Information theory methods for the study of spatial processes and succession*

P. Juhász-Nagy[1] & J. Podani[2,**,***]
[1] *Department of Plant Taxonomy and Ecology, L. Eötvös University, Budapest, Múzeum krt. 4/a, H-1088, Hungary*
[2] *Department of Plant Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada*

Abstract

The use of mathematical methods based on Shannon's entropy function is proposed for the evaluation of the consequences of sampling unit size and for the study of vegetation succession. The concept of diversity is extended to sets of phytosociological relevés under the term florula diversity. It is shown that Shannon's entropy as well as two other related characteristic functions can express the local behaviour and overall relationships of species. Characteristic areas are defined in terms of the maxima and minima of these functions. Several study areas yielded the data which are used in the examples. Some theoretical problems of the methods are discussed and a computer program, written in FORTRAN, is described.

Introduction

In an optimal phytosociological survey one should consider both spatial and temporal characteristics of plant communities. Spatial aspects, especially the different areas characteristic of the communities being studied, are of central importance, since they should be considered in selecting the sampling design and the method of data collection. Although it is generally accepted that the conclusions drawn from a vegetation study are influenced by the sampling design, the optimization of sampling is all but ignored in the phytosociological literature. Very few papers (e.g., Juhász-Nagy, 1967; Moravec, 1973; Fekete & Szöcs, 1974; and

Matthews, 1978, 1979) actually treat methods for the evaluation of the influence of sampling.

The information necessary for the evaluation of sampling can be obtained if successive changes are applied to the sampling procedure. These changes can usually be ordered into a well-defined series, such as the enlargement or the elongation of sample plots, the increase of the number of sampling units, etc. These series will be referred to as *spatial processes*. This definition differs from that of statistical ecology, where stochastic changes for generating spatial patterns are traditionally termed as spatial processes (cf. Cliff & Ord, 1981). However, the different terminology should not cause confusion, since the generation of spatial patterns is not discussed in this paper.

Temporal vegetation processes are never independent of spontaneous processes (e.g., propagation of plants) taking place in topographic space. Succession is therefore both a spatial and temporal process and will be termed as 'spatio-temporal' to clarify the terminology. In the sequel, the term 'spatial' will be restricted to processes related to sampling.

The study of spatio-temporal vegetation changes, based on permanent plots, is very extensive (see e.g., Falinski, 1977; van der Maarel, 1980). The difficulty with these studies is that the sampling properties are not considered, and usually a single, more or less arbitrarily selected plot size is used. This practice, however, can be criticized, since there is no guarantee that the plot size chosen will yield a sufficient and theoretically justifiable data base to reveal the successional trends. Whilst several models exist to describe and analyze succession (e.g., Horn, 1975, 1981; van Hulst, 1979; Gittins, 1981; Orlóci, 1981; Usher, 1981), spatial and spatio-temporal processes are not examined at the same time on the same object. The indirect mathematical methods applicable to the joint study of these processes have received little attention so far.

The present paper is an attempt to resolve some primary problems of the study of processes in vegetation analysis. The functions used are based on well-known information theoretical considerations (Kullback, 1959; Pielou, 1977; Orlóci, 1978; Feoli *et al.*, 1982). The most fruitful concepts of diversity are extended to sets of species, i.e., the floras of sampling units (The terms 'sampling unit' and 'sample plot' will be used interchangeably. 'Sample size', as a standard statistical term, will refer to the number of sampling units) or relevés (called *florulas*). The methods are applicable to floristic (binary) data only, but their extension to other types of data is conceivable. A more formal and mathematically explicit treatment of the topic is given elsewhere (Juhász-Nagy, 1967, 1976, 1982).

Materials

The discussion of the proposed techniques relies on results from five data sets. One of these comes from a study area situated at the edge of the Ravnako Alpine Basin near summit Vichren in the Pirin Mountains, Bulgaria. The sampling was completed in late June, 1962, and was based on ten sets of circular plots. The sampling radii are specified in Table 1. Each set contained 200 sampling units located at random in the study area. Only binary data were recorded. The community is a simple stand of a *Saxifraga stellaris* syntaxon, containing eight vascular species *(Ranunculus montanus, Nardus stricta, Saxifraga stellarius, Pin-*

guicula vulgaris, Viola biflora, Vaccinium vitis-idaea, Carex fusca, and *Deschampsia caespitosa).* This stand (at elevation 2400 m) is a depauperized form of a richer community of more than 100 species (including cryptogams) which occurs below 2200 m. Its syntaxonomic position is unclear, since it consists of a peculiar mixture of a transitional moor and a spring swamp ('Quellflur'). Juhász-Nagy (1963) discusses these problems and presents some results of statistical analyses.

Additional surveys were carried out between 1965 and 1967 based on a similar design in communities from three isolated locations in Hungary. These are: *Puccinellietum limosae* (Máta, Hortobágy National Park), a degraded stand of *Nardetum strictae* (Háromhuta, Sátor Mountains) and *Alopecuretum pratensis* (Rivaly, Beregi-sik). The number of species was 29, 34 and 73, respectively. The sample size was 256 in all cases.

For succession studies two sites at Nagycsere, Hungary, each of 100×200 m² size, were examined. The sites were cleared in the winter of 1956 for sylvicultural purposes. The complete plant cover, including roots, was removed. For a number of reasons, both sites were subsequently abandoned for several years. Sampling began in 1958 and ended in 1962, and resumed again in 1967 and 1968, after the establishment of two *Quercus robur* nurseries. Sampling was carried out twice a year, in mid-May and early September. Quadrats of $.1 \times .1$, $.2 \times .2$, $.3 \times .3$, . . ., 10×10 m² sizes were randomly placed at both sites. All plots were marked by coloured iron bars to ensure exact correspondence in their positions during the years of sampling. The most characteristic type of succession in this region is through the sere: *Brometum tectori* (annual grassland) → *Festucetum vaginatae* (perennial grassland) → *Festuco vaginatae – Quercetum roboris* (a forest-steppe climax). There are, of course, several other possible seres. Differences between the two sites are attributable in part to the fact that they are surrounded by different forest plantations *(Quercus robur* and *Robinia pseudacacia).* The present study is concerned only with the *Brometum tectori* stage.

In addition to these two sites, a nearby mature stand of *Brometum tectori* was sampled with similar plot sizes on one occasion in May, 1968. This was done to contrast the early phases of succession with a well-developed stand of the community.

Basic concepts

The diversity of a community is usually defined based on the assignment of plant individuals to taxa, phenological and life form types, or other classes. Radford *et al.* (1974) give a list of the various possibilities. In the simplest situation, diversity is defined as the number of taxa, life form types, etc., occurring in the whole community or within a unit area (i.e., the entropy of order one, cf. Rényi, 1961). In the ecological literature, diversity is often expressed as a function of the number of categories and the number of individuals within each category (see Pielou, 1975, 1977; Green, 1979, and Grassle *et al.*, 1979, for details). A literature review would certainly support our contention that the use of these measures is largely restricted to the study of taxon/individual diversity. However, as Dévai *et al.* (1971) have pointed out, this is just a special case. Another family of diversity functions may be obtained if the basic unit is not the individual but the set of all categories found in a certain sample area. Such a supraindividual entity is, for example, the florula of sample plots. It will be shown that conventional diversity measures, in particular Shannon's function, can be used to calculate the supraindividual diversity of communities if a sufficient number of plots are available.

It is noted that there have been several other attempts in the past to express community variation over an area or environmental gradient. These methods generally employ resemblance functions instead of diversity measures. The 'beta diversity' proposed by Whittaker (1967) is the amount of the species turnover from one end of a gradient to the other, and is in fact a function of inter-plot similarity. Another technique due to Gounot & Calléja (1962) and Moravec (1973) takes the average of the non-diagonal elements in the resemblance matrix of sample plots as an indicator of community heterogeneity. The averages obtained at different plot sizes are used in an attempt to find the 'minimal area' of the community.

Joint entropy as a measure of simple florula diversity

The concept of florula diversity is illustrated with two species, which is the simplest case. The joint entropy of species A and B is

$$m\hat{H}(A, B) = m \log m - a \log a - b \log b - c \log c - d \log d \tag{1}$$

Symbols a, b, c and d correspond to the notations of a 2×2 contingency table. $m = a + b + c + d$ is sample size. In this case (1) can be considered as the weighted estimate of Shannon's entropy for the set of possible combinations of the two species in question. The generalization of (1) to s species is given by the formula

$$m\hat{H}(A, B, \ldots, S) = m \log m - \sum_{k=1}^{\omega} f_k \log f_k \tag{2}$$

where f_k is the frequency of the kth species combination in the data set and $\omega = 2^s$ is the number of possible species combinations (different potential florulas). The base of the logarithm is arbitrary, but for binary data the use of $\log_2$ seems most appropriate. Log will always imply $\log_2$ in the sequel. Function (2) will be called the *simple florula diversity*, since several more complicated alternatives (e.g., Rényi's entropies of higher orders, see Rényi, 1961) are available to express uncertainty regarding the species composition of relevés. Furthermore, it is simple because no grouping or deletion of species is involved. It will be shown later that changes in florula diversity with changing plot size can be used to detect overall pattern in vegetation.

If all florulas are of the same composition, that is if one particular florula occurs m times, the florula diversity is zero. When all florulas are different, i.e., f_k is never greater than 1 for any k, the florula diversity reaches its local maximum:

$$\max \{m\hat{H}(A, B, \ldots, S)\} = m \log m = m\bar{H} \tag{3}$$

Since $m\hat{H}(A, B, \ldots, S)$ is a function of s, for comparative purposes it may be appropriate to use the relativized form,

$$m\hat{V}(A, B, \ldots, S) = m\hat{H}(A, B, \ldots, S)/m\bar{H} =$$
$$= 1.0 - (\Sigma f_k \log f_k) / (m \log m) \tag{4}$$

which will be termed the *florula evenness*. It ranges from 0 to 1, indicating minimum and maximum florula diversity respectively. It should be noted that the florula evenness estimates for two communities are comparable only if the sample sizes

(m) are equal, since evenness estimates are dependent on sample size (cf. Pielou, 1975).

Pooled entropy as a measure of local distinctiveness

Shannon's entropy function may be used to estimate the uncertainty of the event that a particular species occurs in a randomly chosen relevé. If species I is present in $n_{i.}$ of m sampling units, its entropy estimate is

$$m\hat{H}_i = m \log m - n_{i.} \log n_{i.} - (m-n_{i.}) \log (m-n_{i.}) \tag{5}$$

Thus, (5) yields valuable information concerning the local behaviour of species I. When $m\hat{H}_i$ is zero, i.e., it is present in, or absent from, all sampling units, species I obviously does not distinguish between relevés. Increasing values of $m\hat{H}_i$ indicate a stronger preference of the species to some relevés. The maximum value of (5) is m (provided that the base of the logarithm is 2), and will occur when the species is present in exactly half of the plots.

Summation over all species will give the pooled entropy estimate,

$$m\hat{H}([L]) = m\hat{H}_1 + m\hat{H}_2 + \ldots + m\hat{H}_i + \ldots + \\ + m\hat{H}_s = s \, m \log m - \Sigma \, (n_{i.} \log n_{i.} + (m-n_{i.}) \log (m-n_{i.})) \tag{6}$$

which wil be termed the *local distinctiveness* of species occurring in the study area. Formula (6) is used under a variety of names in the literature on numerical classification, e.g., information content (Cormack, 1971; Williams *et al.*, 1969) or total information (Sneath & Sokal, 1973). Van der Maarel (1979) argued that it is rather a heterogeneity measure; Podani (1980) used the term preferential information heterogeneity.

Contingency information as a measure of multiple association

In information theory the association between two species may be expressed in terms of their mutual information:

$$m\hat{I}(A,B) = m \log m + a \log a + b \log b + c \log c + \\ + d \log d - (a+c) \log (a+c) - (b+d) \log (b+d) - \\ - (a+b) \log (a+b) - (c+d) \log (c+d) \tag{7}$$

The higher its value the greater is the information obtained on species B by determining the score for species A, and vice versa. Function (7) does not distinguish between negative and positive associations; it is an absolute measure.

In general, the mutual information of s species (denoted by $m\hat{I}(\lambda)$) is termed the contingency information involving a table λ containing 2^s cells. This quantity may be simply derived by subtraction of $m\hat{H}(A,B,\ldots,S)$ from $m\hat{H}([L])$:

$$m\hat{I}(\lambda) = m\hat{H}([L]) - m\hat{H}(A,B,\ldots,S) \tag{8}$$

To distinguish between the two species case for which *association* has been used and the general situation, $m\hat{I}(\lambda)$ will be termed the *associatum* of s species. It follows from (6) and (8) that

$$m\hat{I}(\lambda) = (s-1) \, m \log m + \Sigma \, f_k \log f_k - \\ - \Sigma \, (n_{i.} \log n_{i.} + (m-n_{i.}) \log (m-n_{i.})) \tag{9}$$

For the sake of simplicity in writing formulae the following notations are introduced:

$$\Sigma \, n_{i.} \log n_{i.} = x \tag{10}$$

$$\Sigma \, (m-n_{i.}) \log (m-n_{i.}) = \bar{x} \tag{11}$$

and

$$x + \bar{x} = X \tag{12}$$

Spatial processes

In vegetation studies the sampling design is characterized by the size, shape, number and arrangement of sample plots, and the types of data collected. The successive alteration of one or more of the first four characteristics represents a process in real space. This process is static in time and artificial. For example, the elongation of square sampling units to longer and longer rectangles with the same area constitutes a process related to shape. In plotless sampling one can readily define similar processes. In case of a multiple nearest neighbour method (Williams *et al.*, 1969) the increase (or decrease) of neighbourhood radius constitutes a spatial process. The present paper is concerned only with spatial processes related to changing sample plot size.

However, the methods proposed here are also applicable to the others without much modification.

Entropy estimates are unbiased in a statistical sense only if the location of sample plots follows a random arrangement. Let U_0 be a set of distinct points randomized over the community to be sampled. If a series of plots (of a given shape) of sizes $a_0, a_1, \ldots, a_j, \ldots, a_w$ are arranged around each point such that $a_j < a_{j+1}$ for all j, and a_w is not greater than the size of the sampled area itself, $w + 1$ sets of sampling units are obtained:

$$U_0, U_1, \ldots, U_j, \ldots, U_w$$

Each set consists of m units. In this way the sampling yields a three-way, species by size by point matrix. All functions discussed earlier must be indexed by j to indicate their dependence on plot size.

Characteristic areas

In measuring taxon/individual diversity, the size of sample plots is a matter of statistical estimation. The larger the plot size the more precise the estimate will be, provided that the sample size remains unchanged. If the plot size is constant, an increase in sample size will result in higher precision. In case of supraindividual diversity, however, the interchangeability of plot and sample size no longer applies. An increase in plot size increases the probability of finding a given species within the plot; therefore, the florulas (i.e., the basic units themselves) are subject to change. These changes are influenced not only by the number of species and individuals but also by their spatial pattern and the associations between them. Consequently, the relationship between plot size and simple florula diversity is characteristic of the plant community. Figure 1 displays actual changes in simple florula diversity with changing plot size in *Nardetum strictae* and *Alopecuretum pratensis* communities. Table 1 summarizes the results for all three characteristic functions for the Ravnako Basin, showing that local distinctiveness and associatum also have maxima and minima. An idealized case for the three functions is illustrated in Fig. 2. The maxima and minima of these functions indicate particularly important plot sizes, termed the *characteristic areas*.

In fact, the concept of characteristic areas can be generalized to any other measures that are meaningful in reflecting some important property of
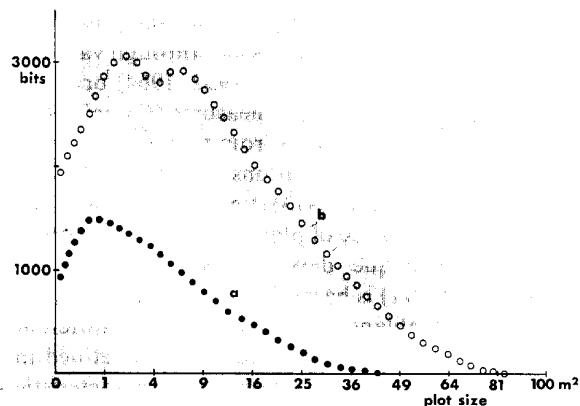

Fig. 1. The change of florula diversity by plot size in a) *Nardetum strictae* and b) *Alopecuretum pratensis*.
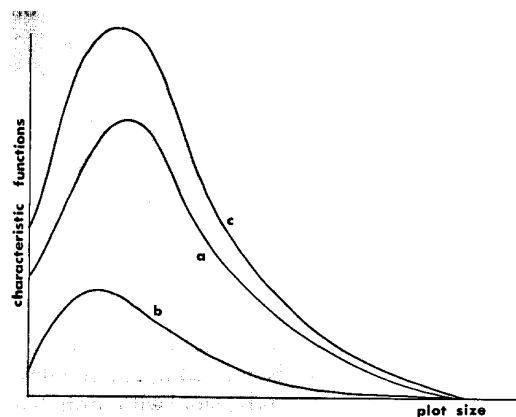

Fig. 2. An idealized diagram for three characteristic functions related to plot size. a) florula diversity, b) associatum, c) local distinctiveness.

Table 1. Values of three characteristic functions for data from Ravnako Basin.

Sampling radius (in meters)	Characteristic functions (in bits)		
	$m\hat{H}_j([L])$	$m\hat{H}_j(A, B, \ldots, S)$	$m\hat{i}_j(\lambda)$
0.0	353.16	326.00	27.16
0.1	729.36	586.69	142.67
0.3	869.87	615.24	254.63
0.5	885.36	656.13	229.23
0.7	879.10	709.01	170.09
0.9	700.86	538.77	162.09
1.1	344.53	319.47	25.06
1.3	168.51	164.61	3.90
1.5	60.00	60.00	0.00
1.7	0.00	0.00	0.00

vegetation. A good example is the analysis of the spatial pattern of a single species through variance estimators (cf. Greig-Smith, 1952, 1964) or information theory divergence measures (Orlóci, 1971; Feoli et al., 1982) derived from various plot sizes. For multi-species populations, Gounot & Calléja (1962) and Moravec (1973) suggested the use of the average resemblance of plots. The multiple pattern analysis technique developed by Noy-Meir & Anderson (1971) is based on the eigenanalysis of a series of resemblance matrices obtained at different plot sizes. Characteristic areas may be defined in many ways depending on the selected characteristic function. Also, there are many definitions in the literature without objective justification, according to the principles of different phytosociological schools. For these reasons, it is necessary to include a detailed discussion of characteristic areas based on florula diversity and related measures.

Minimum area

It was shown above that the lower and upper bounds for simple florula diversity are given by

$$0 \leqslant m\hat{H}_j(A, B, \ldots, S) \leqslant m \log m \qquad (13)$$

The plot size at which florula diversity and the local distinctiveness become 0 will be termed the minimum area, a_{min}, of the community. The minimum area of a species I is defined as the smallest plot size at which the species is found with probability 1. Clearly, a_{min} cannot be smaller than the largest of the s minimum areas. Consequently a_{min} is trivial (i.e., equals the study area) if there is at least one species represented by a single individual occurring at the edge of the study area. In other situations, a_{min} may be affected by the spatial segregation of negatively associated rare species.

The above definition is very conservative since it requires the complete identity of all florulas. Emphasis is placed on the rare species, which are usually regarded as being unimportant in defining characteristic areas. The traditional definitions are not so restrictive; the minimum area is defined as the smallest plot size which 'sufficiently' represents the characteristic structure and floristic composition of the community (cf. Cain & Castro, 1959; Goodall, 1961). In other words, an increase in the minimal area does not yield a 'significant' amount of additional information (Greig-Smith, 1964). Unfortunately, none of the methods developed is fully ob-

jective and generally applicable in determining such representative areas.

It is noted that the area of minimum associatum is not necessarily equal to a_{min}. As the data in Table 1 show, this area may be smaller if by chance $m\hat{H}_j([L]) = m\hat{H}_j(A, B, \ldots, S) > 0$. Despite this possibility, the area of minimum associatum is not discussed in the sequel.

Maximum areas

The objective of determining the minimum or representative area is to find quadrat size such that a relatively homogeneous sample of the community can be obtained for descriptive purposes. However, this characterization cannot be complete without considering another sample which represents the same community, but in its richest and most diverse development. This sample may be obtained by detecting plot sizes with maximal charcteristic functions.

Maximum area of florula diversity

The maximum of simple florula diversity requires the most detailed discussion, since inequality (13) may not be fulfilled in special cases. If $m < \omega$ and, hence, $\log m < s$, the inequality is satisfied. However, if $m \geqslant \omega$, when $\log m \geqslant s$, and the possible florulas have an equifrequency distribution such that $f_k = m/\omega$ for all k, the maximum will be simply $m\overline{H}_j = ms$ (in bits). It is to be noted that $m \geqslant \omega$ is a non-realistic condition since ω is usually a large number (recall that $\omega = 2^s$). Therefore it is very unlikely that this condition will ever be satisfied if, say, $s > 10$. Furthermore, the equifrequency distribution necessary to reach the maximum is even less likely to occur. Thus the upper bound ms, though sound in theory and attractive in its simplicity, can normally be regarded as an overestimate of some more realistic maximum. In trying to find such a maximum, it should be noted that the vast majority of potential florulas cannot manifest themselves empirically, that is, their frequency is zero. If we consider Ω as a random variable,

$$\Omega = \{ \omega_0, \omega_1, \ldots, \omega_j, \ldots, \omega_w \}$$

where ω_j is the number of different empirical florulas with non-zero frequencies at sampling unit size a_j, a better approximation of the upper bound is given by

$$m\hat{H}_j(A, B, \ldots, S) \leqslant \omega_j \log \omega_j \qquad (14)$$

The related florula evenness is

$$m\hat{V}_j(A, B, \ldots, S) = m\hat{H}_j(A, B, \ldots, S)/\omega_j \log \omega_j) \quad (15)$$

Note that (14) is the local maximum at the given sampling unit size. Beyond this, however, the global maximum over all plot sizes is of considerable interest. The sampling unit size where the global maximum occurs is the maximum area of florula diversity, a_{flor}. For the sake of simplicity we assume that there is only one a_{flor} (i.e., the curve is unimodal) and that this maximum is represented by a point rather than an interval (as in Fig. 2).

For small enough sample plots some combinations consisting of a few species will have high frequency values; this will result in low florula diversity. By increasing plot size to a certain point $a_j \simeq a_{flor}$, the number of species in each plot will continue to increase. In this way the previously similar relevés will become more distinct and $m\hat{H}_j(A, B, \ldots, S)$ will approach a global maximum. As the sampling unit size is further increased, the number of species in each relevé will also increase, and differences between relevés will again begin to disappear. This process ends when all sample plots have the same species composition. To clarify this point let us suppose that s is an even number and consider the following set of relations for binomial coefficients,

$$\binom{s}{0} < \binom{s}{1} < \ldots < \binom{s}{s/2} > \ldots > \binom{s}{s} \quad (16)$$

The number of possible combinations is maximal when $s/2$ elements are taken from s elements. It should be pointed out that the relations

$$\omega_0 < \omega_1 < \ldots < \omega_j > \ldots > \omega_w \quad (17)$$

are satisfied if the sampling starts with very small plots and ends with sufficiently large ones. This process may be regarded as a special partitioning process of the combinatorial type. It is the reason why truncated lognormal distributions frequently give a close fit to empirical data, as in Fig. 3 (see also Aitchison & Brown, 1966; Pielou, 1975).

In more complicated cases there may be a mixture of two (curve b, Fig. 1) or more lognormal distributions. When a_{flor} is not a point but an interval, equation (2) implies that it is likely that certain
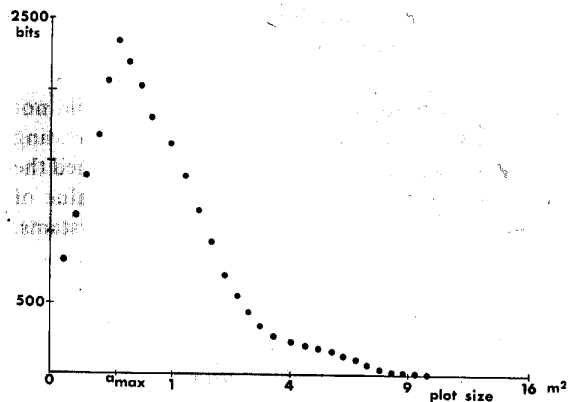

Fig. 3. Relation between florula diversity and plot size in *Puccinellietum limosae*.

species combinations will have persistently high frequency values within this interval (curve a, Fig. 1).

Compensatory area

Let us now examine some properties of local distinctiveness. Estimator (12) has limits

$$s\,m \log (m/2) \leqslant X_j \leqslant s\,m \log m \quad (18)$$

From (6) and (18) it follows that

$$0 \leqslant m\hat{H}_j([L]) \leqslant s\,m \quad (19)$$

Bounds are given in weighted bits for both relations. The lower bound in (18) is obtained only in the unique case when $x_j = \bar{x}_j = 0.5\,(s\,m \log (m/2))$. In this case, all species would be represented by an $m/2$ Raunkiaer frequency, such that

$$n_{i.} = m - n_{i.} = m/2 \quad (20)$$

This corresponds to what might be termed the ideal compensatory area, with all species showing an equifrequency distribution. This is evidently an unrealistic case since species normally show a high degree of variation in their pattern of commonness and rarity. Hence the upper bound in (19) is an overestimate of some more realistic value. Empirical results show that X_j follows a skewed U-distribution under the present sampling conditions. Therefore $m\hat{H}_j([L])$ always has a unique peak. The

point where local distinctiveness reaches its empirical maximum will be called the compensatory area, denoted by a_{comp}.

The discussion of compensatory area would not be complete without considering an interesting study by Williams *et al.* (1969). They examined the dependence of information content (the value of function (6)), on clump size in point clump systems. The clump size at which function (6) is maximal can be thought of as the compensatory clump size in a multiple nearest neighbour analysis.

The area of maximum associatum

Empirical results suggest that function $m\hat{I}_j(\lambda)$ has at least one maximum, at which the highest interspecific associations exist. This plot size corresponds to the area of maximum associatum, a_{ass}. If $m\hat{H}_j (A,B,\ldots,S)$ has two maxima then $m\hat{I}_j(\lambda)$ usually has two also. Thus $a_{ass}^{(1)}$ and $a_{ass}^{(2)}$ may be interpreted at first sight as areas dominated respectively by negative and positive associations.

Characteristic ordering

When characteristics areas have been determined for a given community it is interesting to examine their positional arrangement. In the case of very simple communities (see Table 1) the following relations hold,

$$a_{ass} < a_{comp} < a_{flor} < a_{min} \qquad (21)$$

Such an ordering of characteristic areas will be termed the *characteristic ordering* of a given community. Sometimes, of course the relation $<$ should be replaced by $\leqslant$, especially if the difference between two characteristic areas is very small.

If the curve of $m\hat{H}_j(A,B,\ldots,S)$ and/ or $m\hat{I}_j(\lambda)$ is not unimodal the characteristic ordering becomes fairly complicated. The ordering,

$$a_{ass}^{(1)} < a_{flor}^{(1)} <_{comp} < a_{ass}^{(2)} < a_{flor}^{(2)} < a_{min} \qquad (22)$$

may frequently occur, for instance in some types of meadows. Of course, the possibility of other orderings cannot be excluded.

A satisfactory interpretation of any characteristic ordering is very difficult. A combination of different approaches (comparison of classifications, simulation of patterns, topological study of plant

distributions, etc.) may help in this regard. It is also difficult to predict when to expect simple ordering versus a more complicated one. A hypothesis can be put forward by distinguishing some extreme cases, e.g.,

(1)   The community is open
(2)   The community is closed
(2a)   Patches are presumably classifiable
(2b)   Patches are presumably not classifiable

For the simply structured cases (1 and 2a) we can assume that a simple characteristic ordering exists.

Spatial processes and succession

In succession studies it is of considerable interest to investigate the behaviour of functions (2), (6) and (9). In the simplest case changes in the values of these functions can be studied in a primary succession using a fixed plot size. However, the result of such studies would be misleading if spatial processes were not taken into account; an investigation including both temporal and spatial processes has much more relevance.

Let $t_\ell (\ell = 1, \ldots, \nu)$ denote the point in time at which the ℓth sampling was carried out. The characteristic functions will have a second subscript ℓ to denote time. The result is best shown in three dimensions, but the process is sufficiently illustrated as a series of curves (Fig. 4). These diversity curves were calculated using the May data for the first study site at Nagycsere. Only three points in time (1958, 1962, 1967) will be considered to simplify the illustration.
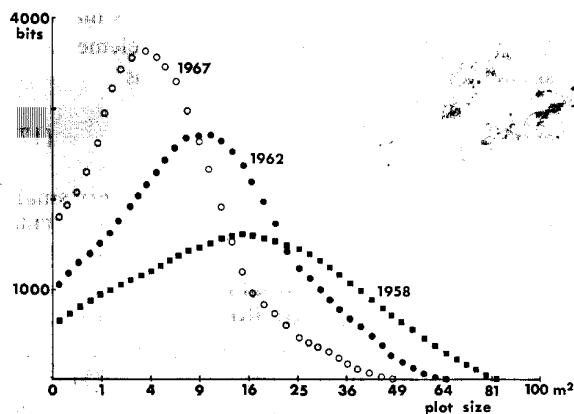

Fig. 4. The change of florula diversity curves during succession in a stand of *Brometum tectori.*

The interpretation of Figure 4 is not difficult. Let us first examine the florula diversities at plot size a_0 ('point quadrat'). It is striking that

$$m\hat{H}_{01} \ (A,B,\ldots,S) < m\hat{H}_{02} \ (A,B,\ldots,S) < m\hat{H}_{03} \ (A, B,\ldots,S) \quad (23)$$

This ordering results because many plots contain no species in the early stages of succession, which leads to low florula diversity. As succession continues the open areas are gradually colonized so that the probability that the point quadrat fails to touch an individual decreases. This results in higher florula diversity. Ordering (23) reflects temporal changes in pioneer communities. This is actually the case in the early *Brometum tectori* stage in grassland succession. It is likely that the ordering becomes more complicated and less interpretable later.

The role of plots lacking plants is of course just one factor to consider. During succession, especially in pioneer communities, the number of species usually increases over time. There is also evidence that species/individual diversity increases (e.g., Loucks, 1970; Auclair & Goff, 1971; Pineda *et al.*, 1981) at least during the early parts of the successional sere. These factors result in higher florula diversities at maximum areas. The results also suggest that the minimum and maximum areas become smaller as succession progresses. Figure 4 shows that as species number and total cover increase, the sampling unit size at which maximum and minimum florula diversity occur decreases. The results for a mature stand of *Brometum tectori* (see Fig. 5)
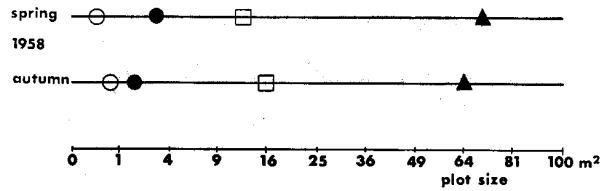
Fig. 6. Seasonal changes of characteristic areas in *Brometum tectori*. See Figure 5 for explanation of symbols.

support this in the case of a_{flor} and, for a_{comp} (not shown in Fig. 4). The area of maximum associatum increases slightly through time. The minimum area is considerably larger than the other characteristic areas. This feature may be explained by rare species.

Figure 5 also illustrates the characteristic ordering. Relation (21) holds true indicating its general validity for simple communities (cf. Table 1). It is also striking that a_{flor}, a_{comp} and a_{ass} become more similar over time, while a_{min} remains considerably larger. This ordering is also valid for the autumn data (see Fig. 6) although two characteristic areas are larger, and two others smaller, than in the spring. This seasonal oscillation is attributable to differences in the physical appearance of spring and autumn vegetation.

Results for the second study site at Nagycsere (not shown) are much more complicated. Florula diversity and associatum usually have two maxima, sometimes three as in $m\hat{I}_{j\varrho}(\lambda)$ for autumn, 1962. The ordering of characteristic areas varies from year to year, and the positional transformations are rather difficult to interpret. The reason is perhaps the extensive spread of ruderals (e.g., *Lamium purpureum*) since 1960. Further analysis will be required to clarify this problem.

Discussion

The present paper shows that the joint information of species in a binary phytosociological matrix has meaning in reflecting the compositional diversity within sets of relevés. Results based on actual data indicate strong dependence of simple florula diversity on plot size. Similar relationships are shown between plot size and two other information theoretical functions. The sampling unit size at which a function is maximal or minimal represents
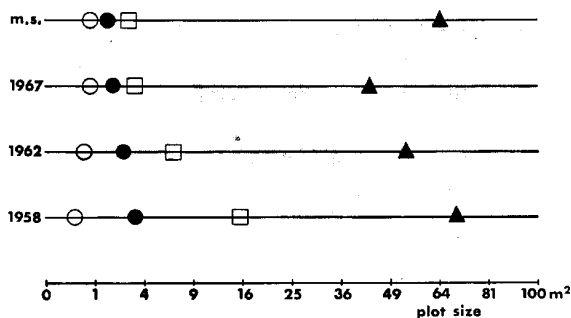

Fig. 5. The characteristic ordering in three points of time during succession in *Brometum tectori*. The uppermost row (m.s.) indicates the characteristic ordering obtained for a mature stand of this community. Symbols identify characteristic areas as follows: O: a_{ass}, ●: a_{comp}, □: a_{flor}, ▲: a_{min}.

a characteristic area of the community with respect to the function. Four characteristic areas seem to be of importance: minimum area, maximum area of florula diversity, compensatory area and maximum area of associatum.

The minimum area is the smallest plot size at which all relevés would contain all species present in the community stand being examined. This is mostly influenced by the spatial distribution of rare species. The maximum areas are ecologically more meaningful since they depend on all species, and the common ones are decisive. At the maximum area of simple florula diversity the chance that two relevés have the same species composition is minimal. The compensatory area corresponds to the plot size at which the pooled entropy of species is maximal, i.e., when the species are most distinctive among plots. The highest overall association of species is indicated at the area of maximum associatum.

The determination of these maximum areas may be useful in pilot studies before ordination and classification studies are performed. It seems that multivariate analyses are most efficient if the sample contains as much information about community heterogeneity as possible. Such an optimal sample can be obtained by using a plot size falling within the interval $[a_{ass}, a_{flor}]$ which usually encloses a_{comp}. The usefulness of information theory functions in determining optimal plot size, however, should be verified by their simultaneous application with other characteristic functions and multivariate analyses.

The ordering of characteristic areas is invariant in the simple communities examined. The smallest characteristic area corresponds to the maximum associatum between species. The plot size with maximum local distinctiveness and florula diversity is usually larger than the area of maximum associatum. The minimum area, when all differences between relevés disappear so that characteristic functions have a zero value, tends to be relatively large.

It is observed that characteristic areas change as succession progresses. However, the characteristic ordering does not change significantly in the early stages of a secondary succession. In some cases the characteristic ordering and its changes in time may be complicated owing to the high complexity of the community.

Some problems may arise when estimating florula diversity. In species poor communities a few hundred plots should suffice. In species rich communities, however, several thousand plots may be needed to obtain a reasonably precise estimate of diversity. Therefore, the extension of florula diversity to subsets of species may be necessary. Such subsets may be chosen at random or based on a ranking procedure.

In addition, it is possible that the use of different sizes of sampling units is insufficient for the proper screening of characteristic areas. In this case, iterative data collection should be carried out. However, it seems very difficult to do any iterative sampling in the field, since sampling may disturb the site. A promising approach to solving this problem is computerized sampling based on digitized field data (e.g., La France, 1972; Szöcs, 1979). Since both traditional and computerized sampling may be very expensive, the optimal choice of sample size in relation to the number of species is of great importance. A method to find such an optimum still awaits development.

Regarding structure and succession, the joint application of characteristic functions with traditional species/individual diversity indices may be more informative than a straight diversity study. Furthermore, the use of characteristic functions should not be restricted to studies in alteration of plot size. Extensions to the change of plot shape and arrangement represent future possibilities.

Appendix

The FORTRAN program INPRO calculates actual values for functions (2), (4), (6) and (9) for phytosociological data matrices. The program accepts any kind of data which is subsequently transformed to binary form according to

$n_{ij} = 1$, if $x_{ij} > 0$
$n_{ij} = 0$, if $x_{ij} \leqslant 0$

Analysis of several data sets may be performed in a single run. The first card contains input format for all matrices. F specification should be used even if the type of data is integer. The following block of input and data cards may be repeated any number of times

Card 2. Cols 1–3   Number of species
          Cols 4–7   Number of sample plots

Card 3. Cols 1–80 Title
Data cards   Each species should start on a
                new card and should be con-
                tinued on additional cards as
                necessary

The last card of the input deck should be a blank
card.

The program was implemented on an IBM 370
computer. A listing is available from the second
author upon request.

References

Aitchison, J. & Brown, J. A. C., 1966. The lognormal distribu-
    tion. Cambridge University Press, London.
Auclair, A. N. & Goff, F. G., 1971. Diversity relations of upland
    forests in the western Great Lakes area. Am. Nat. 105:
    499–528.
Cain, S. A. & Castro, G. M. de O., 1959. Manual of vegetation
    analysis. Harper, New York.
Cliff, A. D. & Ord, J. K., 1981. Spatial processes. models and
    applications. Pion, London.
Cormack, R. M., 1971. A review of classification. J. Roy. Stat.
    Soc. series A 134: 321–367.
Dévai, I., Horváth, K. & Juhász-Nagy, P., 1971. Some problems
    of model building in synbiology. Part 1. Spatial diversity
    process of the binary type in a simple situation. Ann. Univ.
    Sci. Budapest. sectio Biol. 13: 19–32.
Falinski, J. B. (ed।)., 1977. Vegetation dynamics. Proceedings of
    the third symposium of the working group on succession
    research on permanent plots. Phytocoenosis 7.
Fekete, G. & Szöcs, Z., 1971. Studies on interspecific associa-
    tion processes in space. Acta. Bot. Acad. Sci. Hung. 20:
    227–241.
Feoli, E., Lagonegro, M., & Orlóci, L., 1982. Information analy-
    sis in vegetation research (manuscript in press).
Gittins, R., 1981. Towards the analysis of vegetation succession.
    Vegetatio 46: 37–59.
Goodall, D. W., 1961. Objective methods for the classification of
    vegetation. IV. Pattern and minimal area. Aust. J. Bot. 9:
    162–196.
Gounot, M. & Calléja, M., 1962. Coefficient de communauté,
    homogeneité et aire minimale. Bull. Serv. Carte Phytogéogr.
    Ser. B. 7: 181–200.
Grassle, J. F., Patil, G. P., Smith, W. K. & Taillie, C., (eds.).,
    1979. Ecological diversity in theory and practice. Statistical
    ecology, vol. 6. International Co-operative Publ. House,
    Fairland, Maryland.
Green, R. H., 1979. Sampling design and statistical methods for
    environmental biologists. Wiley, New York.
Greig-Smith, P., 1952. The use of random and contiguous quad-
    rats in the study of the structure of plant communities. Ann.
    Bot. 16: 293–316.
Greig-Smith, P., 1964. Quantitative plant ecology. Butter-
    worths, London.

Horn, H. S., 1975. Markovian properties of forest succession.
    In: M. L. Cody & J. M. Diamonds (eds.), Ecology and
    evolution of communities. pp. 196–211. Harvard Univ.
    Press, Cambridge, Massachusetts.
Horn, H. S., 1981. Succession. In: R. M. May (ed.), Theoretical
    ecology. 2nd ed., pp. 253–271. Sinauer, Sunderland.
Hulst, R. van, 1979. On the dynamics of vegetation: Markov
    chains as models of succession. Vegetatio 40: 3–14.
Juhász-Nagy, P., 1963. Investigations on the Bulgarian vegeta-
    tion. Some hygrophilous plant communities I-III. Acta Biol.
    Debrecina 2: 47–70.
Juhász-Nagy, P., 1967. On some 'characteristic areas' of plant
    community stands. Proceedings of the Colloqium on Infor-
    mation Theory, Bolyai Mathematical Society (ed. A. Rényi).
    Akadémiai Kiadó, Budapest. pp. 269–282.
Juhász-Nagy, P., 1976. Spatial dependence of plant popula-
    tions. Part I. Equivalence analysis (an outline for a new
    model). Acta Bot. Acad. Sci. Hung. 22: 61–78.
Juhász-Nagy, P., 1982. Spatial dependence of plant popula-
    tions. Part II. A family of new models. Acta Bot. Acad. Sci.
    Hung. (in press)
Kullback, S., 1959. Information theory and statistics. Wiley,
    New York.
LaFrance, C. R., 1972. Sampling and ordination characteristics
    of computer-simulated individualistic communities. Ecology
    53: 387–397.
Loucks, O. L., 1970. Evolution of diversity, efficiency, and com-
    munity stability. Am. Zool. 10: 17–25.
Maarel, E. van der, 1979. Multivariate methods in phytosociol-
    ogy, with reference to the Netherlands. In: M. J. A. Werger
    (ed.), The study of vegetation. pp. 161–225. Junk, The Hague.
Maarel, E. van der (ed.), 1980. Advances in Vegetation
    Science: Succession. Vegetatio 43: 1–151.
Matthews, J. A., 1978. An application of non-metric multidi-
    mensional scaling to the construction of an improved species
    plexus. J. Ecol. 66: 157–173.
Matthews, J. A., 1979. A study of the variability of some succes-
    sional and climax plant assemblage types using multiple
    discriminant analysis. J. Ecol. 67: 255–271.
Moravec, J., 1973. The determination of minimal area of phyto-
    coenoses. Fol. Geobot. Phytotax. 8: 23–47.
Noy-Meir, I. & Anderson, D. J., 1971. Multiple pattern analysis,
    or multiscale ordination: towards a vegetation hologram? In:
    G. P. Patil, E. C. Pielou & W. E. Waters (eds.), Many species
    populations, ecosystems, and systems analysis. Stat. Ecol. 3.
    pp. 207–225. Pennsylvania State Univ. Press.
Orlóci, L., 1971. An information theory model for pattern anal-
    ysis. J. Ecol. 59: 343–349.
Orlóci, L., 1978. Multivariate analysis in vegetation research.
    2nd ed. Junk, The Hague.
Orlóci, L., 1981. Probing time series vegetation data for evidence
    of succession. Vegetatio 46: 31–35.
Pielou, E. C., 1975. Ecological diversity. Wiley, New York.
Pielou, E. C., 1977. Mathematical ecology. Wiley, New York.
Pineda, F. D., Nicolas, J. P., Ruiz, M., Peco, B. & Bernaldez, F.
    G., 1981. Succession, diversité et amplitude de niche dans les
    pâturages du centre de la péninsule ibérique. Vegetatio 47:
    267–277.

140

Podani, J., 1980. SYN-TAX: Computer program package for cluster analysis in ecology, phytosociology and taxonomy. Abstracta Botanica (Budapest) 6: 1–158. (in Hungarian with English summ.)

Radford, A. E., Dickison, W. C., Massey, J. R. & Bell, C. R., 1974. Vascular plant systematics. Harper and Row, New York.

Rényi, A., 1961. On measures of entropy and information. In: J. Neyman (ed.), Proceedings of the 4th Berkeley Symposium on Mathematical statistics and probability. pp. 547–561. Berkeley Univ. Press.

Sneath, P. H. A. & Sokal, R. R., 1973. Numerical taxonomy. Freeman, San Francisco.

Soó, R., 1964. A magyar flóra és vegetáció rendszertani – növényföldrajzi kézikönyve (Synopsis systematico-geobotanica florae vegetationisque Hungariae). vol. 1. Akadémiai Kiadó, Budapest.

Stojanov, N., Stefanov, B. & Kitanov, B., 1966-67. Flora Bulgarica. 4th ed. Nauka i Iskustvo, Sophia.

Szöcs, Z., 1979. New computer-oriented methods for the study of natural and simulated vegetation structure. In: L. Orlóci, C. R. Rao & W. M. Stiteler (eds.), Multivariate methods in ecological work. Statistical Ecology, vol. 7. pp. 301–308. International Co-operative Publ. House, Fairland, Maryland.

Usher, M. B., 1981. Modelling ecological succession, with particular reference to Markovian models. Vegetatio 46: 11–18.

Whittaker, R. H., 1967. Gradient analysis of vegetation. Biol. Rev. 42: 207–264.

Williams, W. T., Lance, G. N., Webb, L. J., Tracey, J. G. & Connell, J. H., 1969. Studies in the numerical analysis of complex rain-forest communities. IV. A method for the elucidation of small-scale forest pattern. J. Ecol. 57: 635–654.