Spatial Scaling of Structural Complexity in Plant Communities

CARLO RICOTTA1 AND MADHUR ANAND2
1Department of Plant Biology, University of Rome "La Sapienza", Italy
2Canada Research Chair in Biocomplexity of the Environment, Biology Department, Laurentian University, Sudbury, Ontario, Canada
* Author for correspondence (Email: carlo.ricotta@uniroma1.it)

ABSTRACT

In the last decade, it has become well appreciated that disorder-based diversity as a single scalar quantity fails to adequately capture structural complexity in biological communities. Complexity here implies spatial correlation (co-occurrence) between species such that the larger and more intricate the correlations, the more structured the community. Ideally, we should like to measure complexity in terms suitable to compare, say, a rainforest to grassland. But for this, the rules of quantification must be very general. Also, since complexity should summarize the amount of spatial correlation, any complexity measure is by its very nature scale dependent. Taking these into consideration we suggest that Juhász-Nagy information-theoretical coenostate variables offer an adequate basis for the measurement of the structural complexity within plant communities. We examine the notion of complexity in its broad sense and present a worked example using data from an early stage primary succession on an open-cast coal mine in northern Hungary.

Key Words: Diversity, Entropy, Information, Measure, Statistical complexity.

INTRODUCTION

Quantitative ecology has long possessed a general uncertainty or disorder-based measure applicable to sample species abundance vectors: Shannon's (1948) entropy in the underlying distribution, a quantity originally introduced in ecological work by McArthur (1955) and Margalef (1958) almost 60 years ago. Take a plant community composed of N species where the relative abundance of the i-th species (measured as biomass, cover, number of individuals, or any other biological variable that is thought to influence ecosystem functioning) is denoted by \( p_i \), such that \( 0 \leq p_i \leq 1 \), and \( \sum_{i=1}^{N} p_i = 1 \). Shannon's classical entropy function \( H \) is computed from the relative abundance vector \( p = (p_1, p_2, ..., p_N) \) as:

\[
H = -\sum_{i=1}^{N} p_i \log_2 p_i
\]

(1)

It is important to note that this is a special case (order one) of the more general form of entropy definition called Rényi's (1961) generalized entropy. For a given number of species, entropy is maximal \( H_{\text{max}} = \log N \) when \( p_i = 1/N \) for all \( i \) and minimum when there is a dominant species and all other species are represented by one individual \( H_{\text{min}} = 0 \). In Shannon's original definition, the base 2 logarithm is used to measure information content in bits. From an ecological viewpoint, the Shannon entropy \( H \) of a given community is a measure of uncertainty in predicting the relative abundances of species.

The important role of Shannon's entropy in ecological work was already recognized by McArthur (1955) and Margalef (1958). Since then, it has rapidly become one of the most popular measures of disorder-based community structure (Orlóci 1991, Ricotta 2003), not the least for its generalizations by Rényi and suitability for additive partitioning (Orlóci 1978, Orlóci et al. 2002). However, in the last decades, it has become clear that \( H \) as a measure of the complexity in a biological community (or, more generally, of a given system) is less than optimal (Anand and Orlóci 1996). While this is so, when \( H \) is used in solo, the problem can be ameliorated by its suitability for additive partitions and its connectedness with Rényi's information of order 2 that creates a basis upon which to develop an full-fledged system of statistical analysis (Kullback 1959).
 Complexity here is roughly defined as the amount of 'correlation' between the system components, which generates pattern. Intuitively, the larger and more intricate the correlations between the system components, the more structured the analysed system (Feldman and Crutchfield 1998). Nonetheless, species performance being a stochastic variable, structure and correlation are never completely void of randomness. Interestingly, although it is broadly held true that both maximally random and perfectly ordered systems possess no structure (Grassberger 1986, Crutchfield and Young 1989, Anand and Orlóci 1996), it is also true that pattern other than random can arise from under the total reign of chance (the case of the fractal). But ecologists are interested more in the case of stochastics where random variation is superimposed on a deterministic relationship, in which case the former is to be isolated from the latter.

There is still no general agreement as to what defines structural complexity (Calude 2002), however, a number of measures have been proposed to quantify the degree of structure or pattern present in a given system. It has become convenient to distinguish 'statistical complexity' measures from algorithmic complexity measures, such as the Kolmogorov (1965) complexity (also known as 'algorithmic information', 'algorithmic entropy', 'Kolmogorov-Chaitin complexity', 'descriptive complexity', 'algorithmic randomness'), which is defined as the length of the shortest code that describes the sample in the sense as Anand and Orlóci (1996) applied it. Kolmogorov (1965) complexity incorporates accounting for every bit - random or not, which contrasts with other complexity measures that attempt to remove randomness and thus measure the regularity in the residual something along the lines as Feldman and Crutchfield (1998) explained.

Lopez-Ruiz et al. (1995) defined a measure of statistical complexity CLMC as:

\[ C_{\text{LMC}} = H \times D \]  

(2)

where \( H \) is the Shannon entropy and the quantity \( D = \sum_{i=1}^{N} (p_i - 1/N)^2 \) is the "Discernible" measure of the departure of \( p_i \) from uniformity. It is observed by them that \( C_{\text{LMC}} \) vanishes for distributions that correspond to perfect order and maximal randomness. For instance, perfect order corresponds to zero Shannon entropy, while maximal randomness occurs in equiprobable distributions (\( p_i = 1/N \) for all \( i \)). Hence, by Eq. (2) in both cases \( C_{\text{LMC}} \) equals zero.

More recently, Feldman and Crutchfield (1998) criticized the behaviour of the \( C_{\text{LMC}} \) index and proposed an alternative measure of complexity known as the Kullback-Leibler information-theoretical distance. The information gain between two probability distributions \( p \) and \( q \) is defined as the one-way divergence (\( D(p \parallel q) \)) defined by Kullback's (1959) \( D(p \parallel q) = \sum_{i=1}^{N} p_i \log (p_i/q_i) \). This quantity is not a true distance function, neither satisfying the triangle inequality nor being symmetric. If symmetry is required it is provided by Kullback's (1959) \( D(p \parallel q) \), where \( p_i = 1/N \) for all \( i \), provides an adequate measure of the departure of the probability vector \( p \) from uniformity. Thus, the statistical complexity measure proposed by Feldman and Crutchfield (1998) is defined as:

\[ C' = H \times D(p \parallel \hat{p}) \]  

(3)

Again, it is easily shown that by construction \( C' \) vanishes in the extreme ordered and disordered limits. For a review of the physical literature on complexity measures, see Feldman and Crutchfield (1998). For the theory and applications of information analysis in ecology, Feoli et al. (1984), Orlóci (1991), Orlóci et al. (2002), and Ricotta (2003) may be consulted.

As an extension to the entropy-based approach, following Papentin's (1980) division of complexity into 'organized' and 'unorganized' complexity, Anand and Orlóci (1996) proposed a new measure for structural complexity based on communication and coding theory. For a given community characterized by its relative abundance vector \( p \), they defined structural complexity as the difference between Shannon's entropy and Huffman's (1952) average length (\( L \)) of the communication-theoretical parsimonious code required to describe the community:

\[ \Delta = L - H = \sum_{i=1}^{N} p_i l_i + \sum_{i=1}^{N} p_i \log p_i \]  

(4)

where \( l_i \) is the length of the codeword (number of symbols) necessary to encode the relative abundance \( p_i \) (for details, see Anand and Orlóci 1996, 2000). Since \( L \geq H \), where the equality holds for equiprobable distributions, it follows that for \( p_i = 1/N, \Delta = 0 \). This suggests that at maximum disorder there is no room for structure. When \( p_i = 1 \) and \( p_i = 0 \) (i.e.), \( L \) is not defined; that is, the community must have at least two species for the Huffman encoding. Being additive and based on Kolmogorov's notion of algorithmic complexity, the Anand and Orlóci (1996) measure conceptually differs
from the statistical complexity measures developed by Lopez-Ruiz et al. (1995) and Feldman and Crutchfield (1998).

Independently of their mathematical formulation, the complexity measures described above are a function only of the relative abundance vector \( p \). It may not be the case that all systems with the same relative abundance vector \( p \) have the same structural complexity. Orlóci (1971) was the first to use an information-theoretical model, implementing a multiscalar nested decomposition of species data in the study of spatial pattern in plant communities. The pioneering studies of Juhász-Nagy (1976, 1984, 1993) went further and may offer a more pragmatic approach to the measurement of statistical complexity. In the present paper, we discuss an application of the Juhász-Nagy approach with data from an early stage of primary succession on open-cast coal mine on the Mátra foot hills in northern Hungary. Our justification is the suitability of the Juhász-Nagy conceptual framework to summarize various aspects of spatial diversity and spatial dependence in plant communities at different spatial scales (see also Juhász-Nagy and Podani 1983).

**Juhász-Nagy Information-Theoretical Approach**

Juhász-Nagy applied a family of information-theoretical functions to a series of data matrices, each summarizing presence/absence of plant species in a set of sample plots of a given size. One of Juhász-Nagy's most important proposal is the application of Shannon's entropy to the frequency distribution of species combinations. Suppose that a plant community composed of \( N \) species is sampled with \( M \) plots of size \( s \). The realized species combination in each plot is recorded and the compositional diversity or *Florula Diversity* (FD) at plot size \( s \) of the community is estimated as the Shannon entropy in the frequency distribution of the realized species combinations within the \( M \) plots:

\[
FD = \sum_{k=1}^{\omega} p_k \log_2 p_k
\]

(5)

where the term *florula* indicates the realized species combination in the sample plot (distinct from *flora*, which is the collection of all species present), \( \omega = 2^N \) is the number of possible species combinations (i.e., the number of possible florulas), and \( p_k \) is the frequency of the \( k \)-th species combinations in the florula.

*Florula* diversity (a coenological diversity) is a scalar of overall species structure in a community. By changing plot size \( s \), Florula Diversity can be plotted as a function of scale, thus obtaining a *spatial process sensu* Juhász-Nagy and Podani (1983) or *spatial series sensu* Podani (1992). It is easily shown that FD is strongly scale dependent and shows a peaked effect when displayed against plot size, an extension in the direction of the Greig-Smith type pattern analyses (Greig-Smith 1983). For small plots, FD is generally low because most of the plots are either empty or contain only a few species. Likewise, very large plots tend to include almost all species, and the resulting FD will be low again.

Between the extremes, Florula Diversity takes at least one maximum value; the area pertaining to its maximum value reflects the scale point at which the community reaches its highest variation, thus considered optimal for community characterization. Changes of this maximum area over time have been used for example to evaluate different aspects of secondary succession in abandoned fields (Juhász-Nagy and Podani 1983) and of revegetation processes on open-cast coal mine tailings in northern Hungary (Bartha 1992).

Whereas Florula Diversity is suitable to express the compositional diversity of a given plant community, it is not informative on the spatial dependence of species patterns. To overcome this drawback, an additional measure of spatial diversity is used. This measure is defined as the sum of Shannon's \( H_i \) calculated separately for each \( i \)-th species based on its presence and absence in the plots (see Eq. 1):

\[
H_i = -p_i \log_2 p_i -(1 - p_i) \log_2 (1 - p_i)
\]

(6)

where \( p_i \) is frequency of the presence of the \( i \)-th species in the sample containing plots of size \( s \). \( H_i \) is a measure of the uncertainty as to the presence of the \( i \)-th species in a randomly located plot. Hence, the sum of single-species entropies \( \sum_{i=1}^{N} H_i \) gives the pooled entropy of the whole species collection, and is proportional to the uncertainty that any randomly selected plant species is found in a random plot, assuming that the shared I-divergence or mutual information with other species is negligible. Orlóci (1991) discusses the topic in detail and gives partition functions (see also Orlóci et al. 2002) on which the expansion of the analysis to involve interactions with other species, spatial and other environmental sorting criteria can be fashioned.

The pooled entropy, \( \sum_{i=1}^{N} H_i \), reaches its maximum when all species are present in exactly half of the plots, which is obviously an unrealistic constraint, since species normally show a high degree of variation in their commonness and rarity. Mathematically, the pooled entropy is the marginal entropy of the \( 2^N \) contingency table, and its scale dependence as a
function of plot size is similar to that of Florula Diversity, showing a peaked effect at intermediate plot sizes. Two plots, on the average, differ in the greatest number of species at the size where the pooled entropy attains its maximum value. Hence the name Local Distinctiveness (LD) suggested by Juhász-Nagy, but other names such as information content (Williams et al. 1969) or total information (Orlóci 1970, Sneath and Sokal 1973) also appear in the biological literature. Nonetheless, from a mathematical viewpoint, it is perhaps most straightforward to use the term pooled entropy, because this measure is additive, being obtained as the sum of single species entropies (Podani et al. 1993).

Juhász-Nagy further recognized that if the single species patterns are spatially independent, Local Distinctiveness equals Florula Diversity. That is, in a random community (i.e., when there are no associations among the species), the frequencies of species presences and species absences will provide directly the probabilities of species combinations (Bartha et al. 1998). Therefore, the expectation of Florula Diversity in a random community equals the pooled entropy of species.

To the contrary, in actual communities in which randomness is not satisfied because species are spatially segregated or aggregated, Florula Diversity is smaller than Local Distinctiveness. Hence, the information on species spatial dependence, termed Associatum (AS) by Juhász-Nagy, can be conveniently measured as the difference between Local Distinctiveness and Florula Diversity:

\[ AS = LD - FD \] (7)

Mathematically, the Associatum represents the mutual information of N binary variables, or the information of the 2\(^n\) contingency table and reflects the amount of uncertainty on the overall interspecific associations within the community. Empirical results suggest that, like Florula Diversity and Local Distinctiveness, the Associatum also shows a peaked behavior with at least one maximum at intermediate plot sizes, although the maximum values of these three functions are not necessarily positioned at the same plot size.

An Example

The construction of Juhász-Nagy coenological functions is shown using a small data set obtained from an early stage of primary succession on open-cast coal mine tailings in Visonta, northern Hungary. By removing top soil and covering large areas with a mineral substrate, open-cast mining activity creates terra nova conditions for a primary succession of pioneer vegetation (Szegi et al., 1988). Vegetation succession starts with an approximately random pattern of the first colonizing individuals of Tussilago farfara, Lactuca serriola and Matriaria inodora. After 3-4 years, the vegetation develops into a patchwork of annual grasses (for example, Bromus japonicus) and short-lived perennial herbs (mostly legumes, for example, Medicago lupulina and Melilotus officinalis). After 10 years the area is covered by a closed, homogeneous grassland dominated by Agropyron repens (for more details, see Bartha, 1992).

In this paper, micromaps of species distribution within a two-year old stand composed of a 64×96 grid of 20×20 cm cells are used to illustrate the ideas discussed. The position of all individual plants within the grid were recorded and then digitized separately for each species to yield the maps shown in Figure 1. From these micromaps, Juhász-Nagy information-theoretical functions were computed at different block sizes ranging from 20 to 320 cm (Figure 2).

DISCUSSION

As shown in Figure 2, Florula Diversity and Local Distinctiveness reach their highest variation at a block size of 80 cm. Associatum also shows a peaked behaviour, albeit with a maximum at a block size of 160 cm. That is, for the community of our example, the Shannon information on species spatial dependence is maximized at a block size of 160 cm. Within the context of statistical complexity, it is easily demonstrated that, in the ideal case of a random community in which the single species patterns are spatially independent, the mutual information between them vanishes, and AS = 0 (see Eq. 7). On the other hand, if the spatial diversity of the analysed community is known with certainty (i.e., if all cells of a given micromap contain the same species composition), then FD = LD = 0, and, consequently, the Associatum also equals zero. These extremes correspond to AS vanishing in the cases of ideal randomness and trivial predictability, respectively. Between these extremes, for non-random communities in which species are spatially segregated or aggregated, the Associatum is positive and measures the amount of (horizontal) spatial correlation between the system components, also defined as community structural complexity.

As pointed out by Anand and Orlóci (1996), in Koppel's (1987) frame of reference, 'total complexity' includes Shannon's entropy but also another component, 'meaningful complexity', that captures structure. In this view, we may think to AS as structural complexity, to compositional diversity FD as 'disorder-based' complexity, and to LD as 'total complexity'.
Figure 1. Micromaps of species distribution within a two-year old stand composed of a 96×64 grid of 20×20 cm cells on an open-cast coal mine in Visonta, northern Hungary. The numbers show the count of occupied (i.e., black) 20×20 cm cells within the micromaps.

Figure 2. Diagram of three characteristic functions, Florula Diversity (FD), Local Distinctiveness (LD), and Associatum (AS) vs. plot size for a two-year old stand composed of a 96×64 grid of 20×20 cm cells on an open-cast coal mine in Visonta, northern Hungary. The units are in bits. The values of the Associatum are magnified five-fold.
Since AS responds not only to disorder which arises from random assortment, but also to emergent properties which issue from (spatial) organization, it is more in line with our intuitive notion of statistical complexity than any surrogate diversity measure previously used in ecology.

This fact has implication in quantitative evaluations of changes in community composition, restoration and recovery in the sense that changes in species abundance and occurrence be considered in terms of their potential consequences not just on Shannon diversity, but also on 'structural complexity' (Anand and Orłóci 1996). We believe that effects on structural complexity are potentially as important as effects on diversity.

Finally, it should be noted that, although in ideal plant communities of infinite spatial extent, the Associatum is an unbiased measure of species spatial dependence; actual communities have limits. Due to this finite-size effect, additional artificial constraints on the local coexistence of species are generally introduced. For example, in actual communities "accidental" rare species combinations are usually overrepresented, such that even in randomly simulated species assemblages some significant species associations will inevitably emerge. Therefore, as suggested by Podani et al. (1993), the deviation of an actual value from the Associatum calculated for randomly simulated communities should be preferably used.

ACKNOWLEDGMENTS

M.A. thanks the Natural Science and Engineering Research Council (NSERC) of Canada, the Canada Research Chairs Program and the Premier's Research Excellence Award (Ontario) for funding. We thank S. Bartha for use of data, J. Podani for a helpful review and L. Orłóci for constructive comments.

REFERENCES


