

A SHORT OVERVIEW OF MEASURING DIVERSITY

Tóthmérész, B.

Ecological Institute, University of Debrecen, Debrecen, POB. 71, H-4010, Hungary
Tel. (+36) 52 316 666 / 2616; Fax. (+36) 52 431 148; E-mail: tothmerb@terra.ecol.klte.hu

Summary

In ecology, there is a considerable interest in measuring diversity. A short review of the recent developments are presented with a special emphasis both on the evolution of the techniques and the abstract presentation. Diversity is introduced as an average rarity statistics. The importance of scale-dependent diversity characterization through one-parametric diversity index families is stressed. Finally, a fairly general definition of diversity is provided, which is based on the idea of ordering through (weak) majorization and (weak) Schur-concavity. Its biological relevance as density independent and density dependent representation is pointed out.

Keywords: Diversity, Entropy, Scale-dependence, Majorization, Schur-concave function, Ecology.

Introduction and historical notes

Diversity is certainly one of the most important concept in ecology. The literature discussing the idea of diversity is quite voluminous. I mention just two recently published monographs: Huston (1994), Rosenzweig (1995). Species richness and diversity are also important in conservation management; they are frequently used as indicators of the well-being of ecological system (Magurran 1988). Diversity is also widely used in environmental monitoring (Washington 1984).

The problem of *index choice* is well known in the classical diversity literature. Peet (1974) discusses the need for a theory of index response to help the choice of diversity indices:

one may wish the index to be sensitive to the composition of the dominant species but relatively indifferent to that of the rare ones, etc. Unfortunately, the solution is not as well known as the problem. It was proposed by Patil and Taillie (1979) and it is based on the use of one-parametric diversity index families, where the diversity of a community is characterized by a (scale-dependent) diversity profile instead of a numerical value. The first of these techniques, a *generalized entropy*, was published by Rényi (1961). As a special case, Rényi diversity includes the number of species, Shannon diversity, quadratic diversity, and Berger-Parker diversity. Nowadays, there is a quite large family of the methods, which can be used for scale-dependent diversity characterization; a review of them is provided by Tóthmérész (1995, 1998).

Nowadays, the most popular diversity indices are based on the frequency distribution of individuals. Numerous diversity indices have been proposed. The two most frequently used are the *Shannon* index and the *Simpson* or *quadratic* index (see e.g. Pielou 1975). Patil and Taillie (1979) stressed the view that community diversity can be defined to be the *average species rarity*. Depending on the rarity functions, great many diversity functions can be defined.

The number of species and elementary diversity indices

A community A may be described by the *abundance vector of the community*: $\mathbf{n}=(n_1, n_2, \dots, n_S)$, where S is the number of species that are present, and n_i is the abundance of the i -th species of the community. When we sum up all the individuals, we receive the total number of individuals of the community, which is denoted by N :

$$N = n_1 + n_2 + \dots + n_i + \dots + n_S = \sum_{i=1}^S n_i.$$

For our purposes it is frequently enough to know the relative abundances of species: $\mathbf{p}=(p_1, p_2, \dots, p_S)$, where \mathbf{p} is the relative abundance vector of the community, and $p_i = n_i/N$.

Frequently we would like to know which one is the most frequent species, or the second most frequent, etc. In this case, when the species are arranged in descending order, we use the following notation:

$$\mathbf{p}^\downarrow = (p_{[1]}, p_{[2]}, \dots, p_{[1]}, \dots, p_{[S]}) ,$$

where $p_{[1]}$ is the relative frequency of the most frequent species, $p_{[2]}$ is the relative frequency of

the second most frequent, ..., and $p_{[S]}$ is the relative frequency of the rarest species. The sign "[]" in the subscript means that elements of the vector is arranged in descending order. Therefore:

$$p_{[1]} \geq p_{[2]} \geq \dots \geq p_{[i]} \geq \dots \geq p_{[S]}.$$

Figure 1 shows the evolutionary tree of the methods to characterize diversity; this is used as a guide during the paper. The *number of species* is certainly the oldest measure of diversity. However, it depends on the number of individuals in the sample and/or the area to be sampled. This is the basic motivation of the standardizations: the number of species is divided by the number of individuals or by the area to be sampled (e.g. plot size). The number of species does not increase linearly by the number of individuals. Therefore, the standardization is more correct, when the number of species is divided by the logarithm of the number of individuals, because the number of species increases linearly or almost linearly with the logarithm of the number of individuals. A standardization of the number of species by the number of individuals was proposed by Margalef (1958): S/N . Gleason (1922) proposed $S/\log N$ and Menhinick (1964) suggested S/\sqrt{N} as a measure of species richness. These simple, richness-type measures of diversity are frequently useful in many situation, although they may be criticized. These simple indices do not take into account the abundance-dominance structure of the communities. This problem is overcome by the *traditional diversity statistics*, like *Shannon diversity* or *quadratic diversity*. These methods utilize the information about the *relative frequencies* of the species of the communities.

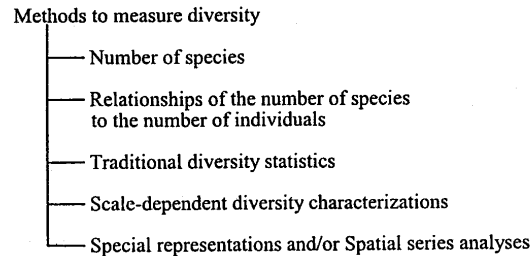


Figure 1. Tree diagram of the diversity measuring methods.

A formal definition of diversity and classical diversity statistics

The "average rarity interpretation" proposed by Patil and Taillie (1979) is especially useful to define the notion of diversity. The classical diversity statistics are presented here from this viewpoint.

Diversity can be defined to be the *average species rarity* of the individuals of a community. Denote the rarity of species i of a studied community by $R(i; \mathbf{p})$; i.e. a numerical measure of rarity is to be associated with each species. Therefore, the diversity measure of a community is defined as its average rarity:

$$D(\mathbf{p}) = \sum_{i=1}^S p_i R(i; \mathbf{p}).$$

A rarity function assigns greater rarity to physically rarer species:

$$R(i; \mathbf{p}) \leq R(j; \mathbf{p}) \text{ when } p_i \geq p_j.$$

Depending on the rarity functions, great many diversity functions can be defined. There are two types of rarity functions: *Rank-type rarity measures* and *dichotomous-type rarity measures*. In the case of ranking the rarity of species depends only on its descending rank. For dichotomy, the rarity of the species i depends only on the numerical value of p_i . The above mentioned criteria for rarity functions is sufficient for rank-type rarities; for dichotomous rarities this is not the case, although we do not present more strict criteria because it would involve more mathematics. Generally, it is related to some *monotonicity* requirement or more biologically to the *forward transfer of abundance* (see Patil and Taillie 1982).

Rarity functions can be created on the basis of biological and/or statistical ideas. The relative frequency, p_i , is high for a frequent species. Therefore, it is reasonable to characterize the rarity of a species by $R(i; \mathbf{p}) := (1 - p_i)$. Then we receive the *quadratic diversity*, DQ :

$$DQ = \sum_{i=1}^S p_i (1 - p_i) = 1 - \sum_{i=1}^S p_i^2.$$

When the rarity of a species is characterized by $R(i; \mathbf{p}) := -\log p_i$, the *Shannon diversity* is received:

$$HS = \sum_{i=1}^S p_i (-\log p_i) = -\sum_{i=1}^S p_i \log p_i.$$

It is natural to standardize $(1 - p_i)$ by p_i . Then the rarity is characterized by $(1 - p_i)/p_i$, and the diversity function is

$$DS_n = \sum_{i=1}^S p_i \frac{1-p_i}{p_i} = S-1.$$

This is exactly the *species richness* of the community. I prefer to say, that S is the *number of species*, while $S-1$ is the *species richness*. This is motivated by the "axiomatic" treatment of diversity concept, which is not discussed here; see e.g. Aczél and Daróczy's monography (1975) of the axiomatic definition of entropy. There are many other ways to choose rarity functions. Further diversity functions produced that way, is not discussed here.

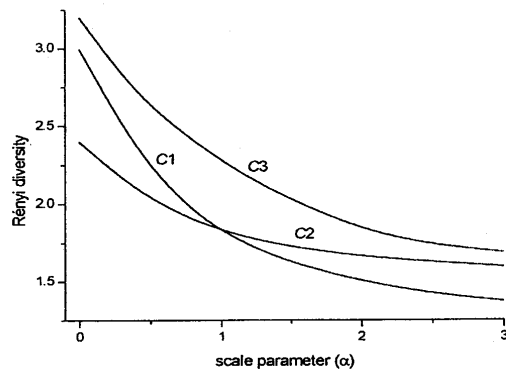


Figure 2. Characterization of scale-dependent diversity by Rényi diversity profiles for the hypothetical C1, C2 and C3 communities.

Characterization of scale-dependent diversity

Figure 2 shows the diversity profiles of three communities by the Rényi diversity index family. I would like to stress three important special scale parameter values. (i) When the value of the scale parameter is 0, then the value of the Rényi diversity is the logarithm of the number of species of the community; i.e. $HR(0)=\log S$. In this case the method is extremely sensitive to the contribution of the rare species to the diversity of the community. Here, the C1 community is more diverse than the C2.

(ii) When the value of the scale parameter is 1, then the Rényi diversity is identical with the Shannon diversity. In this case the diversity is sensitive to the rare species, although the sensitivity is not as high as it was when the value of the scale parameter was 0. Now, the diversities of the C1 and C2 communities are identical.

(iii) When the value of the scale parameter is 2, then the Rényi diversity is related to the quadratic diversity; it is $HR(2)$. In this case the method is more sensitive to the frequent species than to the rare ones; and now, community C2 is more diverse than the community C1.

Therefore, community C1 is more diverse for the rare species, while community C2 is more diverse for the frequent ones; therefore, in the provided example the communities cannot be ordered by diversity. Community C3 is more diverse than both the C1 and the C2 on the whole range of the scale parameter.

This method can be used in a graphical form to visualize the diversity relations of communities as it is demonstrated by Figure 2. When we are using a one-parametric family $\{D(\alpha): \alpha \text{ real}\}$ of diversity indices, then the family may be portrayed graphically by plotting diversities $D(\alpha)$ against the (scale) parameter α . This curve, the graph of the $\{D(\alpha): \alpha \text{ real}\}$ family, frequently mentioned as the *diversity profile of the community*. Basically, α serves as a *scale parameter*, and members of the $D(\alpha)$ family have varying sensitivities to the rare and abundant species as α changes. This curve, the graph of the $\{D_\alpha: \alpha \text{ real}\}$ family, frequently described as the *diversity profile of the community* (Patil and Taillie 1979, 1982). In essence, α serves as a *scale parameter* and members of the D_α family have varying sensitivities to the rare and abundant species as α changes.

Using diversity profiles we can define the *diversity ordering of communities* in the following way: Community A is *more diverse* than community B (written $A > B$) when the diversity profile of A is above or equal to the diversity profile of B on the whole range of the scale parameter. It can be shown that diversity ordering is a partial order so that if $A > B$ and $B > C$ then $A > C$. However, it is not true that for every communities A , and B , either $A > B$ or $B > A$; i.e. curves of two diversity profiles may intersect. This situation may reflect important ecological processes which can be interpreted clearly.

It is evidently a *partial order*, because for the diversity profiles of A , B , and C communities (i) $A < A$, (ii) $A < B$ and $B < A$ imply $A = B$, and (iii) $A < B$ and $B < C$ imply $A < C$. However, it is not true that for every A , B , either $A > B$ or $B > A$; i.e. curves of two diversity profiles may intersect. In this case the two communities are not comparable; this means that we can find two diversity indices which order the communities differently. Usually, this situation reflects important ecological processes which can be interpreted clearly.

Rényi (1961) has extended the concept of Shannon entropy by defining the *entropy of order α* or *Rényi diversity* ($\alpha \geq 0, \alpha \neq 1$):

$$HR(\alpha) = \frac{1}{1-\alpha} \left(\log \sum_{i=1}^S p_i^\alpha \right).$$

It was the first published family of diversity indices. In the original definition the base number of the logarithm was 2; in ecological applications natural logarithm is the most frequently used. It is important to know some special cases of diversity index families to interpret the result of diversity orderings. For the Rényi diversity ordering the following relations are valid.

$HR(0)$ = logarithm of the total number of species;

$HR(\alpha \rightarrow 1)$ = Shannon diversity;

$HR(2)$ = related to the quadratic or Simpson diversity;

$HR(\alpha \rightarrow +\infty)$ = logarithm of the reciprocal of the relative abundance of the commonest species. This is the logarithm of the reciprocal of Berger-Parker diversity (Berger and Parker 1970).

Nowadays a large family of one-parametric diversities is known. Their family-tree is presented by Figure 3; Tóthmérész 1993. Rényi diversity is a typical member of the *generalized entropies*. *RTS* diversity (*Right-Tail-Sum* diversity) plays also a central role in scale-dependent diversity characterizations; Patil and Taillie 1979, Solomon 1979. *RTS* diversity is a typical member of the *cumulative relative abundance plots*, and it is defined as follows:

$$RTS(i) = p_{[i+1]} + \dots + p_{[S]},$$

where $p_{[1]}, \dots, p_{[S]}$ are the relative abundances of the species of a community arranged in *descending order*. Tóthmérész (1995) has demonstrated that Rényi diversity can be used very effectively in ecological studies. *RTS* diversity is more important from theoretical than from practical point of view. In the form of a *logarithmic dominance plot* is advised to use it as a diversity profile (Tóthmérész 1993, 1995).

Diversity index families

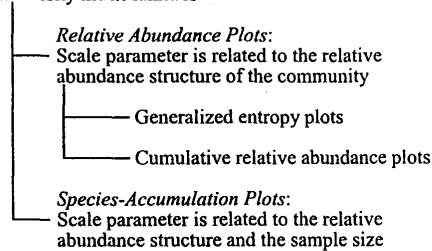


Figure 3. Tree diagram of one-parametric diversity index families.

There is a long tradition of *species-area* and *species-individual curves* in biology (Engen 1978, Fisher et al 1943). Generally, I prefer to mention them as *species-accumulation plots*. These curves also can be used for scale-dependent diversity characterization. Therefore, these methods scale along the abundance-dominance. One of them is the so-called *ES(m)* diversity:

$$ES(m) = S - \sum_{i=1}^S (1 - p_i)^m.$$

It produces the expected number of species present when m individuals are drawn at random from the population. Conceptually m is an integer, but real values make mathematical sense. *ES(m)* also used to be mentioned for a particular value of m as *expected species-individual diversity*. The minimum variance unbiased estimator for *ES(m)* was devised by Smith and Grassle (1977):

$$S - \sum_{i=1}^S \binom{N - n_i}{m} / \binom{N}{m},$$

and

$$\binom{N}{m} = \frac{N(N-1) \cdot \dots \cdot (N-m+1)}{1 \cdot 2 \cdot \dots \cdot m}.$$

The expected number of individuals on an area is proportional to the size of the area. Therefore, we can calculate the *expected species-area curve* using the following relationship

$$m = N \frac{\text{plot size}}{\text{total area}},$$

where N is the total number of individuals on the area.

Species-accumulation plots can be used producing density dependent and density independent representations of the diversity profiles (Tóthmérész 1998). These methods are also related to spatial series analysis through pattern dependent representations (Tóthmérész 1994).

The effective number of species

Diversity statistics introduced in the previous section are clearly showing the effect of abundance-dominance structure of the community. At first sight, however, there is not much biological meaning of these numerical figures. We would like to have a diversity characterization which has a straightforward biological meaning. This is provided by the *effective number of species*. It is defined as the number of species, having the same number of individuals for each species, produced the same diversity as the studied one; i.e. it is the number of species that would be

found in a hypothetical community of perfect evenness having the same diversity as the community whose diversity is to be measured.

For the Shannon diversity the effective number of species is defined as

$$SHS = \exp \{HS\},$$

where \exp is the exponential function. Shannon diversity receives its maximum, when all the species are present with the same number of individuals. (This simple statement should be proved mathematically, but we omit it.) In this case the diversity is

$$\max \{HS\} = \log S.$$

Therefore, the effective number of species is exactly S for that community; it is less than S for any other communities.

Quadratic diversity looks slightly different from Shannon diversity. It is based on $\sum p_i^2$, and then a "converse" of it is created by subtracting it from one; i.e.

$$DQ = 1 - \sum_{i=1}^S p_i^2$$

is devised. Actually, $\sum p_i^2$ is a measure of *concentration* and, in this particular case, diversity is defined as its "opposite". It also can be created in the following ways:

$$SDQ = 1 / \sum_{i=1}^S p_i^2,$$

$$HR(2) = -\log \sum_{i=1}^S p_i^2 = \log \frac{1}{\sum p_i^2}.$$

DQ and SDQ are trivial. $HR(2)$ is based on the properties of the logarithmic function. SDQ can be used for measuring effective number of species related to the quadratic diversity. SDQ is closely related to $HR(2)$, because $SDQ = \exp \{HR(2)\}$. SHS and SDQ are also strongly related, because they are related to the Rényi diversity index family. It is also evident from the relationships, that

$$SHS \geq SDQ,$$

and the equality is valid just in the case of perfectly even community (all the species have the same number of individuals). The interpretation of the values of the effective number of species are more straightforward. It is evident from these particular values of the Rényi diversity index family, that it can be used in the form of an effective number of species for the other values of the scale parameter α as:

$$EHR(\alpha) = \exp \{HR(\alpha)\}.$$

Further perspectives

The definition of diversity as the average rarity of the community is fairly general, which induced an important development of diversity measurements. By nowadays, even this definition looks a bit obsolete and a more general definition is inevitable based on the notion of *Schur-concavity* or the idea of *majorization* and *weak majorization* (Tóthmérész and Katona 1996). It is evident, that the Rényi diversity cannot be written in the form of average rarity. Although, $\exp \{HR(\alpha)\}$ can be represented in the form of average rarity. There are other difficulties. These are related to the idea of *density dependent representation of diversity* (Tóthmérész 1994, 1998). In this case, the possibility of the average rarity interpretation is also questionable.

It can be verified that each diversity index mentioned in the paper is a Schur-concave function. A function is said to be *Schur-concave* if it preserves the (backwards) orderings, i.e. for two abundance vectors, \mathbf{p} and \mathbf{q} , $\mathbf{p} > \mathbf{q}$ implies $D(\mathbf{p}) \leq D(\mathbf{q})$ (Hardy et al 1952, Marshall and Olkin 1979). For the relative abundance vectors $\mathbf{p} > \mathbf{q}$ means that \mathbf{p} majorize \mathbf{q} , which is true when

$$\sum_{i=1}^j p_i \geq \sum_{i=1}^j q_i \quad \text{and} \quad \sum_{i=1}^S p_i = \sum_{i=1}^S q_i$$

for each $j=1,2,\dots,S-1$. When the diversities of two communities are comparable through majorization of their relative abundance vectors, then they can be partially ordered through any of these diversity indices. Actually, all the other diversity indices used in the ecological literature fulfill this requirement. Therefore, it looks natural to use these criteria as a definition of diversity. There is an additional benefit of these definitions. The usual diversity indices provide density independent representation of diversity. These indices are calculated from the relative abundance vectors of the communities, therefore they naturally fulfill the

$$\sum_{i=1}^S p_i = \sum_{i=1}^S q_i$$

requirement of the majorization for any \mathbf{p} , \mathbf{q} relative abundance vectors, because $\sum_{i=1}^S p_i = \sum_{i=1}^S q_i = 1$. In the case of the species-area curves, when a density dependent representation is used, this criterion is valid when the compared communities have the same density. Even in this case, however, fulfills the diversity indices the requirements of weak majorization. Therefore the following, fairly general, new definition of diversity may be proposed:

Diversity is a real-valued non-negative function defined on \mathbb{R}_+^S , which preserves in

backwards the order of *majorization* or *weak majorization*. When the order of weak majorization is preserved we speak about *density dependent representation of diversity* while the other case is mentioned as *density independent*.

Let

$$P = \{x \in \mathbb{R}_+^S : x_{[1]} \geq x_{[2]} \geq \dots \geq x_{[S]}\}$$

represent the set of species abundance vectors for an S -species community, where $x_{[j]}$ is the abundance of the most abundant species, etc. We shall write that x majorize y , $x > y$, when

$$\sum_{i=1}^j x_{[i]} \leq \sum_{i=1}^j y_{[i]}$$

for each $j=1,2,\dots,S-1$. For measures of species diversity in biology, a function D may be used as a measure of diversity when D is *Schur-concave*, i.e.

$$x < y \text{ then } D(x) \leq D(y),$$

or *strictly Schur-concave*:

$$x < y \text{ then } D(x) < D(y),$$

but x is not a permutation of y .

It can be shown that majorization is a *partial order*. A highly useful feature of partially ordered sets is that they can be drawn. To see how, we need the idea of *covering*. Let $(P, <)$ be a partially ordered set and let $x, y \in P$. We say x is *covered by* y and write $x \prec y$ if $x < y$ and $x < z < y$ implies $z = x$; i.e. if there is no element z with $x < z < y$. Suppose $(P, <)$ is a finite partially ordered set. (i) To each $x \in P$ associate a point $P(x)$, depicted by a small circle centered at $P(x)$ and (ii) for each covering pair $x \prec y$ in P , take a line segment *line*(x, y) joining $P(x)$ to $P(y)$. It is easily proved by induction on P that this can be done in such a way that (i) if $x \prec y$, then $P(x)$ is "lower" than $P(y)$, and (ii) $P(z)$ does not lie on *line*(x, y) if $z \neq x$ and $z \neq y$. Such a representation is called *Hasse diagram* of $(P, <)$. This representation also may be very useful for representing the diversity changes of sophisticated ecological situations, e.g. diversity relations of successional stages.

Acknowledgements - The research was supported by Hungarian Scientific Research Fund (OTKA research grant no. T25888).

Literature

Aczél, J. & Daróczy, Z. 1975: On Measures of Information and their Characterizations, Academic Press, New York, San Francisco, London.

Berger, W.H. and Parker, F.L. 1970: Diversity of planktonic Foraminifera in deep sea sediments. Science, 168: 1345-7.

Engen, S. 1978: Stochastic Abundance Models. Chapman and Hall, London.

Fisher, R.A., A.S. Corbet, and C.B. Williams 1943: The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol., 12: 42-58.

Gleason, H.A. 1922: On the relation between species and area. Ecology, 3: 156-162.

Huston, M. A. 1994: Biological Diversity. Cambridge University Press, Cambridge.

Hardy, G.H., Littlewood, J. E., and Pólya, G. 1952: Inequalities. 2nd ed. Cambridge Univ. Press, London and New York.

Tóthmérész, B. and Katona, É. 1996: Methods to characterize and display diversity relations. 18th International Biometric Conference.

Magurran, A.E. 1988. Ecological Diversity and Its Measurement. Croom Helm, London.

Margalef, R. 1958: Information theory in ecology. General Systems 3, 36-71.

Marshall, A. W. and Olkin, I. 1979: Inequalities: Theory of Majorization and Its Applications. Academic Press, New York.

Menhinick, E.F. 1964: A comparison of some species-individuals diversity indices applied to samples of field insects. Ecology, 45: 859-861.

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*, pp. 3-27. International Cooperative Publishing House, Fairland, Maryland.

Patil, G.P. and Taillie, C. 1982: Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77: 548-567.

Peet, R.K. 1974: The measurement of species diversity. *Annual Review of Ecology and Systematics*, 5: 285-307.

Perry, J.N. 1986: Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology*, 79: 1149-1155.

Pieou, E.C. 1975: *Ecological Diversity*. Wiley, New York.

Rényi, A. 1961: On measure of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability (Vol. I)* (Ed. by J. Neyman), pp. 547-561. University of California Press, Berkeley.

Routledge, R.D. 1977: On Whittaker's components of diversity. *Ecology*, 58: 1120-1127.

Rosenzweig, M. L. 1995: *Species diversity in space and time*. Cambridge Univ. Press, Cambridge.

Smith, W. and Grassle, F.J. 1977: Sampling properties of a family of diversity measures. *Biometrics*, 33: 283-292.

Solomon, D.L. 1979: A comparative approach to species diversity. *Ecological Diversity in Theory and Practice* (Ed. by J.F. Grassle, G.P. Patil, W. Smith, & C. Taillie), pp. 29-35. International Cooperative Publishing House, Fairland, Maryland.

Tóthmérész B. 1993: DivOrd 1.50: A Program for Diversity Ordering. *Tiscia*, 27: 33-44.

Tóthmérész, B. 1994: 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. 1998: On the characterization of scale-dependent diversity. *Abstracta Botanica*, 22: 149-156.

Washington, H.G. 1984: Diversity, biotic and similarity indices. A review with special reference to aquatic ecosystems. *Water Research*, 18: 653-694.