TISCIA 31, 71-84

INFOTHEM PROGRAM: NEW POSSIBILITIES OF SPATIAL SERIES ANALYSIS BASED ON INFORMATION THEORY METHODS

A. Horváth

Horváth, A. (1998): INFOTHEM program: new possibilities of spatial series analysis based on information theory methods. — *Tiscia 31, 71-84.*

Abstract: The aim of the INFOTHEM 1.0 program is to analyze coenological patterns based on information theory models in a spatial series approach. The analysis, concept of which was developed by Juhász-Nagy, was applied to describe coexistence structures of vegetation. The facilities of program include numerous functions of both syncretic and diacretic models. It is also possible to use Rényi's general entropy formula rather than Shannon's formula. The spatial scaling of information theory functions is integral part of program operation, so there are some procedures to organize fusion of primary sampling units in steps of the spatial series. For the statistical evaluation, to estimate deviation from null hypothesis, the program is able to generate many types of random patterns considering different random references, and to calculate the significance levels using Monte-Carlo simulation. The features and operation of the program are discussed with some remarks on the application of information theory models.

Keywords: coexistence structures, information theory models, loess grasslands, random reference, spatial series analysis

A. Horváth, Department of Ecology, JATE University, H-6701 Szeged, Pf. 51, Hungary

Introduction

The analysis of coexistence structures of vegetation was developed by Juhász-Nagy (1967, 1972a, 1972b, 1973, 1976, 1980a, 1980b, 1984, 1985) on the basis of information theory mainly in the sixties and seventies. The theory concentrates upon the most cardinal question of plant community research: is there any spatial dependence of populations existing in a given stand (Juhász-Nagy 1976, 1984)? In contrast to traditional approaches focusing only on the pairwise association between species, Juhász-Nagy's methods apply to the entire community (associatum as total association, Juhász-Nagy 1967, 1972a). Since the pioneer works by Juhász-Nagy, a new measure, the so-called mean compositional information (Podani and Czárán 1997) has been developed as a connecting link between the individualistic approach and the investigation of coexistence structures.

The information theory methods represent a unique tool to describe the coexistence structures of multispecies pattern (cf. Podani *et al.* 1993; Podani and Czárán 1997), by placing the basic coenological phenomena (preference, diversity and resemblance) into a coherent framework (Juhász-Nagy 1986). Information theoretical functions are additive, well-manageable and programmable. The spatial scaling is inherent and essential part of the models, which use the concept of characteristic areas in a more comprehensive sense than minimum area (Juhász-Nagy and Podani 1983). Consequently, the problem can only be approached by spatial series sampling (spatial process: Podani 1984a, space series: Podani 1992, spatial series: Erdei and Tóthmérész 1993). Finally, the models give possibility to study the spatial pattern on both community and population levels (syncretic and diacretic models: Juhász-Nagy, 1973, 1980a), even reflecting to coalitions.

Although the fundamentals were developed more than twenty years ago, and the test of models started at that time, too, extensive applications could not begin until computers were available (due to the many calculations, cf. Erdei *et al.* 1994). The first published analyses (e.g. Juhász-Nagy and Podani 1983) were calculated with the SYN-TAX program package (Podani 1980, 1988). Other computer programs have been made since then, such as MULTI-PATTERN (Erdei and Tóthmérész 1993; Tóthmérész and Erdei 1995), and JNP-MODELLEK (Bartha *et al.* 1994). There were several field studies using the full capabilities offered by the programs. These suggest some new methodical and methodological problems which did not arise earlier. One of these problems is the question of primary sampling (e.g. planning of the sampling area and arrangement of sampling units in the field), another is the problem of spatial scaling (in fact: execution of secondary sampling by fusion of primary sampling units, that is the organization of spatial series steps usually by the computer, see below). The next problem is the question of random references and significance tests, and finally the role of rare species (cf. Tóthmérész and Erdei 1992) should be mentioned.

In this study the first three problems are surveyed. In addition to these theoretical problems, there is one more requirement in practical computer work, namely the well-structured, arranged and transportable files of results. The INFOTHEM program has been developed with this in mind, in IBM DOS compatible environment. First the methodical problems are discussed (summarizing the known possibilities and proposing the new procedures), then the use of the program is explained. As an example we illustrate the results of own field data collected in several types of loess plant community from Mezõföld region.

Methodical comments

Primary sampling

Sampling experiments for testing the information theory methods were carried out based on theoretical considerations (Juhász-Nagy 1972a). In this way there are several sampling sets corresponding to specified spatial series steps, and each set contains a certain number of sampling units of a given size. The shape of sampling units is, for example, circular and plots are located at random in the study area (Juhász-Nagy 1980a; Juhász-Nagy and Podani 1983). Thus, spatial scaling was realized directly in the field. If all plots are located separately and randomly, we can speak about *independent plots*, whereas the *nested plots* mean that we laid down first the plots belonging to the largest area set, and these units will completely include plots of other smaller sets (Podani 1984a). Usually we use isodiametric units of any shape (circular or square especially, cf. Juhász-Nagy and Podani 1983; Podani *et al.* 1993).

Since a large number of sample plots is necessary for the analysis in each steps of the spatial series, the above sampling technique requires exhaustive field work. To reduce sampling effort, we can use a *grid* or a *transect*, in which the plots are in contact with each other by their four or two sides. (A *map* can be considered as a grid consisting of infinitely small units, namely continuous X, Y coordinates.) In this case the spatial scaling will be realized after field sampling in the *secondary sampling*.

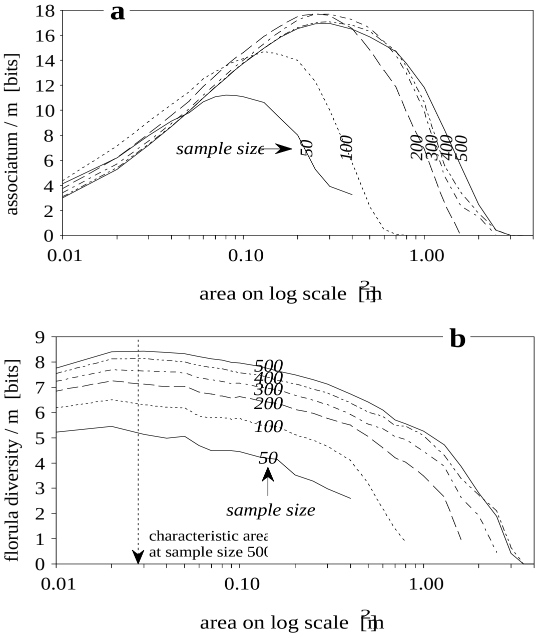


Fig. 1. The effects of unit size and sample size on the characteristic functions. If sample size is too small (in this case less than about 300), both the maximum value and area of maximum value is translocated, so the curve of associatum becomes deformed (**a**). We can see on diagram **b**, that using larger than 0.03 m^2 unit size we lose the real global maximum value of florula diversity curve, and the characteristic scaling is false or is not possible. The data result from a loess grassland community consisting of 31 species with more than 1% relative frequncies.

When we plan the field (primary) sampling, we have to choose either a grid or a long transect, and define primary unit size, and the number of units (sample size). *Sample size* depends on the number, composition and association of species; the empirical ratio of the number of units and species is at least 5-10 (e.g. 200/8=25 in the *Saxifraga stellaris* coenotaxon [Juhász-Nagy 1980a], 2500/37=68 in a steppe community [Bartha and Horváth 1987], between 2750/3=917 and 2750/21=393 in primary succession [Bartha 1992]). If sample size is too small, the shape of the characteristic functions can become distorted, and characteristic scaling will be imperfect (Fig. 1a). *Unit size* depends also on composition of species; in a diverse, intact grassland it is about 25-100 cm^2 (e.g. 5×5 cm in dolomite grassland communities [Szollát and Bartha 1991], 10×10 cm in loess grassland [Bartha and Horváth 1987], 20×20 cm in pioneer communities [Bartha 1992]). If unit size is too large, the characteristic areas can be outside of the surveyed spatial scale (see Fig. 1b). To choose between grid and transect, we consider two aspects. The long transect (consisting of the same number of units as a grid) spans a larger area of the stand, and yields less redundant data. On the other hand, when applying a long transect, influence of elongated secondary sampling unit must be considered (see Nosek 1976; Podani 1984a, 1984b; Bartha and Horváth 1987 for more details).

Secondary sampling

Secondary sampling is unavoidable for spatial scaling if a grid (or transect) is used to collect data. (For mapped point patterns of species, the secondary sampling usually involves circular plots [cf. Podani 1984b; Podani and Czárán 1997].) Primary plots will be fused in a predetermined arrangement, producing *secondary sampling units*. Fusion means that in the secondary unit a species is present, if it occurs in any primary unit fused, in other words logical "or" operation will be performed among binary values of primary plots. In each step a certain number of primary plots are fused, in a regular manner. This means that the units to be fused must be contiguous. The shape of secondary unit is usually isodiametric in case of grids, and elongated in case of transects. (Sometimes we merge units that are positioned in the grid randomly; this is a procedure to make a type of random references.) There are two important alternatives to define secondary plots on a grid: randomly (*random sampling*) or regularly (*systematic sampling*). In random sampling we use a constant number of plots in each spatial series step. In this case all area of the grid is sampled more or less uniformly (except edges) supposing that sample size is sufficiently large. Using systematic sampling we have to shift units on the grid by a given number of primary units. This number differs from each other with the size of secondary plots. In this case we must guarantee (by programming appropriate offset) that the complete area will be sampled uniformly. If the offset is only one primary unit long, we have the so-called *complete sampling* (cf. Bartha *et al.* 1995), since the grid is completely sampled from all possible positions in all spatial series steps. In case of complete sampling, the number of secondary units varies in the different spatial series steps, so it must be standardized with sample size.

Although the first analyses were executed with random secondary sampling, it can be shown that the curve fitted to values of any information statistical function depending on sample size reaches the expected value in the infinite, whereas complete sampling gives good approximation (see Fig. 2). The explanation is that in complete sampling plots are located to all possible place exactly once. Because it is the computer program that performs secondary sampling, complete sampling is easily accomplished, so that standardization with sample size is always performed.

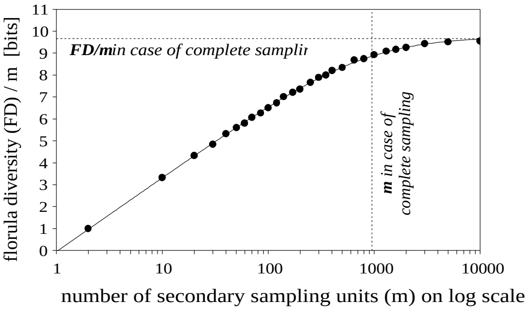


Fig. 2. Comparison of random and complete sampling for standardized florula diversity. Field sampling was made in a loess grassland community, using a 50×25 grid of 20×20 cm units. The size of the secondary sampling unit is 1×1 m. Applying random sampling the equation of curve (solid line) fitted to empirical values (points): y=9.75+3.38log(x/x+790), so the expected value in infinite: 9.75 (correlation coefficient of fitting: 0.9998). At sample size (m) 10000 we get 9.56 that is the nearest to the expected value. Using complete sampling, this value (exactly 9.63) results at m=966 (dotted lines).

For the spatial series the determination of the size of secondary units is needed. The increment of size along the spatial scale depends on the requested accuracy of measures; fine resolution of scale is important near characteristic areas (which are unknown before analysis). Usually we use a logarithmic spatial scale, so the increment will cause the exponential enlargement of the secondary unit. To determine the area of the largest secondary unit we must consider that its area must be smaller than about the quarter of total sampling area (because of the edge effect).

Random references

Studying the coexistence structures of communities, we expect that there is spatial dependence (pairwise and multiple) among species. So, our null hypothesis is that populations living in a stand are completely independent from one another (cf. metamethodological quadruplet: Juhász-Nagy 1986). This means that they are combined randomly with each other at all spatial scales. Therefore, to prove the existence of coexistence structures (patterns), we have to use significance tests to show the difference from randomly combined species. There are some problems regarding the hypothesis test of associatum and other functions. First of all, larger than zero associatum does not necessarily represent real spatial dependence because of the textural and structural constraints derived from the abundance and spatial pattern of species (see Podani 1982, 1984b; Szollát and Bartha 1991; Bartha 1992; Tóthmérész and Erdei 1992). Secondly, if secondary sampling units overlap, namely some regions of the study area may take part in many sample plots, the studied area is “over sampled” (this is valid in most cases), and thus the criterion of independently placed sample plots does not satisfy, therefore conventional significance tests do not apply (Podani 1984; Podani *et al.* 1993). Thirdly, conventional significance tests have not been developed for information theory methods, because there is a multitude of parameters of species (abundance, dispersion type) which should be considered for the null model. (We can produce random references by direct calculation of florula frequencies assuming the Poisson distribution for all species; cf. Podani *et al.* 1993; Erdei *et al.* 1994; Podani and Czárán 1997.) Usually there is only one possibility: to compare the functions derived from field data to *random references*, and to use *Monte-Carlo simulations* as a basis of the significance test.

Considering different null models, there are several *types of random references*. In any case, we transform one or more parameters of the population or pattern into random. Table 1 shows some types (marked with numbers and names) according to unchanged or altered (random) parameters. (The role of variable called “RRType” (Random Reference Type) will be detailed later.) In case of first type we randomize the location of the primary sampling units in a grid. In case of types 2–4 we relocate species occurrences in a grid, so the species will combine randomly. In case of the fourth type only species frequencies will be constant; this is the so-called *complete randomization* (Bartha 1990, 1992; Tóthmérész 1994b; Bartha *et al.* 1995; Margóczi 1995). In case of the third type the distribution of species number in sampling units will also be held constant, while in the second type, in addition, the species number of each plots in the given position of sampling area will be unchanged. In the types 5–8 the species frequencies are fitted to a certain species-abundance distribution model, and species are mixed randomly.

If we want to preserve species distribution type (pattern) in space, we must choose the ninth type of random references (see Table 1). In this case we shift the pattern (the patches) of a species along two dimensions of grid, or alongside transect by randomly generated number of primary units (Fig. 3a-b). This process does not change the distribution of species if we have circular transect, because the opposite ends of ring-shaped transect are connected (Fig. 3c). Now, the random shift occurs as a random rotation. This sampling design is introduced as “trainsect” (Palmer and van der Maarel 1995), and its

Table 1. Types of random references. Symbol ≡ shows if a parameter agrees with the value in the original pattern. See text for more details.

Name of the random reference	*Number of type (RRType)*	*Frequncy of species*	*Distribution of species number in plots*	*Distribution of species number in space*	*Species combination in plots*	*Pattern of distribution of species*
plot randomization	1	≡	≡	random	≡	random
	2	≡	≡	≡	random	random
	3	≡	≡	random	random	random
complete randomization	4	≡	random	random	random	random
making of random patterns with species frequencies deriving from some species-abundance models	5	random	random	random	random	random
	6	geometric distribution model	random	random	random	random
	7	linear distribution model	random	random	random	random
	8	broken stick distribution model	random	random	random	random
random shift	9	≡	random	random	random	≡

application is detailed by Bartha and Kertész (1997). Although in grid or transect the edge effect appears (patches of aggregated species will be separated or fused, see Fig. 3a-b), if patches are not too big compared to sampling area, and there are many patches, this method will more or less preserve the original distribution of species. Since shift length varies randomly with different species, populations will be combined randomly. We can generate patterns of multispecies community with the MULTI-PATTERN program package (Erdei and Tóthmérész 1993), supposing that we know all species parameters.

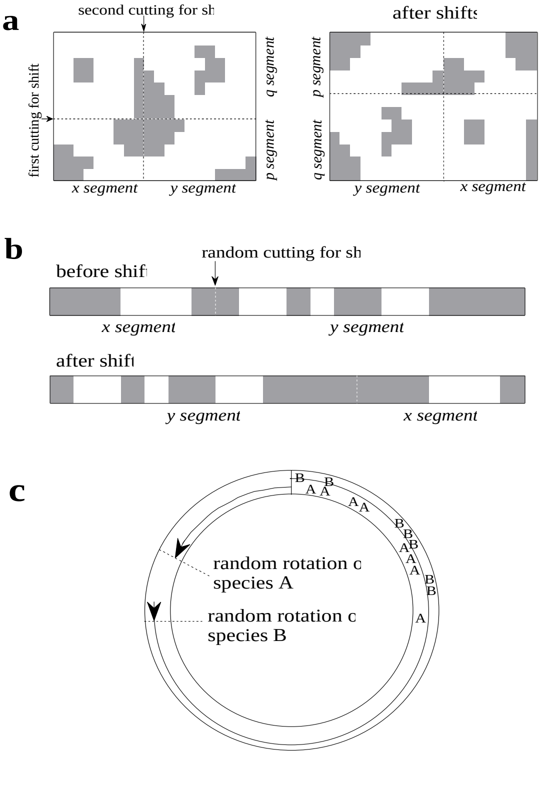


Fig. 3. The possibilities of random shift to generate random references in several types of sampling arrangement. In **a** and **b** the dark patches indicate the pattern of a species, while in **c** two populations (A and B) occur along just the quarter part of circular transect together or separately. In case of a grid two random shifts are needed at right angles to each other (**a**), and only one in case of transect (**b**). In the circular transect the random shift appears as random rotation (**c**). The **c** diagram shows that the lengths of shifts may differ by species, so the original species combination will be converted into random.

Because different random references indicate different null models, the evaluation of results has to consider the transformed and unchanged parameters of original pattern. The random shift (type 9) represents the strongest limitations for randomization, so the differences between the values of field and random data are probably the smallest in this case. The preservation of distribution of species number in sampling units is also relatively strong limitation (types 1–3), in absence of it (but not changing the species frequencies: type 4), we have less defined random references. Among the other listed types (5–8), that type is the nearest to the field situation which contains the most fittable species–abundance model to original frequency distribution. On Fig. 4 we can see an example to study the results applying the different types.

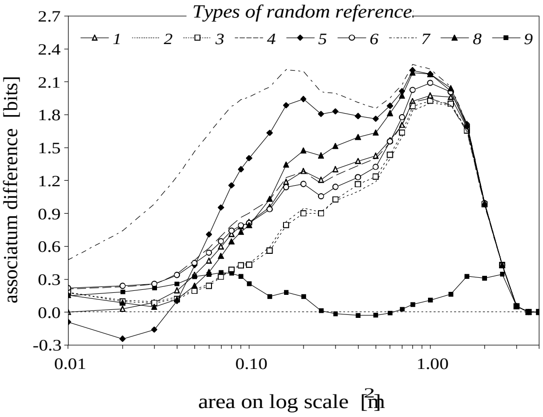


Fig. 4. Curves of differences between associatum values derived from field data and several types of random patterns. Data of 500 primary units originate from a pioneer loess plant community, species number is 11 (without rare species). We have 29 random references at all types, and perform complete sampling. In all cases we can detect positive differences in nearly the total interval of the spatial scale, so the organization in pattern of community has been demonstrated. The level of difference refers to distance between the null model and the field situation. For example, the curve of type 8 (broken stick distribution) has higher values than curve of type 6 (geometric distribution), which confirms that the species-abundance relation represents an initial stage in the succession of communities.

For significance test we apply a Monte-Carlo simulation (cf. Galiano *et al.* 1987; Bartha 1992; Podani *et al.* 1993; Bartha *et al.* 1995; Podani and Czárán 1997). We make many random patterns of a given type of random references, and analyze them with the same method used for the field data. The simulated values determine a random envelope along the spatial scale. If the original curve runs inside the envelope the field pattern probably does not differ from random patterns (cf. Fig. 5a). To calculate significance level, in case of each random pattern we examine whether the original value is under or above (or equal to) the random value. We do not have predetermined hypothesis in case of a concrete function, so if the number of positive differences is higher than the number of negative differences, the hypothesis is that we have a larger value than which derive from the random reference; and vice versa.

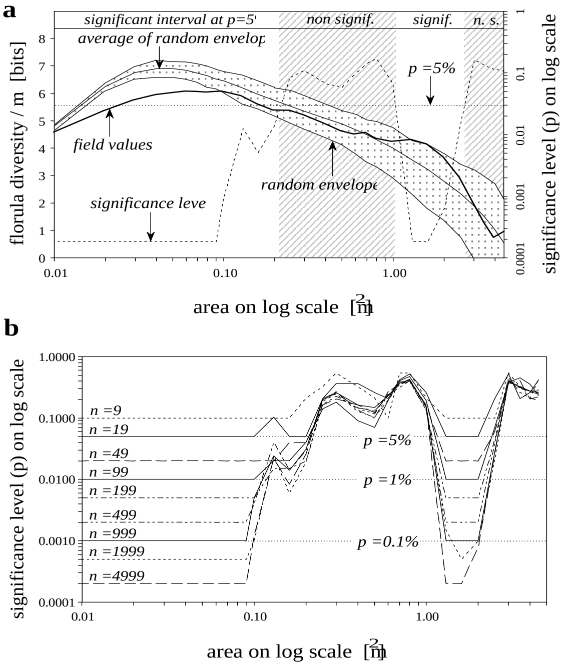


Fig. 5. An example to evaluate the florula diversity curves of random references refering to the significance levels. The diagrams were calculated from the data of a pioneer loess plant community, species number was 18. We use the type 4 for the randomization. The number of random references was 4999 in case of **a**, while in **b** we have a set of different numbers (*n*). In **a** we can see two significant (non shaded areas) and two non significant (shaded areas) spatial scale intervals at *p*=5%. The **b** diagram shows that this intervals do not vary with *n*, if *n* is greater than 19 (the curves cross the pointed line of p=5% level at approximately same locations). Reducing the number of random references there are three results: (1) in significant intervals *p* increases, but it remains inside the accettable domain (if *n* is not too low!), (2) in non significant intervals *p* keeps its high and more or less constant, non-accettable value, (3) in transitional interval (now: $0.1\text{m}^2<\text{area}<0.2\text{m}^2$) *p* seems to be partly unpredictable (varies between about 1% and 5%). Of course, increasing *n*, the reliability of statistical measure increases also.

The level of significance, *p* (probability of type I error) is given by:

$$p=\frac{n-ND+1}{n+1}$$

where *n* is the number of randomizations, *ND* is the number of positive or negative (the higher) differences between field and random values. (For example, if we have 99 random references, and the number of positive differences is 3, number of negative differences is 95, and there is no difference in one case, the actual hypothesis is that field value is lower than the random, and

$$p=\left[99-95+1\right]/\left[99+1\right]=0.05\,.)$$

It is necessary to take numerous random references to produce an acceptable significance level, while considering the available time for calculations (cf. Fig. 5b).

The incidental effects of rare species appear mainly with random references, causing a widening of random envelope at larger values of spatial scale (Podani *et al.* 1993). It can be seen in Fig. 6 that in

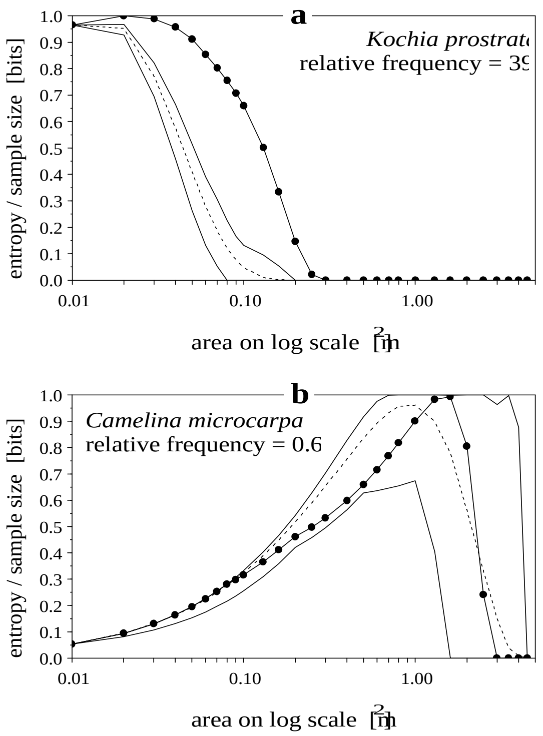


Fig. 6. Entropy curves of two species with different frequency from a pioneer loess community. The number of primary plots is 500, and complete sampling was performed using 99 random references. The line with full circles represents the curve derived from field data, the solid lines indicate the maximum and minimum values of random references; and the dotted line is the average of them. In case of high frequency the random envelope is narrow (**a**), while for very small frequency it becomes wider in which the field values are found (**b**). This means that the concrete location of a rare species has strong, but vagueness-increasing influence on entropy curves. Because the local distinctiveness is the sum of entropies, rare species could produce disadvantageous effect in information theory functions.

Table 2. Formulae to calculate the required memory capacity for the program. The meaning of the variables can be found in Table 4. *Notes:* (1): If every parameter has its maximum value; (2): If QN=maxQO=1000, PN=50, NSP=25.

Necessary result	*Output file extension*	*Memory requirement (bytes)*	*Maximum required memory (1)*	*An example of required memory (2)*
synchretic values	RSA	$QN \cdot 25 + \max QO \cdot 50 + NSP \cdot 30$	753 kB	75.75 kB
diachretic values	RSB	$\max QO \cdot 25 + PN \cdot NSP \cdot 30$	850 kB	62.5 kB
Rényi’s entropy	RSC	$NSP \cdot 738$	73.8 kB	18.45 kB
pairwise association	RSD			
lists of florulas	RSE			
If random references are needed:		$QN \cdot 25$	250 kB	25 kB
• synchretic values	RSF	$NSP \cdot 90$	9 kB	2.25 kB
• diachretic values	RSG	$PN \cdot NSP \cdot 66$	1.32 MB	82.5 kB
• pairwise association	RSH	$PN \cdot PN \cdot NSP \cdot 14$	56 MB	875 kB
		sum total:	60 MB	1.14 MB

case of a frequent and aggregated species the entropy curve is shifted right from the random reference, but the entropy curve of a rare species has a wide random envelope along the whole spatial scale. Therefore it could be necessary to neglect rare species with relative frequency less than about 1% (cf. Bartha and Horváth 1987; Bartha 1992; Tóthmérész and Erdei 1992). Tóthmérész suggests to solve the problem of rare species using Rényi’s general entropy functions (Tóthmérész 1994a).

Description of INFOTHEM 1.0 program

General specifications

The INFOTHEM program has been developed by reflecting the methodical problems detailed above. The most important new facilities are that the program can realize various types of the secondary sampling procedures and random references, calculates the significance level for randomizations, and creates well structured output files.

The program is a DOS application, so it can be started in most user interfaces (e.g. DOS, Norton or Windows environment) on IBM PC-s and compatible machines. Because calculations may be time consuming for a big data matrix, a fast computer with numerical coprocessor and about 8 MB RAM can be useful. To count the required memory capacity we can refer to Table 2. If capacity is not enough (the program aborts), we have two possibilities: (1) to quit the „shell” program which uses much memory (e.g. Windows), and to work in plain DOS environment; (2) to get the different output (result) files one by one (this is realizable by setting the Res parameter in the parameter file; see Table 3).

A data file and a parameter file are necessary for running the program. The type of all files (output files also) is *text file* (ASCII-file). The input (data) file is just like a data matrix with sampling units as rows and species as columns. The values may be binary (presence/absence) or quantitative abundance (frequency), but the program transforms all data types into binary. At least one space or tabulator character has to separate each value, the number of delimiters is not fixed. If the sampling area (grid) consists of *m* rows and *n* columns, the units are given row after row in the matrix.

The parameter file consists of at least six rows. The first five rows contain the general parameters for program running and analysis; the parameters of every spatial series step are specified in the next one or more rows (Table 3). Therefore, the number of rows in the parameter file is the number of spatial series steps plus five. The meanings and limits of parameters are given in Table 4, and Fig. 7 gives an example for parameter file structure. If the parameters are not adequate, the program aborts. We can make the parameter file directly using any editor program (e.g. DOS or Norton Editor), or interactively by the program. If we have a parameter file ready, we can start the program by entering the name of the parameter file after the program name.

Table 3. The structure of the parameter file. The content of sixth rows is repeated for each step of the spatial series.

row 1	FI
row 2	FO
row 3	QN PN
row 4	FType QS RN NSP Res
row 5	RRType RRNumb MaxFr MinFr
row 6	Area QO QR RF SH BR BD

Table 4. The description and limits of parameters.

parameter	*description*	*notes*
FI	name of input (data) file	with path, if necessary
FO	name of output (result) files	without extension!
QN	number of plots in data file	$1 < QN \le 10000$
PN	number of species (populations)	$1 < PN \le 200$
FType	type of fusion of primary sampling units for secondary sampling	=1: random, plot-repeated fusion =2: random, non-repeated fusion =3: regular fusion, systematic sampling =4: regular fusion, random sampling
QS	number of plots in one row of grid	in the sampling area
RN	number of rows of grid	in the sampling area
NSP	number of spatial series steps	$1 \le NSP \le 100$
Res	code to set the required output (result) files	serial number of character of an output file type: RSA(RSF)=1, RSB(RSG)=2, RSC=3, RSD(RSH)=4, RSE=5 (e.g. Res=11010)
RRType	type of random references	=0 ... 9 (see Table 1 for more details)
RRNumb	number of random references	$1 \le RRNumb \le 10000$
MaxFr	frequency of the most frequent species	$1 \le MaxFr \le QN$
MinFr	frequency of the rarest species	$1 \le MinFr \le QN$
Area	area of a secondary plot	decimal number is allowed
QO	number of secondary plots	$1 \le QO \le 10000$
QR	width of a secondary plot	in number of unit of primary plots
RF	height of a secondary plot	in number of unit of primary plots; $1\le QR\cdot RF \le 10000$
SH	length of shift	in number of unit of primary plots
BR	number of shift to skip	see text for more details
BD	length of skip	in number of unit of primary plots

While the program is running, screen displays the name and extension of required output files under the heading, and in the next four rows the number of steps left. The program will finish the analysis when each value becomes zero.

```
DATA.DAT
RESULTS
25 6
3 5 5 3 11111
4 99 1 1
0.01 25 1 1 1 1 0
0.02 20 2 1 1 4 1
0.04 16 2 2 1 4 1
```

Fig. 7. The contents of the example parameter file named PAR.

Setting of parameter values

In one row of the data (input) file the values of all species follow one another. The maximum number of species is 200, and we can set it by PN parameter in the third row of parameter file (see Table 3). The QN parameter represents the number of primary sampling units; its maximum value is 10000. For grids the value of QS (number of plots in one row of grid) and RN (number of rows) is more than 1, but in case of a transect RN=1.

The name of the input file with extension and path is entered in the first row of parameter file. The second row contains the name of the output file without extension, because the extensions will be attached by the program according to the required results. The meaning of different extensions and the roles of output files belonging to each extension are given in the first two columns of Table 2, but the details will be explained below. The required results (selected types of analysis) can be determined by the Res parameter. This parameter consists of five characters referring to the first five output files (with extensions RSA, RSB, RSC, RSD and RSE). If a given character of Res is 1, the given output file is necessary, otherwise it is 0.

The secondary sampling type is set by the FType parameter (cf. Table 4). FType controls the arrangement of primary plots for fusion. The fused plots form the secondary unit. The fusion may be of four kinds. If FType =1, the positions of plots for fusion are random in the grid, and some plots may be fused again many times. In case of FType =2, the difference is that now a plot can be fused only once, so the maximum number of secondary units comes from QN/QO. These two values of FType give possibility to make two types of random references. However, if we want to compare values of a model derived from field data to random references, we have to choose other options as given below. In the case of FType =3 or 4, systematic or random secondary sampling will be performed using regularly shaped secondary plots.

The value of the NSP parameter equals to the number of steps in the spatial series. The parameters of each step succeed row by row in the parameter file started at the sixth row. The Area (area of secondary sampling unit, actually the identifier of a given step) is the only informative parameter for the user, the program does not calculate with it, but it will be listed in the output files. QO determines the number of secondary plots in a given step; QO=QR·RF. The dimension of a secondary unit is given by the QR and RF parameters, similarly QS and RN specify the dimensions of sampling area. QR and RF have meaning only if FType=3 or 4, otherwise their product is important. If we have a transect, RF=1.

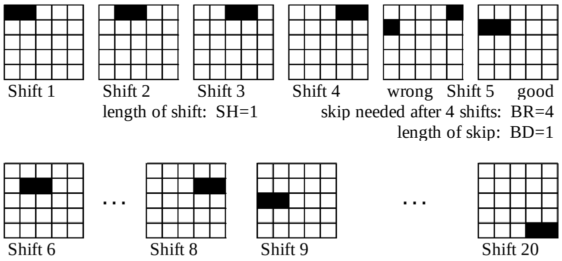


Fig. 8. Demonstration to set parameters of the systematic secondary sampling according to the second spatial series step listed in seventh row in parameter file in Fig. 7.

The SH, BR and BD parameters are larger than zero only in case of systematic sampling (Ftype=3). SH indicates the length of shift on the grid by a given number of primary units. In complete sampling SH=1. If we work on a grid, and the dimension of the secondary unit is larger than 1×1 primary plots, it has to perform skip when the secondary plot reaches the margin of grid. The length of skip is given by BD, and the number of shifts before each skip is given by BR. Fig. 8 explains the meaning of these parameters. To design spatial series steps using visual control, the determination of parameters can be performed by the SPATPROC program developed by the author. If we adopt complete sampling, the preparation of the parameter file is recommended by the INFOTHEM program (started without parameters), because the program will calculate the SH, BR, BD parameters in all steps of the spatial series.

When random references are not needed, the value of RRType is set to zero, otherwise it falls between 1 and 9. The different types of random references are listed in Table 1 (see also “Random references”). If species frequencies are fitted to species abundance models (RRType=5–8), the value of MaxFr is equivalent to the frequency of the most frequent species, and MinFr indicates the frequency of the rarest species, otherwise their value is zero. The number of random references is adjusted by the RRNumb parameter. If RRType>0, then files with RSF, RSG and RSH extension will be output, if RSA, RSB and RSD output files are needed (cf. setting the Res parameter).

The output files

There are five output files for the results, and another three for the versions of random references (cf. the first and second columns of Table 2). The output files are standard ASCII-(text) files, and their data are well arranged and structured, so it can be easily transported to any graph or chart editor program (e.g. Excel 5.0). The first seven rows of each output file are constant, they contain the general parameters of the analysis (cf. Fig. 9). The files with random references have additional three rows (e.g. Fig. 10).

```
Name of input file: DATA.DAT
Name of parameter file: PAR
Number of quadrat in input file: 25
Number of species: 6
Dimension of sampling area: 5*5
Type of quadrat fusion: 3
Number of spatial process steps: 3

  Area    QO  QR  RF  SH  BR  BD    FDiv      FEv      LD       LEv      Ass
-------------------------------------------------------------------------------------
0.01000   25   1   1   1   1   0   102.59   0.88363   143.42   0.95612   40.83
0.02000   20   2   1   1   4   1    70.93   0.82057    88.91   0.74094   17.98
0.04000   16   2   2   1   4   1    26.03   0.40665    30.63   0.31906    4.60
```

Fig. 9. A detail from the RESULTS.RSA sample output file.

The function of the RSA output file is to store the values of the most common syncretic models. The designation, meaning and formula of these models are found in Table 5. The source of variables and formulae used is Juhász-Nagy, 1976, 1984, 1985 and Juhász-Nagy and Podani, 1983. In the file the results of each single spatial series step follow row by row after the heading. The width of rows is 354 characters. An example can be found in Fig. 9.

Table 5. The variables listed in the RSA file. ST=standardized by number of plots. The basic variables: *m*=number of plots, *s*=number of species, *i*=a given species (*i*=1..*s*), *j*=a given step of the spatial series, *k*=a given species combination (florula), *f*=frequency of a species combination, *g*=a given plot (*g*=1..*m*), *n*=frequency of a species.

designation	*description and formula*
FDiv	florula diversity: $mH_j^{(\varphi)} = m\log m - \sum_k f_{jk} \log f_{jk}$
FEv	florula evenness: $mV_j^{(\varphi)} = mH_j^{(\varphi)} \Big/ \left(m\log m\right)$
LD	local distinctiveness: $mH_j\left(\left[L\right]\right) = sm\log m - \sum_i \left[n_{ij}\log n_{ij} + \left(m - n_{ij}\right)\log\left(m - n_{ij}\right)\right]$
LEv	relative local distinctiveness: $mH_j\left(\left[L\right]\right) \Big/ sm$
Ass	associatum: $mI_j\left(\lambda\right) = mH_j\left(\left[L\right]\right) - mH_j^{(\varphi)}$
Com	number of realized species combinations: ω
FDiv/Q	ST florula diversity: $H_j^{(\varphi)}$
Ld/Q	ST local distinctiveness: $H_j\left(\left[L\right]\right)$
Ass/Q	ST associatum: $I_j\left(\lambda\right)$
Com/Q	ST number of realized species combinations: $\omega \big/ m$
ESV	entropy of species valences: $N_j H_j\left(V_q\right) = N_j \log N_j - \sum_i n_{ij} \log n_{ij}$
ESI	entropy of species invalences: $n_j H_j\left(v_q\right) = n_j \log n_j - \sum_i \left[\left(m - n_{ij}\right)\log\left(m - n_{ij}\right)\right]$
EQV	entropy of plot valences: $N_j H_j\left(V_t\right) = N_j \log N_j - \sum_g n_{jg} \log n_{jg}$
EQI	entropy of plot invalences: $n_j H_j\left(v_t\right) = n_j \log n_j - \sum_g \left[\left(m - n_{jg}\right)\log\left(m - n_{jg}\right)\right]$
ESV/Q	ST entropy of species valences: $N_j H_j\left(V_q\right) \Big/ m$
ESI/Q	ST entropy of species invalences: $n_j H_j\left(v_q\right) \Big/ m$
EQV/Q	ST entropy of plot valences: $N_j H_j\left(V_t\right) \Big/ m$
EQI/Q	ST entropy of plot invalences: $n_j H_j\left(v_t\right) \Big/ m$
Diss	dissociatum: $mH_j\left\{\delta_\lambda^{(s)}\right\} = mH_j\left(\left\{A\right\}\right) + mH_j\left(\left\{B\right\}\right) + \ldots + mH_j\left(\left\{S\right\}\right)$
Diss/Q	ST dissociatum: $H_j\left\{\delta_\lambda^{(s)}\right\}$
SumPosAss	sum of positive pairwise associations
SumNegAss	sum of negative pairwise associations
Diff(P-N)	difference between sum of positive and negative associations
SumPosAss/Q	ST sum of positive pairwise associations
SumNegAss/Q	ST sum of negative pairwise associations
Diff(P-N)/Q	ST difference between sum of positive and negative associations

The RSB file is the source of diacretic functions listed in Table 6. The file contains the results step by step in separated units according to species. The width of this file is 141 characters.

The output file with RSC extension includes the three most common syncretic models and their values standardized by the number of sample plots, but now based on Rényi’s general entropy function. This process makes an ordering of characteristic curves by the α parameter (in addition to spatial scaling), as in case of diversity ordering (cf. Patil and Taillie, 1979; Tóthmérész, 1993, 1995). In the file there are six units according to these functions and their standardized forms, and different α values can be found in columns. α ranges from 0 to 4, and the increment is 0.1. If α=1, the program calculates the functions with Shannon entropy formula. The width of rows is 593 characters, and the file structure helps to make a three-dimensional representation. The values listed in the file can be seen in Table 7.

Table 6. Diacretic functions of RSB file. ST=standardized by number of plots.

designation	*description and formula*
Entr	local entropy of species i (or A): $mH_{ij} = mH_j(A) = m\log m - n_{ij}\log n_{ij} - (m-n_{ij})\log(m-n_{ij})$
TAss	total associativity of species A: $mI_j(\langle A\rangle) = mH_j(A) - mH_j(\{A\})$
TDiss	total dissociativity of species A: $mH_j(\{A\}) \equiv mH_j(A\|[B,C,\ldots,S]) = mH_j^{(\varphi)} - mH_j^{(\varphi\|A)}$
Diss%	total dissociativity of species A in percentage of dissociatum: $100\cdot mH_j(\{A\}) \big/ mH_j\{\delta_\lambda^{(s)}\}$
SubDiv	subflorula diversity without species A: $mH_j^{(\varphi\|A)} \equiv mH_j(B,C,\ldots,S)$
SubAss	subassociatum without species A: $mI_j([\overline{A}]) = mI_j(\lambda) - mI_j(\langle A\rangle)$
Entr/Q	ST local entropy of species A: $H_j(A)$
TAss/Q	ST total associativity of species A: $I_j(\langle A\rangle)$
TDiss/Q	ST total dissociativity of species A: $H_j(\{A\})$
SubDiv/Q	ST subflorula diversity without species A: $H_j^{(\varphi\|A)}$
SubAss/Q	ST subassociatum without species A: $I_j([\overline{A}])$

Table 7. The functions using Rényi’s entropy functions, listed in RSC file. ST=standardized by number of plots.

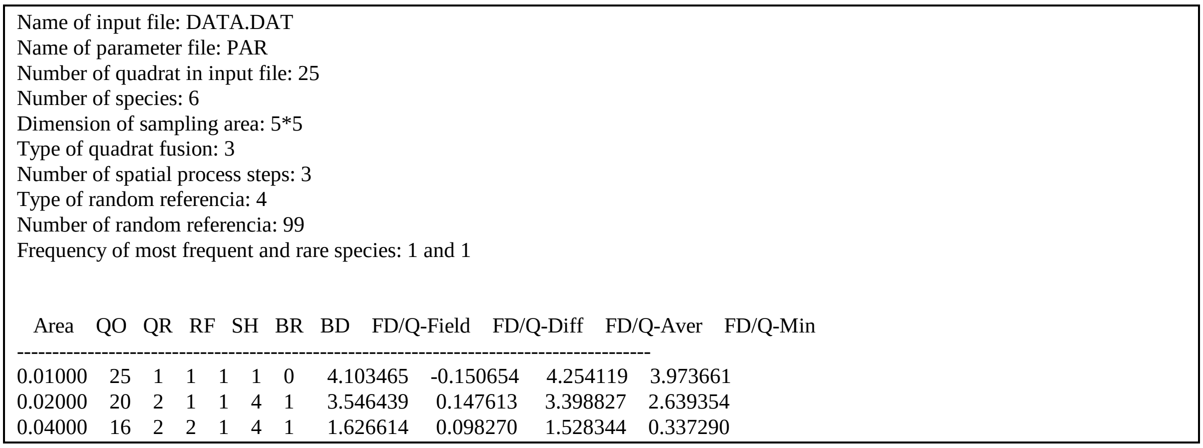
designation	*description and formula*
FDiv	florula diversity: $mH(\alpha)_j^{(\varphi)} = m\log\sum_k (f_{jk}/m)^\alpha \big/ (1-\alpha)$
LD	local distinctiveness: $mH(\alpha)_j([L]) = m\sum_i \left\{\log\left[(n_{ij}/m)^\alpha + ((m-n_{ij})/m)^\alpha\right]\right\} \big/ (1-\alpha)$
Ass	associatum: $mI(\alpha)_j(\lambda) = mH(\alpha)_j([L]) - mH(\alpha)_j^{(\varphi)}$
FDiv/Q	ST florula diversity: $H(\alpha)_j^{(\varphi)}$
Ld/Q	ST local distinctiveness: $H(\alpha)_j([L])$
Ass/Q	ST associatum: $I(\alpha)_j(\lambda)$

The RSD file contains the association values. The program calculates two types of pairwise association, both of them are on the basis of the contingency table derived from the occurrences of two species. One of them is the χ^2-value (which is not part of information theory methods), and the other is the association on information theory. This latter equals to half of the G-score (cf. Zar, 1984). If interest lies in calculating the association in cases when any cell of the contingency table has zero value, the program replaces 0 with 1, so the denominator of the formula for χ^2 will not contain zero. Hereby the association will be changed a little, but if the number of sampling plots is sufficiently large, this difference is negligible. In case of the χ^2-test the significance is examined at probability levels 5%, 1% and 0.1%. Note that conventional significance test does not apply in most cases, because the requirement of independence of sampling units is not met (cf. Podani, 1984; Podani *et al.*, 1993), so the application of random references is recommended (RSH file). The RSD file is divided into units the number of which equals to the number of spatial series steps. The variables of the file are listed in Table 8. The width of rows is 86 characters.

File RSE lists the realized species combinations (florulas). In each unit, according to the spatial series steps, the florulas appear row by row. Row contains the frequency of the florula, followed by the species number of the florula, finally the code with characters 0 and 1. Further processing of data of the RSE file (e.g. collecting all florulas for all spatial series steps to apply indirect global spatial series analysis, cf. Tóthmérész, 1994a) can be performed with the COMSUM program developed by the author.

Table 8. Variables in RSD file.

designation	*description and formula*
Sp1, Sp2	series number of two compared species
A, B, C, D	fields of contingency table derived from occurrence of two species. Occurences: A: 11, B: 10, C: 01, D:00.
A*D-B*C	$ad - bc$
Chi2-Ass	$\chi^2 = \dfrac{(m-1)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$
Sign	the significance level at which the χ^2-value is significant
Inf-Ass	$mI(A,B) = m\log m + b\log b + c\log c + d\log d - (a+c)\log(a+c) -$ $-(b+d)\log(b+d) - (a+b)\log(a+b) - (c+d)\log(c+d)$



```
Name of input file: DATA.DAT
Name of parameter file: PAR
Number of quadrat in input file: 25
Number of species: 6
Dimension of sampling area: 5*5
Type of quadrat fusion: 3
Number of spatial process steps: 3
Type of random referencia: 4
Number of random referencia: 99
Frequency of most frequent and rare species: 1 and 1


  Area    QO  QR  RF  SH  BR  BD   FD/Q-Field  FD/Q-Diff  FD/Q-Aver  FD/Q-Min
-------------------------------------------------------------------------------------
0.01000   25   1   1   1   1   0    4.103465   -0.150654   4.254119   3.973661
0.02000   20   2   1   1   4   1    3.546439    0.147613   3.398827   2.639354
0.04000   16   2   2   1   4   1    1.626614    0.098270   1.528344   0.337290
```

Fig. 10. A detail from the RESULTS.RSF sample output file.

The RSF, RSG and RSH files are in accordance with the RSA, RSB and RSD output files with the difference that they contain the comparison of functions with random references. The variables of these files are listed in Table 9, and an example for the RSF file can be found in Fig. 10. In a graph representing an information statistical function, the ...-Min and ...-Max values point out the boundaries of the random envelope (cf. Fig 5a). Note that the signs of ...-Sign functions do not relate to the sign of a particular function but show that the field value was larger or smaller several times than a random value (cf. “Random references”). If value of a ...-Sign function exceeds 0.50, it means that the field value was equal to a random one’s in several cases (in all cases if it is 1.00).

Table 9. Functions listed in RSF, RSG and RSH file. ST=standardized by number of plots.

designation	*description and formula*
...-Field	equal to a functions derived from field data
...-Diff	difference in a function between the field and average of random references: (...-Diff)=(...-Field)—(...-Aver)
...-Aver	average of values of random references standardized by number of sampling
...-Min	minimum value from random references
...-Max	maximum value from random references
...-Sign	*its sign*: the sign of difference between field and random values in the cases, *its value*: the value of significance level (*p*). $p = \frac{n - ND + 1}{n + 1}$, where *n*: number of randomizations, *ND*: number of differences between field and random values signed below
	in RSF file:
FD/Q	ST florula diversity: $H_j^{(\varphi)}$
Ld/Q	ST local distinctiveness: $H_j([L])$
Ass/Q	ST associatum: $I_j(\lambda)$
Com/Q	ST number of realized species combination: ω/m
Dis/Q	ST dissociatum: $H_j\{\delta_\lambda^{(s)}\}$
	in RSG file:
Entr/Q	ST local entropy of species A: $H_j(A)$
TAss/Q	ST total associativity of species A: $I_j(\langle A\rangle)$
	in RSH file:
Ass	pairwise association based on information theory: $mI(A,B)$ (as Inf-Ass in Table 8)

Acknowledgment

I am especially thankful to Sándor Bartha who helped my research from the initial steps in attainment knowledge of Juhász-Nagy’s methods, and showed me many methodical problems. I am also very thankful to János Podani, László Gallé and László Körmöczi for recension of the manuscript, and their useful notions. I am grateful to Szilvia Fóti and Nándor Horváth for participation in the field work, and to all my collegues, firstly Miklós Dombos who, using the program, helped its development. The research was supported by the Hungarian Scientific Research Fund (OTKA F020105 and T025335).

References

Bartha S. (1990): Spatial process in developing plant communities: pattern formation detected using information theory. - In: Krahulec F., Agnew, A. D. Q., Agnew, S. and Willems J. H. (eds.): Spatial processes in plant communities. pp. 31-47. Academic, Prague.

Bartha, S. (1992): Preliminary scaling for multi-species coalitions in primary succession. - Abstracta Botanica *16*, 31-41.

Bartha, S., Collins, S. L., Glenn, S. M. and Kertész, M. (1995): Fine-scale spatial organization of tallgrass prairie vegetation along a topographic gradient. - Folia Geobot. Phytotax., Praha *30*, 169-184.

Bartha, S., Czárán, T., Oborny, B., Podani, J. and Kertész, M. (1994): JNP-MODELLEK 1.0 - Számítógépes programcsomag a cönológia koegzisztenciális mintázatainak detektálására Juhász-Nagy Pál információstatisztikai modellcsaládjával. (JNP-MODELLEK 1.0 - Computer program package to detect coexistence patterns of cenology by Juhász-Nagy's information statistical model-family.) III. Magyar Ökológus Kongresszus, Szeged. Elõadások és poszterek összefoglalói, p.17.

Bartha, S., Czárán, T. and Oborny, B. (1995): Spatial constraints masking community assembly rules: a simulation study. - Folia Geobot. Phytotax., Praha *30*, 471-482.

Bartha, S. and Horváth, F. (1987): Application of long transects and information theoretical functions to pattern detection I. Transects versus isodiametric sampling units. - Abstracta Botanica *11*, 9-26.

Bartha, S. and Kertész, M. (1998): The importance of neutral-models in detecting interspecific spatial associations from 'trainsect' data. -Tiscia *31*, 85-98.

Erdei, Zs. and Tóthmérész, B. (1993): MULTI-PATTERN 1.00. Program package to analyze and simulate community-wide patterns. Tiscia *27*, 45-48.

Erdei, Zs., Tóthmérész, B. and Erdei, A. (1994): Linear algorithm to calculate indirect spatial statistics for completely random multi-species communities. - Tiscia *28*, 67-71.

Galiano, E. F., Castro, I. and Sterling, A. (1987): A test for spatial pattern in vegetation using a Monte-Carlo simulation. - Journal of Ecology *75*, 915-924.

Juhász-Nagy, P. (1967): On association among plant populations I. - Acta Biol. Debr. *5*, 43-56.

Juhász-Nagy, P. (1972a): Elemi preferenciális folyamatok információelméleti modellezése szünbotanikai objektumokon. (Information theory models of elemantary preferential processes on synbotanical objects.) Kandidátusi értekezés. Budapest.

Juhász-Nagy, P. (1972b): A növényzet szerkezetvizsgálata: új modellek. 1. rész. Bevezetés. (The structure of vegetation: new models. Part 1. Introduction.) - Bot. Közl. *59*, 1-5.

Juhász-Nagy, P. (1973): A növényzet szerkezetvizsgálata: új modellek. 2. rész. Elemi beskálázás a florális diverzitás szerint. (The structure of vegetation: new models. Part 2. Elementary scaling according to floral diversity.) - Bot. Közl. *60*, 35-41.

Juhász-Nagy, P. (1976): Spatial dependence of plant populations. Part 1. Equivalence analysis (an outline for a new model). - Acta Bot. Acad. Sci. Hung. *22*, 61-78.

Juhász-Nagy, P. (1980a): A cönológia koegzisztenciális szerkezeteinek modellezése. (Models of the cenological coexistence structures.) Akad. Dokt. Ért. Budapest.

Juhász-Nagy, P. (1980b): A növényzet szerkezetvizsgálata: új modellek. 3. rész. Florális diverzitás: elemek. (The structure of vegetation: new models. Part 3. The properties of floral diversity.) - Bot. Közl. *67*, 185-193.

Juhász-Nagy, P. (1984): Spatial dependence of plant populations. Part 2. A family of new models. - Acta Bot. Acad. Sci. Hung. *30*, 363-402.

Juhász-Nagy, P. (1985): A növényzet szerkezetvizsgálata: új modellek. 4. rész. Problémafelvetés az autocönológiában. (The structure of vegetation: new models. Part 4. Some problems of autocoenology.) - Bot. Közl. *72*, 1-15.

Juhász-Nagy, P. (1986): Egy operatív ökológia hiánya, szükséglete és feladatai. (Absence, necessity and tasks of an operative ecology.) Akadémiai Kiadó, Budapest.

Juhász-Nagy, P. and Podani, J. (1983): Information theory methods for the study of spatial processes and succession. - Vegetatio *51*, 129-140.

Margóczi, K. (1995): Interspecific associations in different successional stages of the vegetation in a Hungarian sandy area. - Tiscia *29*, 19-26.

Nosek, J. N. (1976): Comparative analysis of some diversity functions under different conditions of sampling in sandy meadow. - Acta Bot. Acad. Sci. Hung. *22*, 415-436.

Patil, G. P. and Taillie, C. (1979): An overview of diversity. - In: Grassle, J. F., Patil, G. P., Smith, W. and Taillie, C. (eds): Ecological Diversity in Theory and Practice. Internat. pp. 3-27.Cooper. Publ. House, Fairland, Maryland.

Palmer, M. W. and van der Maarel, E. (1995): Variance in species richness, species association, and niche limitation. - Oikos *73*, 203-213.

Podani, J. (1980): SYN-TAX: Számítógépes programcsomag ökológiai, cönológiai és taxonómiai osztályozások végrehajtására. (SYN-TAX: Computer program package for ordering data of ecology, cenology and taxonomy.) - Abstracta Botanica *6*, 1-158.

Podani, J. (1984a): Spatial processes in the analysis of vegetation: theory and review. - Acta Bot. Acad. Sci. Hung. *30*, 75-118.

Podani, J. (1984b): Analysis of mapped and simulated vegetation patterns by means of computerized sampling techniques. Acta Bot. Acad. Sci. Hung. *30*, 403-425.

Podani, J. (1988): SYN-TAX III. User's Manual. - Abstracta Botanica *12* Suppl. 1, 1-183.

Podani, J. (1992): Space series analysis of vegetation: processes reconsidered. - Abstracta Botanica *16*, 25-29.

Podani, J., Czárán, T. and Bartha, S. (1993): Pattern, area and diversity: the importance of spatial scale in species assemblages. - Abstracta Botanica *17*, 37-51.

Podani, J. and Czárán, T. (1997): Individual-centered analysis of mapped point patterns representing multi-species assemblages. - Journal of Vegetation Science *8*, 259-270.

Szollát, Gy. and Bartha, S. (1991): Pattern analysis of dolomite grassland communities using information theory models. - Abstracta Botanica *15*, 47-60.

Tóthmérész, B. (1993): DivOrd 1.50: a program for diversity ordering. - Tiscia *27*, 33-44.

Tóthmérész, B. (1994a): Diverzitási rendezések és térsorozatok. (Diversity ordering and spatial series analyses.) - Akad. Dokt. Ért. (DSC dissertation), Debrecen

Tóthmérész, B. (1994b): Statistical analysis of spatial pattern in plant communities. - Coenoses *9*, 33-41.

Tóthmérész, B. (1995): Comparison of different methods for diversity ordering. - Journal of Vegetation Science *6*, 283-290.

Tóthmérész, B. and Erdei, Zs. (1992): The effect of species dominance on information theory characteristics of plant communities. - Abstracta Botanica *16*, 43-47.

Tóthmérész, B. and Erdei, Zs. (1995): New features of MULTI-PATTERN 1.10: robust nonlinear smoothing. - Tiscia *29*, 33-36.

Zar, J. H. (1984): Biostatistical Analysis. 2nd edition. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.