices were used in numerical ecology.

The study of this complex relationship is impossible without solid methodological basis. Juhász-Nagy (1967a, b, 1976, 1980, 1984, 1993, Juhász-Nagy and Podani 1983) developed a model family to describe all aspects of spatial dependence of plant species within the same framework. In spite of this obvious advantage, their method has not become widely used; most of the researcher use ad hoc methods (see Dale 1999 for review of these methods). The possible reason is that the model family is more complicated than the widely
used ad hoc methods. It is more complicated, partly because it is more complex (and the complexity of vegetation can be described only by such complex methods), partly because most of the ecologist unfamiliar with its mathematical background (i.e. information theory).

The log-linear contingency table analysis was developed for describing and testing relations among categorical variables at 1960s (Kullback 1959, Darroch 1962, Birch 1963, Bishop 1969). It was shown in detail by Gokhale and Kullback (1978), and nowadays it has been treated as special case of GLM in statistical books for biologists (e.g. Lindsey 1997, Dobson 2002). Statistical handbooks (e.g. Sokal and Rohlf 1981, Zar 1999) also deal with it and it is implemented by the generally used statistical programs. Fienberg (1970) presented the theory and some possible biological application, and it is used in other field of biology (e.g. Herben and Liška 1988). The application of the log-linear analysis of two-dimensional contingency tables in the vegetation science (incl. describing spatial pattern of plant communities) was discussed by Feoli et al. (1984) and Orlóci (1991a) in detail. Thus, it could be regarded as known by ecologists. In spite of this fact, I found only one example (Dale et al. 1991) to use it for describing associations among species, and even Dale et al. (1991) used only one (i.e. test of total independence hypothesis) from the numerous possibilities of this method. Its reason may be that biological meaning of these possibilities is not clear yet.

This paper aims at showing that:

(1) From mathematical point of view, there are no substantial difference between G-statistics used in the log-linear contingency table analysis and information theory functions used by Juhász-Nagy.

(2) Due to this fact, G-statistics can be incorporated easily into the biological framework of multi-species patterns developed by Juhász-Nagy.

I hope the paper helps (1) to learn Juhász-Nagy’s models and (2) to apply log-linear contingency table analysis in the study of associations among species.

In the first part of the paper, the most important features of log-linear contingency table analysis and information theory are shortly summarised. Then I show how the associatum (the overall association) is measured in the two approaches and how the three most important functions in Juhász-Nagy’s approach are related to G-statistics. Finally, the differences between the two approaches are discussed.
G-STATISTIC AND LOG-LINEAR CONTINGENCY TABLE ANALYSIS

Although log-linear contingency table analysis is a standard statistical method, in this section I shortly summarise the most important information about it, which is enough to understand the later sections. More details can be found for example in Fienberg (1970), Gokhale and Kullback (1978), and Sokal and Rohlf (1981).

The G-statistic was proposed first by Wilks (1935) as an alternative of the classical test of independence for two-way contingency tables by Pearson’s chi-square (Legendre and Legendre 1983):

\[
G = 2 \sum_i O_i \ln \frac{O_i}{E_i}
\]

where: \(O_i\) = observed frequency in the i-th cell of two-way contingency table, and \(E_i\) = expected frequency in the i-th cell of two-way contingency table.

However, its application is not restricted to this case. G-statistic also can be used when the dimensions of a contingency table are more than two, and not only total independence of variables, but any other assumption (e.g. there are pair-wise associations among species, but there are not any higher order association) can be tested by this form. In case of simple assumptions (e.g. there are no association among species), expected values can be calculated easily from the margins of the contingency tables. Otherwise (e.g. there are associations among species) an iterative algorithm has to be used (Fienberg 1970).

In this paper, I consider only the contingency tables with binary variables (species). Value in any cell in these contingency tables depend on:

- number of plots (grand total of the contingency table),
- difference of frequencies of species from the half number of plots, and
- pair-wise and higher order associations among species.

The log-linear contingency table analysis supposes that the value in any cell is the product of these effects. Thus, the logarithm of the value in any cell is the sum of logarithm of effects. This terminology is similar to the one used in the analysis of variance (ANOVA). (Both methods are special cases of generalised linear models (cf. e.g. Lindsey 1997, Dobson 2002)).

During the analysis reference contingency tables are generated by considering different subsets of possible effects. These reference contingency tables are called models. The importance of effects can be measured by comparison of models and observed contingency tables. Let us consider the possible models in a simple example with two species (A and B):
The observed data set is the following:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Σ</td>
<td>a + c</td>
<td>b + d</td>
<td>n</td>
</tr>
</tbody>
</table>

The simplest model (it is called model 0) assumes that all species combinations are equiprobable. The expected frequencies based on this model are:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>n/4</td>
<td>n/4</td>
<td>n/2</td>
</tr>
<tr>
<td></td>
<td>n/4</td>
<td>n/4</td>
<td>n/2</td>
</tr>
<tr>
<td>Σ</td>
<td>n/2</td>
<td>n/2</td>
<td>n</td>
</tr>
</tbody>
</table>

The next model (it is called model A) assumes that frequency of species A equals the observed frequency, species B occurs in the half of plots, and the two species are independent from each other. In this case the expected frequencies are:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>(a + c)/2</td>
<td>(b + d)/2</td>
<td>n/2</td>
</tr>
<tr>
<td></td>
<td>(a + c)/2</td>
<td>(b + d)/2</td>
<td>n/2</td>
</tr>
<tr>
<td>Σ</td>
<td>a + c</td>
<td>b + d</td>
<td>n</td>
</tr>
</tbody>
</table>

Model B is analogous to the previous one: here frequency of species B equals the observed frequency, species A occurs in the half of plots, and the two species are independent from each other. The expected frequencies are:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>(a + b)/2</td>
<td>(a + b)/2</td>
<td>a + b</td>
</tr>
<tr>
<td></td>
<td>(c + d)/2</td>
<td>(c + d)/2</td>
<td>c + d</td>
</tr>
<tr>
<td>Σ</td>
<td>n/2</td>
<td>n/2</td>
<td>n</td>
</tr>
</tbody>
</table>
Model A and model B differ from model 0 in that it considers frequency of either species. Model AB combines assumptions of the previous two models: both species frequencies equal the observed frequencies and species are independent from each other. It is the most commonly used assumption in the two-way contingency tables. The expected frequencies are:

The last model (model AB) contains all effects: frequencies of both species and the association between them. In the two-way contingency table, its expected values equal the observed values. Feoli et al. (1984) and Orlóci (1991a) discussed the application of the above-mentioned models for $m \times n$ contingency tables in the vegetation science, in detail.

In the case of more than two species, the logic of model building is similar. The models consider first the species frequencies, then the pair-wise associations, then the third-order associations and so on. Table 1 shows the possible models and the considered effect(s) in the case of three species.

As the number of dimensions of the contingency table (the number of species) increases, the number of effects (first of all, the number of associations) increases exponentially. For example, in the case of ten species there are 45 species pairs, 120 species triplets and so on; there are 1024 possible effects. One aim of the analysis is the choice of important effects. As it was showed above, each model corresponds to an expected contingency table. It can be tested by G-statistic, whether this expected contingency table significantly differs from the observed one or not. Let $G(X)$ denote the G-statistic calculated from the expected contingency table corresponding to model X. For example, the $G(A,B)$ is calculated from the expected values in table <5>:

$$G(A,B) = 2 \left[ a \ln \frac{a}{(a+b)(a+c)/n} + b \ln \frac{b}{(a+b)(b+d)/n} + 
+ c \ln \frac{c}{(a+c)(c+d)/n} + d \ln \frac{d}{(b+d)(c+d)/n} \right]$$

(2)

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. For example, in the case of two species, if $G(A,B)$
does not exceed the critical value, there is no significant association between species A and B.

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. For example, the importance of association among A and B is measured by $G(A,B) - G(AB)$, and the importance of pair-wise associations in the case of three species is measured by $G(A,B,C) - G(AB,AC,BC)$ (for the meaning of notations, see Table 1).

Table 1
Possible models and considered effects in the log-linear contingency table analysis in case of 3 species (A, B, C)

<table>
<thead>
<tr>
<th>Models</th>
<th>Considered effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>number of plots,</td>
</tr>
<tr>
<td>A</td>
<td>number of plots, frequency of species A</td>
</tr>
<tr>
<td>B</td>
<td>number of plots, frequency of species B</td>
</tr>
<tr>
<td>C</td>
<td>number of plots, frequency of species C</td>
</tr>
<tr>
<td>A, B</td>
<td>number of plots, frequency of species A and B</td>
</tr>
<tr>
<td>AB</td>
<td>number of plots, frequency of species A, frequency of species B, association</td>
</tr>
<tr>
<td></td>
<td>between species A and B</td>
</tr>
<tr>
<td>AB, C</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and B</td>
</tr>
<tr>
<td>AC</td>
<td>number of plots, frequency of species A, frequency of species C, association</td>
</tr>
<tr>
<td></td>
<td>between species A and C</td>
</tr>
<tr>
<td>AC, B</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and C</td>
</tr>
<tr>
<td>BC</td>
<td>number of plots, frequency of species B, frequency of species C, association</td>
</tr>
<tr>
<td></td>
<td>between species B and C</td>
</tr>
<tr>
<td>BC, A</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species B and C</td>
</tr>
<tr>
<td>AB, AC</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and B, association between species A</td>
</tr>
<tr>
<td></td>
<td>and C</td>
</tr>
<tr>
<td>AB, BC</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and B, association between species B</td>
</tr>
<tr>
<td></td>
<td>and C</td>
</tr>
<tr>
<td>AC, BC</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and C, association between species B</td>
</tr>
<tr>
<td></td>
<td>and C</td>
</tr>
<tr>
<td>AB, AC, BC</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and B, association between species A</td>
</tr>
<tr>
<td></td>
<td>and C</td>
</tr>
<tr>
<td></td>
<td>association between species B and C</td>
</tr>
<tr>
<td>ABC</td>
<td>number of plots, frequency of species A, frequency of species B, frequency of</td>
</tr>
<tr>
<td></td>
<td>species C, association between species A and B, association between species A</td>
</tr>
<tr>
<td></td>
<td>and C</td>
</tr>
<tr>
<td></td>
<td>association between species B and C</td>
</tr>
<tr>
<td></td>
<td>third-order interactions among species.</td>
</tr>
</tbody>
</table>
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.

**ENTROPY AND INFORMATION**

Juhász-Nagy’s approach uses information theory functions (e.g. entropy, joint entropy, information, and conditional entropy). In this section, I show the meaning of these functions considering the simplest case: two binary variables (species) only.

Further details can be found in the following books and papers: Rényi’s (1987) book is an excellent introduction to the mathematics of information theory, Shannon’s classical work (Shannon 1948, reprinted in Shannon and Weaver 1999) explains well the logic of his entropy function, Kullback (1959) treats the statistical aspects of the information theory in detail, Tanaka (1989, 2001) gives a good overview on the properties of Shannon’s entropy and the generalized entropy functions, biological applications of information theory were discussed, e.g. by Feoli et al. (1984) and Orloci (1991a, b).

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 $2 \times 2$ contingency table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>0</td>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>$a + c$</td>
<td>$b + d$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

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 predictions. 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 can be defined by the same way.

The joint entropy of the two species (its symbol is $H(AB)$) is the uncertainty of our prediction about the composition (i.e. species combination) 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) 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 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:

$$I(A, B) = H(A) - H(A|B) = H(B) - H(B|A)$$  (3)

The mutual information is a symmetric measure, as it is shown in equation (3). 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.

Figure 1 demonstrates the means of these terms by Venn-diagrams. It clearly shows that mutual information can be calculated from the entropies of species A and B, and from their joint entropy:

$$I(A, B) = H(A) + H(B) - H(AB)$$  (4)

With some rearrangement of equation (4) 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, Rényi 1961), however Shannon-function is the most commonly used (Pielou 1969):

$$H(A) = -rac{a+c}{n} \ln \frac{a+c}{n} - \frac{b+d}{n} \ln \frac{b+d}{n} =$$

$$= \ln n - \frac{1}{n} \left( (a+c) \ln(a+c) + (b+d) \ln(b+d) \right)$$

$$H(B) = -rac{a+b}{n} \ln \frac{a+b}{n} - \frac{c+d}{n} \ln \frac{c+d}{n} =$$

$$= \ln n - \frac{1}{n} \left( (a+b) \ln(a+b) + (c+d) \ln(c+d) \right)$$  (5)

* For simplicity, notations of entropy and its estimation are the same here.
Fig. 1. Demonstration of meaning the information theory functions in the case of two species by Venn-diagrams. Areas of circles represent our uncertainty in the prediction of occurrence species A and B, respectively. Patterned areas show the amount of information theory measures. A = entropy of species A; B = entropy of species B; C = joint entropy of two species; D = conditional entropy of species A to species B; E = conditional entropy of species B to species A; F = mutual information of the two species.
The forms become simpler if both sides of the equation are multiplied by \( n \). Therefore in the Juhász-Nagy’s approach these (so-called weighted) forms are used. For example:

\[
\text{nl}(AB) = n \ln n + a \ln a + b \ln b + c \ln c + d \ln d - (a + c) \ln(a + c) - (b + d) \ln(b + d) - (a - b) \ln(a + b) - (c + d) \ln(c + d)
\]  

or

\[
nl(A, B) = a \ln \frac{a}{(a + b)(a + c)} + b \ln \frac{b}{(a + b)(b + d)} + c \ln \frac{c}{(a + c)(c + d)} + d \ln \frac{d}{(b + d)(c + d)}
\]  

It can be seen that \( nl(A, B) \) in equation (9) is the half of \( G(A, B) \) in equation (2). It will be showed below that G-statistics (or difference between G-statistics) equal to information theory functions multiplied by two in many cases. This multiplication by two is essential to use \( \chi^2 \)-distribution to establish critical values (cf. Kullback 1959, pp. 113–114).

**MEASURING OVERALL ASSOCIATION (ASSOCIATUM) IN THE TWO APPROACHES**

In multi-species communities, there are many different associations among species: e.g. pair-wise associations, associations among species triplets and so on. To handle these many associations an operative terminology had to be developed (Juhász-Nagy 1967a, 1980). Here we concentrate on the most important of the possible associations, i.e. the overall association in community. In Juhász-Nagy’s terminology (Juhász-Nagy 1976, 1980, 1984, Juhász-Nagy and Podani 1983) it is called associatum. In this section I compare how the two approaches measure the associatum in multidimensional contingency tables.

**Log-linear analysis**

Let us consider a community with \( s \) species: A, B, C, ..., S. In the previous section it was showed that model A, B, C, ..., S contains the number of plots

\[
H(AB) = -\frac{a}{n} \ln \frac{a}{n} - \frac{b}{n} \ln \frac{b}{n} - \frac{c}{n} \ln \frac{c}{n} - \frac{d}{n} \ln \frac{d}{n}
\]  

\[
= \ln n - \frac{1}{n} (a \ln a + b \ln b + c \ln c + d \ln d)
\]  

(7)
and frequencies of species, but does not contain any association. Therefore, the corresponding expected contingency table differs from the observed one due to associations among species. Thus, \( G(A, B, C, \ldots, S) \) measures the associatum.

**Juhász-Nagy’s approach**

The pair-wise association between two species (A and B) can be measured by their mutual information (e.g. Juhász-Nagy 1967, Legendre and Legendre 1983). The computation of mutual information can be generalised for more than two species by the following simple way (Juhász-Nagy and Podani 1983): mutual information of species = sum of their entropies – their joint entropies.

The three most important functions describing overall (community level) patterns of Juhász-Nagy are the associatum (\( AC \)), the local distinctiveness (\( LD \)), and the diversity of species combinations (\( DSC \)). They can be measured by mutual information of species, sum of species’ entropies, and joint entropy of the species, respectively. These functions measure the similarity among the species behaviour (\( AC \)), the diversity of species coexistence (\( DSC \)), and the local preference of species (\( LD \)).

Let \( n_A = \) the number of plots where species A is present and \( n - n_A = \) the number of plots where species A is absent. Weighted entropy of species A can be estimated by

\[ nH(A) = n \ln n - n_A \ln n_A - (n - n_A) \ln(n - n_A) \quad (10) \]

Let us assume that the community consists of \( s \) species: A, B, C, \ldots, S. The local distinctiveness is measured by the sum of entropy estimation of all species:

\[ LD = nH(A) + nH(B) + \ldots + nH(S) \quad (11) \]

Diversity of species combinations is measured by the entropy of species combinations, which equals the joint entropy of all species:

\[ DSC = nH(AB\ldots S) = n \ln n - \sum f_i \ln f_i \quad (12) \]

where: \( f_i = \) frequency of \( i \)-th species combinations.

In the previous section I showed that if the species are independent, their joint entropy (=\( DSC \)) equals the sum of their entropies (=\( LD \)), while otherwise joint entropy is smaller than the sum of entropies. It means that \( LD \) can be interpreted as the possible maximum of \( DSC \) with fixed species frequencies. \( DSC \) equals \( LD \) if species are spatially independent.
The measure of associatum is calculated from the other two quantities:

\[ AC = LD - DSC \]  

(13)

**Relationship between G-statistics and Juhász-Nagy’s functions**

In this section I will show that \[ G(A, B, C, \ldots, S) = 2AC \]. It is not my discovery. It is known in the literature of information theory (see e.g. Kullback 1959) that overall information of the contingency table multiplied by two equals G-test of total independence of variables. In spite of this fact, I think it is useful to show this relationship in our special case (i.e. binary contingency tables of species) in detail.

Let us introduce two notations:

- \( f_i \) = observed frequency of the \( i \)-th species combination
- \( F(X)_i \) = expected frequency of the \( i \)-th species combination in model \( X \)

By these marks the form of G-statistic is:

\[ G(X) = 2 \sum_i f_i \ln \frac{f_i}{F(X)_i} \]  

(14)

**Diversity of species combinations.** In the log-linear contingency table analysis the simplest model is the model 0 where all species combinations are equiprobable:

\[ F(0)_i = \frac{n}{2^s} \]  

(15)

where: \( n \) = number of plots, \( s \) = number of species

The corresponding G-statistic is:

\[ G(0) = 2 \left( \sum_i f_i \ln f_i - \sum_i f_i \ln \frac{n}{2^s} \right) = \]  

\[ = 2 \left( \sum_i f_i \ln f_i - n \ln n + sn \ln 2 \right). \]  

(16)

and the diversity of species combination is:

\[ DSC = n \ln n - \sum_i f_i \ln f_i, \]  

(17)

thus:

\[ G(0) = 2(sn \ln 2 - DSC). \]  

(18)
If the number of species and the number of plots are constant, the maximum of $DSC$ equals $sn \ln 2$. Thus, $G(0)$ measures the difference between actual value of $DSC$ and its potential maximum, multiplied by 2.

**Entropy of species and local distinctiveness.** Let us now consider the model $A$. In this case, the expected frequency of species $A$ is $n_A$ (equals the observed frequency), while expected frequency of all other species is $n/2$. There are $2^{S-1}$ species combinations that contain species $A$. In this model, they are equiprobable and their expected frequency is $\frac{n_A}{2^{S-1}}$. Similarly, there are $2^{S-1}$ species combinations which do not contain species $A$ and their expected frequency is $\frac{n-n_A}{2^{S-1}}$.

The importance of effect $A$ is measured by $G(0)-G(A)$ statistic:

$$G(0) - G(A) = 2 \sum_i f_i \ln \frac{F(A)_i}{F(0)_i} = 2 \left( n_A \ln \frac{2n_A}{n} + (n - n_A) \ln \frac{2(n - n_A)}{n} \right) =$$

$$= 2 \left[ n_A \ln n_A + (n - n_A) \ln(n - n_A) - n \ln n + n \ln 2 \right]$$

(19)

The weighted entropy of species $A$ is:

$$nH(A) = n \ln n - n_A \ln n_A - (n - n_A) \ln(n - n_A)$$

(20)

Its maximum is $n \ln 2$. Based on equations (19) and (20):

$$G(0) - G(A) = 2 \left[ n \ln 2 - H(A) \right]$$

(21)

Thus, $G(0)-G(A)$ statistic measures the difference between entropy of species $A$ and its maximum, multiplied by two.

Let us calculate and sum both sides of equation (21) for all species:

$$[G(0) - G(A)] + [G(0) - G(B)] + \ldots + [G(0) - G(S)] =$$

$$= 2 \left[ sn \ln 2 - H(A) - H(B) - \ldots - H(S) \right]$$

(22)

In the right side of equation (22), the sum of entropies can be replaced by $LD$. The left side of this equation can be simplified using the following properties of $G$-statistics:

$$[G(0) - G(A)] + [G(0) - G(B)] + \ldots + [G(0) - G(S)] = G(0) - G(A, B, \ldots, S)$$

(23)

Thus, we get the following equation:

$$G(0) - G(A, B, \ldots, S) = 2 \left( sn \ln 2 - LD \right)$$

(24)
It means that $G(0) - G(A, B, \ldots, S)$ statistic measures the difference between $LD$ and its maximum, multiplied by 2.

Associatum. As I mentioned above, associatum is measured by the information content of the $s$-dimensional contingency table, which can be calculated as the difference between $LD$ and $DSC$ (see equation 13).

Based on equations (13), (18) and (24):

$$2AC = G(A, B, \ldots, S)$$

(25)

**DISCUSSION: COMPARISON OF APPROACHES**

In this paper I show that in the two approaches associatum is measured by the same way (the multiplication by two does not change the tendencies). However, in addition to the similarities in the mathematical forms, the differences between the two approaches should be emphasised.

**General (statistical) vs. special (biological) models**

Any statistical method is developed to solve more or less general problem not an individual problem, even if researcher intended to solve an individual problem (e.g. Student’s $t$-test was developed to compare averages of two samples with normal distribution and not to compare, say, average alcohol content of two cask of beer). Consequently, statistical methods use general terms (e.g. means rather than means of something). When we use these methods, we have to correspond our special terms (e.g. average height of plants in samples from the two studied populations) to the general terms (e.g. averages of samples from two populations* with normal distribution).

Statistical methods consider the properties of studied phenomenon which important from statistical point of view and neglect all other properties. The advantage of this generalization is that the number of commonly used statistical methods is not large, while the number of biological problems can be solved by them is infinite. However, the properties neglected by the statistical methods may be important during the interpretation of results. Biological methods consist of the related statistical method, rules for relating biological and mathematical terms and rules for interpretation of results.

These rules may be so simple or self-evident that researcher uses them unconsciously. In the case of spatial pattern analysis these rules are complicated due to spatial constraints. Therefore they differ from the other applica-

---

* It should be noted that in the previous example “population” is used in biological sense, while here it is used in statistical sense.
tions of statistics. Juhász-Nagy built up a model to describe spatial pattern of vegetation. The main advantage of his model is that it can describe all possible spatial dependence among species and not only some aspects of the spatial pattern. He used information statistics (Kullback 1959) as statistical method and accomplished the corresponding biological method. During this work he connected mathematical terms with biological ones (e.g. mutual information between two binary variable is a mathematical term, while pair-wise association between two species is the corresponding biological one) and created rules for biological interpretation of results (cf. marginal idealisations in Juhász-Nagy 1976).

The log-linear contingency table analyses is a mathematical method, which can be used to solve various problems, among others describing spatial dependence of species. However, as mentioned above, before biological application, the relationship between mathematical and biological terms, and the rules for biological interpretation of results have to be revealed. In the case of describing spatial dependence of species, these relationships and rules are neither simple nor self-evident. Fortunately, the relation between information statistics and log-linear contingency table analyses (the latter was evolved from the former) make it possible to use Juhász-Nagy’s and his followers’ results (e.g. Juhász-Nagy 1976, 1980, 1984, Juhász-Nagy and Podani 1983, Bartha 1992, Tóthmérész and Erdei 1992, Bartha et al. 1997, 1998, Horváth 1998). Based on these results, in the next section we shortly point out the main differences between the analysis of spatial pattern and other applications of log-linear contingency table analysis.

**Special features of pattern analysis of vegetation**

Beyond to above-mentioned differences between the two approaches, it should be noted that analysis of co-existential relations among species differs from other applications of log-linear contingency table analysis in that not only associations, but also diversity of species combinations are of interest. In other applications of log-linear contingency tables, researchers are often interested in the relationship among the variables only (distribution of one or more variables is often controlled by the researcher). By contrast, in plant communities the diversity plays a central role (c.f. relationships among diversity and ecosystem functions; see reviews in Johnson et al. 1996, Tilman 1999, Loreau et al. 2001). In these studies diversity most often means species-abundance diversity or number of species, but Bartha et al. (1997) pointed out that DSC is more appropriate measure of diversity than the species-abundance diversity.

Interactions among species constrain the possibilities of their coexistence. Log-linear contingency table analysis concentrates on this limitation, the dif-
ference between the potential and the actual value. However, in the biology
the potential diversity of coexistence (\(LD\)), the actual diversity of coexistence
\((DSC)\), and the difference between them \((AC)\) are all important.

This fact is generally true for the analysis of any community, but vegeta-
tion (and communities of other sessile organs) has other important features. Because of the sessile life form of plants, and limitation of the dispersal of their
offspring, interactions are always local (Harper 1977, Czárán and Bartha 1992).
Due to the considerably different abundances (cf. Tóthmérész and Erdei 1992)
and size of individuals (e.g. size of the clonal genet can considerably exceed
the size of corresponding aclonal individual; Oborny and Bartha 1995), the dif-
ferent species has different spatial grain. Consequently, there is not a single
relevant scale in the community; the detectable relationships depend on the
plot size (Podani et al. 1993).

Juhász-Nagy (1967b, 1976, 1984, 1993) strongly emphasised that associa-
tions among species depended on plot size. Therefore, the values of his func-
tions (or \(G\)-statistics) should be calculated at many plot sizes (Podani et al.
1993). In this respect, analysis of associations among plant species departs
from the other applications of log-linear contingency table analysis, where
only one table (representing one scale) is analysed. This scale-dependence
of associations is often neglected. For example, Dale et al. (1991) considered only
one plot size in the analysis of associatum, although they showed the
scale-dependence of pair-wise associations in an earlier analysis of the same
data set (Dale and Blundon 1991).

Plots with different sizes can be sampled in the field, but it can be done
more easily in two steps: first in the field the vegetation is recorded (e.g. maps,
point patterns, grids or transects), and then this primary reference (sensu

Juhász-Nagy (1967b, 1976, 1984) suggested that the calculated values
should be plotted against plot size, and the shapes of the curves should be
compared, not only values at an arbitrarily chosen plot size. The typical shape
of the curves is showed in Figure 2. It can be characterised by maximum of the
function \((V_{\text{max}})\), the plot size where maximum is observed \((A_{\text{max}})\), and the plot
size where the value decreases to zero \((A_{\text{min}})\) (Juhász-Nagy 1967b).

It is mentioned above, that \(G\)-statistics and differences between \(G\)-statis-
tics have chi-square distribution. However, this is true only if data in the con-
tingency table come from independent observations. Due to the finite size of
sampled area, our data are spatially autocorrelated. This autocorrelation pre-
vents us from using statistical tables. Values of the statistics have to be deflated
(Dale et al. 1991) or significance levels have to be established by randomisation
(e.g. Palmer and van der Maarel 1995, Roxburgh and Chesson 1998, Horváth
When the frequency of species combinations depend only on the frequency of species and there are no interactions among species, the associatum should be zero. However, there are so-called textural constraints (e.g. the limited size of community, finite size, shape and non-random distribution of individuals) which restrict the coexistence of species even if there is not any biological interaction (cf. Bartha 1992, Bartha and Kertész 1998). Thus, positive associatum does not mean that there are biological interactions among species. The effect of these textural constraints can be eliminated from the results by using different randomisation techniques. For example, complete randomisation eliminates the effect of limited size of the community, but does not eliminates the non-random distribution of individuals, while random shift eliminate both effect (Palmer and van der Maarel 1995, Bartha and Kertész 1998). It should be emphasised that these textural constraints are not only the noise, which should be eliminated during the analysis. The uneven frequencies of species combinations, whether they result from biological interactions or textural constraints, may have vegetation dynamical consequences (Czárán and Bartha 1992, Czárán 1998).

CONCLUSIONS

Both log-linear contingency table analysis and Juhász-Nagy’s model family are derived from the information statistics. Consequently, they use many
common mathematical forms. 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 al. (1991), and they used only one of the numerous possibilities of this method. One aim of this paper is to begin the translation (I considered only the three most important characteristics of spatial pattern), 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, e.g. the importance of diversity, which cannot 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 experience, and these terms always have biological meaning. Unfortunately, many biologists are not familiar with the information theory functions used by Juhász-Nagy, while G-statistics and log-linear contingency table analysis are treated in the statistical textbooks for biologist (e.g. Sokal and Rohlf 1981, Zar 1999). The other aim of this paper was to show the mathematical relationship between the two approaches, and by this way to make the understanding of Juhász-Nagy’s models easier.

Appendix: 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, etc.). There are some drawbacks of these programs:

– computerised sampling and randomisation have to be done with another program,
– the maximum 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 (e.g. Podani 1991, Erdei and Tóthmérész 1993, Horváth 1998, Bartha in appendix of Bartha et al. 1998). 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, there are some G-statistics (first of all third- and higher-order interactions), which cannot be calculated from Juhász-Nagy’s functions, because they are calculated by iterative procedure.

REFERENCES


