

## *Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy*

C. J. Keylock, *Earth and Biosphere Institute and School of Geography, Univ. of Leeds, Woodhouse Lane, Leeds, UK* (c.j.keylock@leeds.ac.uk).

Many indices for measuring species diversity have been proposed. In this article, a link is noted between a common family of diversity indices and non-additive statistical mechanics. This makes the Shannon index and the Simpson diversity (or Gini coefficient) special cases of a more general index. The general index includes a parameter  $q$  that can be interpreted from a statistical mechanics perspective for systems with an underlying (multi)fractal structure. A  $q$ -generalised version of the Zipf–Mandelbrot distribution sometimes used to characterise rank–abundance relationships may be obtained by maximising this entropy.

The study of the causes and effects of differences in biodiversity between communities requires suitable measures of species richness and diversity. The former is the number of a species in a community, while the latter is a function of the relative frequency of different species. There has been a great deal of debate in the literature over the most appropriate measures for diversity. Lande (1996) listed a set of desirable properties of an index (*italics in original*):

A measure of species diversity should ideally be *nonparametric* and *statistically accurate*. It should be applicable to any community independent of species abundance distribution, and should have small bias and sampling variance in samples of moderate size. An important property for a diversity measure, first discussed by Lewontin (1972) for genetic diversity, is *strict concavity*. This means that the total diversity in a pooled set of communities equals or exceeds the average diversity within communities, with equality only for identical communities.

Thus, the chosen index must be formulated carefully to account for a variety of theoretical constraints.

### **The Hill family of indices**

Routledge (1979) suggested that of the various indices proposed for measuring species diversity, the admissible

forms are those that belong to the Hill family of indices (Hill 1973):

$$I_k = \left[ \sum_{i=1}^n p_i^k \right]^{1/(1-k)} \quad (1)$$

where  $k$  is a coefficient and  $p_i$  is the relative abundance of species  $i$  in a sample of  $n$  species. Important special cases of the Hill family include appropriate transformations of the Shannon–Wiener index (Pielou 1966, Whittaker 1972, Ludwig and Reynolds 1988, Spellerberg and Fedor 2003) and the Simpson concentration (Simpson 1949, May 1975). The former is based upon the Boltzmann–Gibbs–Shannon entropic form:

$$S = -k_B \sum_{i=1}^n p_i \ln p_i \quad (2)$$

where  $k_B$  is the Boltzmann constant, which is taken to be unity in ecological applications. An evenness index can be constructed from Eq. 2 by dividing by  $\ln n$ . The entropy has a maximum for a uniform distribution, where it can be made equal to  $n$  by calculating  $I_1 = \exp S$ . The Simpson concentration is given by:

$$\Psi = \sum_{i=1}^n p_i^2 \quad (3)$$

and  $1 - \Psi$  is the Gini coefficient or Simpson diversity (Lande 1996). A comparison of Eq. 1 and 3 reveals that  $I_2 = 1/\Psi$ . Hence, the Hill family includes the commonly adopted Shannon–Wiener index and Simpson concentration. However, it has been shown (Lande 1996) that although  $I_1$  and  $1 - \Psi$  are concave everywhere,  $1/\Psi$  ( $I_2$ ) is not. Hence, the concavity constraint invalidates the use of the Hill family of indices as a general set of measures of diversity. The fact that  $1 - \Psi$  can be interpreted as a variance and the probability that two randomly chosen

individuals from a particular community are different species (Lande 1996) makes  $1 - \Psi$  a particularly useful measure. Lande et al. (2000) have shown that the Simpson diversity is particularly useful for rapidly assessing areas for conservation because of its rapid convergence towards the limit diversity value for small sample sizes.

### A family of indices based on a generalized entropy

It does not appear to have been noted that, based on the results presented above, a preferable family of indices to the general form given by Eq. 1 can be written as:

$$S_q = \frac{1 - \sum_{i=1}^n p_i^q}{q - 1} \quad (4)$$

It is clear that  $S_2 = 1 - \Psi$  is a special case of Eq. 4 for a value for the exponent  $q$  of 2. However, Eq. 2 is also a limiting case for Eq. 4 as  $q$  tends to 1. The general measure given by Eq. 4 is concave for  $q > 0$  and has a maximum under conditions of equiprobability of

$$S_q(\max) = \frac{n^{1-q} - 1}{1 - q} \quad (5)$$

Recently, in statistical physics there has been a concerted research effort to explore the properties of Eq. 4, which is a generalization of the entropic form given by Eq. 2. These works have shown that Eq. 4 leads to a statistical mechanics that satisfies many of the properties of the standard theory (Tsallis 1988, Curado and Tsallis 1991, Plastino and Plastino 1993, Tsallis et al. 1998). In addition, it has been applied to a range of phenomena including the Bak–Sneppen model of biological evolution (Tamarit et al. 1998), and a model for the defence of territory by birds (Papa and Tsallis 1998).

The standard statistical mechanics, due to Boltzmann and Gibbs, for which Eq. 2 is the appropriate entropic form has restricted applicability. For systems that involve interactions that take place over long distances, which have long memories of perturbations or, which have an underpinning structure that is fractal or multi-fractal in nature, another form of statistical mechanics is required. Tsallis (1988) proposed Eq. 4 as an entropy that could be used to analyse such systems. In the ecological literature, several authors have suggested that the distribution and abundance of species may be self-similar or fractal (Harte et al. 1999, Li 2000, Brown et al. 2002), while vegetation patterns and the dynamics of rain forests have been considered as multifractals (Scheuring and Riedi

1994, Solé et al. 1994). Furthermore, complex interactions between species and individuals at a range of scales and the persistence of species in a region support the view that ecological systems contain significant memory (Okland et al. 2003). Hence, the Tsallis entropy (Eq. 4) would appear to be at least as relevant to characterising diversity as the Boltzmann–Gibbs–Shannon entropy (Eq. 2).

The exponent  $q$  that underpins the Tsallis entropy is linked to the underlying dynamics and measures the amount of nonadditivity in the system. That is, for two independent systems  $G$  and  $H$ , the joint probability  $p_{ij}(G+H) = p_i(G)p_j(H)$ . However, the additivity of the entropy is related to the value for  $q$ :

$$\frac{S_q(G+H)}{k_B} = \frac{S_q(G)}{k_B} + \frac{S_q(H)}{k_B} + (1-q) \frac{S_q(G)}{k_B} \frac{S_q(H)}{k_B} \quad (6)$$

Thus, if  $q=1$  the entropies are additive, while for  $q > 1$  they are subadditive and for  $q < 1$  they are super additive. When  $q > 1$  the difference between high and low probability states is enhanced, while it is reduced for  $q < 1$ .

Routledge (1979) suggested that the Simpson Concentration and Shannon index were the only members of the Hill family that were worthy of consideration, with the latter retained due to its links to information theory and entropy. The study by Lande (1996) has shown that Simpson diversity is preferable to both the Simpson concentration and Shannon index. From Eq. 4 and the associated literature on non-additive statistical mechanics it follows that both the Simpson diversity and the Shannon–Wiener index can be interpreted from an entropic perspective. Hence, the reasons Routledge gave for retaining the Shannon–Wiener measure apply just as readily to Simpson diversity, removing this justification for the use of the former.

### Sampling properties of the generalized index

An important dataset that has recently been used in several studies concerned with measuring species abundance is the Barro Colorado Island tree dataset (Condit et al. 2002). In particular, this dataset has been used to test the unified neutral theory of biodiversity (Hubbell 2001, McGill 2003, Volkov et al. 2003). This dataset consisting of 21 457 individuals and 225 species is plotted in Fig. 1 using the method of Preston (1948). Samples of between 100 and 10 000 individuals were allocated to a species based on the relative probability of occurrence of the species shown in Fig. 1. Diversity indices for different values for  $q$  were then calculated using Eq. 4 and normalised by Eq. 5, except in the case of  $q=1$ , where Eq. 2 was used and normalisation was by  $\ln n$ . Note that  $n$  is the

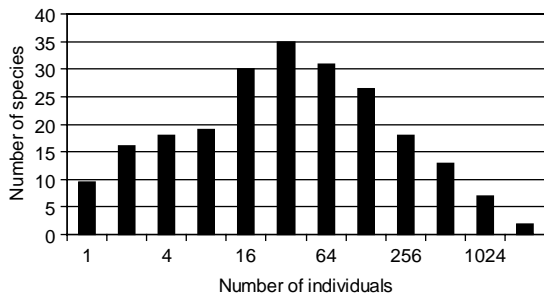


Fig. 1. The tree species abundances at Barro Colorado Island, Panama (Condit et al. 2002). The bars show the observed number of species grouped by the number of times they were observed. The log2 abundance categories used follow Preston (1948).

number of species that emerges from a particular set of random trials and has an upper bound of 225, which is approached as the number of sampled individuals increases. Twenty replicate trials were used for each sample size. The mean square error between the sample diversities and the observed value as a function of sample size and  $q$  is given in Fig. 2. Improved convergence is clearly seen for higher values of  $q$  in agreement with related analysis in Lande (1996) and the fact that indices with a higher value for  $q$  more clearly distinguish between high and low probability states. However, setting too high a value for  $q$  reduces the sensitivity of the index, decreasing the potential to distinguish between real differences in species abundance structure between communities. For example, when sampling 100 individuals from the dataset in Fig. 1 with 100 replicate trials, there was less than 2.0 standard deviations between the mean index at  $q=3.0$  and the mean for 100 replicate trials from a uniform distribution of 225 species. In contrast, there was a difference of 2.8 standard deviations at  $q=2.0$  and when sampling 1000 individuals with 100 replicate trials, the distance between cases increased to 6.5 and 12.6 standard deviations, respectively.

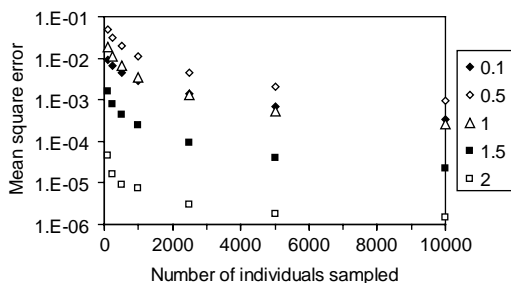


Fig. 2. The mean square error between the standardized diversity indices for the data in Fig. 1 and twenty random samples from this dataset as a function of the number of individuals sampled and the entropic index  $q$ . Five values for  $q$  are shown from 0.1 to 2.0.

## The Zipf–Mandelbrot distribution

It may be noted that several ecological studies have considered the Zipf–Mandelbrot distribution to be an appropriate form for the rank-abundance relationship (Frontier 1985, Aoki 1995, Wilson et al. 1996, Mouillot and Leprêtre 1999) although it is perhaps more usual at present to consider the zero-sum multinomial distribution, with the log-normal as a suitable null hypothesis (McGill 2003, Volkov et al. 2003). The Zipf–Mandelbrot distribution is given by:

$$p_r = N_1 (r + \beta)^{-\gamma} \quad (7)$$

where  $p_r$  is the relative frequency of a particular species of rank  $r$  when ranked in decreasing order of relative abundance,  $\beta$  and  $\gamma$  are model parameters and  $N_1$  is a normalising constant:

$$N_1 = \frac{1}{\sum_{i=1}^n (i + \beta)^{-\gamma}} \quad (8)$$

where  $n$  is the total number of species. It is possible to express Eq. 7 and 8 in a form that is compatible with the  $q$  framework (Denisov 1997, Montemurro 2001):

$$p_r = \frac{[1 - (1 - q)\lambda' r]^{1/(1-q)}}{\sum_{i=1}^n [1 - (1 - q)\lambda' i]^{1/(1-q)}} \quad (9)$$

where  $q = (1/\gamma) + 1$  and  $\lambda' = \gamma/\beta$ . Because values for  $\gamma$  are always greater than zero (Mouillot and Leprêtre 1999), the concavity of Eq. 4 is enforced. The distribution given by Eq. 9 emerges naturally from an entropy maximisation of Eq. 4. When working with the standard Shannon entropy (Eq. 2) and imposing appropriate constraints on the energy  $\varepsilon$  and probability the canonical distribution (Eq. 10) is obtained.

$$p_i = \frac{e^{-\lambda \varepsilon_i}}{\sum_{i=1}^{\infty} e^{-\lambda \varepsilon_i}} \quad (10)$$

Unfortunately, if the Shannon entropy is replaced by the generalized Tsallis entropy (Eq. 4) and entropy is maximised using the same constraints then there are mathematical difficulties with the resulting form of the distribution. A modification to the energy constraint was introduced by Tsallis et al. (1998) to resolve this, leading to the probability distribution given by Eq. 11:

$$p_i = \frac{[1 - (1 - q)\lambda' \varepsilon_i]^{1/(1-q)}}{\sum_{i=1}^{\infty} [1 - (1 - q)\lambda' i]^{1/(1-q)}} \quad (11)$$

Following Denisov (1997) and equating rank with energy and thus, normalising between 1 and  $n$  rather than 1 and

$\infty$ , Eq. 11 is equivalent to Eq. 9. The connection to Eq. 10 is perhaps clearer to see with the definition of a  $q$  exponential function:

$$e_q^x \equiv [1 + (1 - q)x]^{1/(1-q)} \quad (12)$$

which means that Eq. 11 can be re-written as:

$$p_i = \frac{e_q^{-\lambda_i \varepsilon_i}}{\sum_{i=1}^{\infty} e_q^{-\lambda_i \varepsilon_i}} \quad (13)$$

## Conclusion

Previous work has suggested that measures based on the Simpson index have a number of desirable properties (Lande 1996). Routledge (1979) stated that consideration should also be taken of the Shannon index due to its central role in information theory and statistical physics. The generalized entropy introduced by Tsallis (1988) and presented here suggests that these indices are special cases of Eq. 4, and with  $q > 0$ , can both be linked to a more general information theory and statistical physics. Consequently, from the perspective of this generalized measure of entropy, Eq. 2 is no more justifiable than  $1 - \Psi$  as a diversity measure because both can be linked to a generalised formulation of entropy. This means that the improved sampling properties of  $1 - \Psi$  (Lande 1996, Lande et al. 2000) make it the preferable index. As well as providing a useful practical tool for field ecology, the non-additive form for entropy has implications for the macro-ecological study of the energetics of communities. This may have a more general utility with respect to other fractal or multifractal properties of ecological systems than has been demonstrated in this study.

*Acknowledgements* – I am grateful to Russell Lande, Oliver Phillips and Drew Purves for critical comments on various drafts of this manuscript and some extremely useful suggestions. Some of this work was undertaken while the author was at the Nagaoka Institute of Snow and Ice Studies, funded by JSPS short term fellowship PE 04511.

## References

Aoki, I. 1995. Diversity and rank–abundance relationship concerning biotic compartments. – *Ecol. Model.* 82: 21–26.  
 Brown, J. H., Gupta, V. J., Li, B. et al. 2002. The fractal nature of nature: power-laws, ecological complexity and biodiversity. – *Philos. Trans. R. Soc. B.* 357: 619–626.  
 Condit, R., Pitman, N., Leigh Jr, E. G. et al. 2002. Beta diversity in tropical forest trees. – *Science* 295: 666–669.  
 Curado, E. M. F. and Tsallis, C. 1991. Generalized statistical mechanics-connection with thermodynamics. – *J. Phys. A-Math. Gen.* 24: L69–L72.  
 Denisov, S. 1997. Fractal binary sequences: Tsallis thermodynamics and the Zipf law. – *Phys. Lett. A* 235: 447–451.  
 Frontier, S. 1985. Diversity and structure in aquatic ecosystems. – *Mar. Biol. Annu. Rev.* 23: 253–312.

Harte, J., Kinzig, A. P. and Green, J. 1999. Self-similarity in the distribution and abundance of species. – *Science* 284: 334–336.  
 Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – *Ecology* 54: 427–432.  
 Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.  
 Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – *Oikos* 76: 5–13.  
 Lande, R., DeVries, P. J. and Walla, T. R. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. – *Oikos* 89: 601–605.  
 Lewontin, R. C. 1972. The apportionment of human diversity. – *Evol. Biol.* 6: 381–398.  
 Li, B.-L. 2000. Fractal geometry applications in description and analysis of patch patterns and patch dynamics. – *Ecol. Model.* 132: 33–50.  
 Ludwig, J. A. and Reynolds, J. F. 1988. Statistical ecology. A primer on methods and computing. – Wiley.  
 May, R. M. 1975. Patterns of species abundance and diversity. – In: Cody, M. L. and Diamond, J. M. (eds), *Ecology and evolution of communities*. Harvard Univ. Press, pp. 81–120.  
 McGill, B. J. 2003. A test of the unified neutral theory of biodiversity. – *Nature* 422: 881–885.  
 Montemurro, M. A. 2001. Beyond the Zipf–Mandelbrot law in quantitative linguistics. – *Physica A* 300: 567–578.  
 Mouillot, D. and Leprêtre, A. 1999. A comparison of species diversity estimators. – *Res. Popul. Ecol.* 41: 203–215.  
 Okland, R. H., Rydgren, K. and Okland, T. 2003. Plant species composition of boreal spruce swamp forests: closed doors and windows of opportunity. – *Ecology* 84: 1909–1919.  
 Papa, A. R. R. and Tsallis, C. 1998. Imitation games: power-law sensitivity to initial conditions and nonextensivity. – *Phys. Rev. E* 57: 3923–3927.  
 Pielou, E. C. 1966. Shannon's formula as a measure of specific diversity: its use and misuse. – *Am. Nat.* 100: 463–465.  
 Plastino, A. R. and Plastino, A. 1993. Tsallis entropy, ehrenfest theorem and information-theory. – *Phys. Lett. A* 177: 177–179.  
 Preston, F. W. 1948. The commonness and rarity of species. – *Ecology* 29: 254–283.  
 Routledge, R. D. 1979. Diversity indices: which ones are admissible? – *J. Theor. Biol.* 76: 503–515.  
 Scheuring, I. and Riedi, R. H. 1994. Application of multifractals to the analysis of vegetation pattern. – *J. Veg. Sci.* 5: 489–496.  
 Simpson, E. H. 1949. Measurement of species diversity. – *Nature* 163: 688.  
 Solé, R. V., Manrubia, S. C. and Luque, B. 1994. Multifractals in rainforest ecosystems: modelling and simulation. – In: Novak, M. M. (ed.), *Fractals in the natural and applied sciences*. North-Holland, pp. 397–407.  
 Spellerberg, I. F. and Fedor, P. J. 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' index. – *Global Ecol. Biogeogr.* 12: 177–179.  
 Tamarit, F. A., Cannas, S. A. and Tsallis, C. 1998. Sensitivity to initial conditions in the Bak-Sneppen model of biological evolution. – *Eur. Phys. J. B* 1: 545–548.  
 Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. – *J. Stat. Phys.* 52: 479–487.  
 Tsallis, C., Mendes, R. S. and Plastino, A. R. 1998. The role of constraints within generalized nonextensive statistics. – *Physica A* 261: 534–554.

- Volkov, I., Banavar, J. R., Hubbell, S. P. et al. 2003. Neutral theory and relative species abundance in ecology. – *Nature* 424: 1035–1037.
- Whittaker, R. H. 1972. Evolution and measurement of species diversity. – *Taxon* 21: 213–251.
- Wilson, J. B., Wells, T. C. E., Trueman, I. C. et al. 1996. Are there assembly rules for plant species abundance? An investigation in relation to soil resources and successional trends. – *J. Ecol.* 84: 527–538.