Ecological Indicators 37 (2014) 90–98

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Computer intensive methods for controlling bias in a generalized species diversity index

Davi Butturi-Gomes[a,*,1], Miguel Petrere Junior[b,c], Henrique C. Giacomini[d,2], Paulo De Marco Junior[e]

[a] Universidade Estadual Paulista "Julio de Mesquita Filho", Instituto de Biociências, Departamento de Bioestatística, Programa de Pós-Graduação em Biometria, Distrito de Rubião Junior S/N, Botucatu, SP, Brazil
[b] Centro de Ciências e Tecnologias para a Sustentabilidade (CCTS), Programa de Pós-graduação em Diversidade Biológica e Conservação (PPGDBC), Universidade Federal de São Carlos (UFSCar), Rod João Leme dos Santos, km 110, 18052-780 Sorocaba, SP, Brazil
[c] UNISANTA, Programa de Pós-Graduação em Sustentabilidade de Ecossistemas Costeiros e Marinhos, Rua Oswaldo Cruz, 277 (Boqueirão), 11045-907 Santos, SP, Brazil
[d] Department of Ecology and Evolutionary Biology, University of Toronto, Office RW 520B, 25 Harbord St., Toronto, ON, Canada M5S 3G5
[e] Universidade Federal de Goiás, Campus II, Laboratório de Ecologia Teórica e Síntese, Goiânia, GO, Brazil

ARTICLE INFO

Article history:
Received 19 March 2013
Received in revised form 8 September 2013
Accepted 7 October 2013

Keywords:
Quadrat sampling
Monte Carlo simulations
Resampling methods
Shannon–Wiener index
Simpson index
Species richness

ABSTRACT

The use of diversity indices is a common practice in studies of community ecology. Historically, the main indices were derived by Shannon and Simpson. Currently, these two indices are recognized as part of families of entropy-based indices, which generally include species richness as another particular case. This paper evaluates the statistical properties of one of these families, the Tsallis index, as dependent on four factors: (i) spatial distribution of individuals; (ii) species-abundance distributions; (iii) sampling method and (iv) the estimator. To do so, we carried out computer simulations. The maximum likelihood estimator under all scenarios produced more biased estimates than the two computationally intensive estimation methods (i.e., Jackknife and bootstrap). The Broken-Stick was the species-abundance distribution that led to lowest bias, particularly in the species richness estimation. Intermediate levels of spatial aggregation of individuals were also related to less biased estimations of diversity. The effect of quadrat size upon the bias of estimation was weak, despite the fact that such sampling method often produces a non-random sample of individuals. On the one hand, the Jackknife method was more accurate than the bootstrap, although both methods have shown poor performances for diversity indices that emphasize species richness. On the other hand, if confidence intervals are needed for individual community samples, the bootstrap is strongly recommended over the Jackknife.

1. Introduction

Species diversity is one of the most important measures of community organization and is frequently used in theoretical and applied studies. It has direct implications for decisions concerning ecosystem management and conservation, as it is widely adopted as a goal or as an indicator of ecosystem health and function (Chapin et al., 2000; Chiarucci et al., 2011; Hooper et al., 2012; Myers et al., 2000). Despite its central role in ecology, there is still controversy about what mathematical formulation should be used to represent diversity, and a large number of indices have been proposed (Chiarucci et al., 2011; Magurran, 2004; Peet, 1974). The most commonly used indices are species richness (S), the Shannon–Wiener index (H'), and the Gini–Simpson index ($1-D$). Species richness is the simplest and most intuitive measure; however, it does not account for differences in abundance and is the most sensible to sample size (Gotelli and Colwell, 2001; Lande, 1996). H' measures the amount of entropy or information in a system (Margalef, 1958). It is perhaps the most popular amongst the so called heterogeneity indices (Magurran, 2004), although criticized for not being easily interpretable, for being sensitive to sample size and completeness, or for producing counter-intuitive community ordering in some cases (Hurlbert, 1971; Lande, 1996; May, 1975, but see Marcon et al., 2012). The Gini–Simpson index gives a probability of interspecific encounters if we assume infinite population sizes (Hurlbert, 1971 presents the index version for finite populations), and several authors have recommended its use due to favorable statistical properties (Lande et al., 2000; Mouillot and Lepretre, 1999; Routledge, 1979).

* Corresponding author. Tel.: +55 19 3429 4127; fax: +55 19 3429 4346.
E-mail addresses: davi.butturi@gmail.com, davibg@usp.br (D. Butturi-Gomes), hgiacomini@gmail.com (H.C. Giacomini), pdemarco@icb.ufg.br (P.D.M. Junior).
[1] Present address: Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciências Exatas, PPG em Estatística e Experimentação Agronômica. Av. Padua Dias 11, Piracicaba, SP, Brazil, 13418-900.
[2] Tel: +1 416 978 7338; fax: +1 416 978 5878.

1470-160X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2013.10.004

Despite an apparent disparity of interpretation and mathematical formulation, these indices differ essentially on the relative emphasis they give to abundant versus rare species or, in other words, to differences in species abundance (evenness) versus species richness, which form the two major components of species diversity. So, in the last decades, several authors have sought to establish a common link between indices while recognizing their differences, deriving or adopting generalized models that include S, H' and D as part of a richness-dominance continuum (Good, 1953; Hill, 1973; Leinster and Cobbold, 2012; Patil and Taillie, 1982; Ricotta, 2005; Shamia, 2013; Tóthmérész, 1995). One particularly interesting model is the Tsallis entropy family, namely S_q, derived as a generalization of Boltzmann–Gibbs entropy (Tsallis, 1988) and brought to community ecology as measures of diversity by Keylock (2005a). It can be formulated as:

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

where W is the total number of observed states (species i.e., $W = S$), p_i is the estimated probability of state "i" (i.e., relative density of species "i"), and q is an arbitrary real value, usually non-negative, defining the relative contribution of species richness versus evenness. By varying q, S_q renders different diversity indices along the richness-dominance continuum. If $q < 1$, the index emphasizes species richness (S) by reducing relative differences between abundant and rare species and, if $q > 1$, it emphasizes dominance within the community by exacerbating such differences. More specifically, when $q = 0$, $S_q = S_0 = S - 1$; when $q \to 1$, $S_q \to S_1 = H'$; when $q = 2$, $S_q = S_2 = 1 - D$ (Keylock, 2005a; Mendes et al., 2008).

Interestingly, an equivalent equation was derived independently by Patil and Taillie (1979, 1982) from considerations on interspecific encounters between individuals, highlighting that all three indices S, H', and $1-D$ can be interpreted in terms of both entropy and encounter probabilities. In addition, when compared to other families of indices, such as the Rényi entropy (Rényi, 1961), the Tsallis family has the desirable property of concavity for the entire range of biologically meaningful indices ($q \geq 0$), which means that the diversity of a pooled group of communities will be always equal or larger than the average diversity within each community (Keylock, 2005a; Lande, 1996, but see Jost, 2006 for a different view on the problem). More recently, a number of authors have suggested the use of equivalent species numbers or Hill's numbers (i.e., the number of identically abundant species required to produce the same entropy value of the original community) as a more adequate measure of diversity instead of entropy-based indices (Chao et al., 2012; Jost, 2006, 2007; Tuomisto, 2010). Here we chose to focus on the entropy index version as it has a longer tradition in ecology. Moreover, the Tsallis index and other entropy-based families can be converted into equivalent species numbers using simple equations as presented by Jost (2006).

The desirable mathematical properties and its intuitive value make S_q widely applicable to diversity studies in a number of fields such as physics, economics, and ecology (Bentes et al., 2008; Evangelista et al., 2012; Keylock, 2005b; Mendes et al., 2008; Tóthmérész, 1995; Tsallis, 1988). However, the statistical performance of estimators of S_q was still a pending study, which we intend to provide in this paper. Assessing the statistical robustness of a diversity estimator is important for disentangling true differences in diversity from those caused by sampling artifacts, which is fundamental given that (i) all diversity surveys rely on limited samples and (ii) virtually all natural communities do not strictly follow the assumptions underlying commonly used indices. For instance, one problem concerning the maximum likelihood estimator (MLE) of Eq. (1) is the assumption of random sampling of individual organisms, which is rarely achieved by field ecologists. Instead, quadrat sampling is commonly used in community studies, especially of sessile organisms such as plants (Routledge, 1980). Since the aggregated pattern, or patchy distribution, is the most common in nature (Taylor et al., 1978), a quadrat sampling will often produce non-random samples of individuals, which can be a source of bias and lack of precision in estimating diversity indices. If individuals tend to be located next to their conspecifics, the relative density of dominant species in the community may be overestimated or underestimated, depending on the frequency with which large aggregates are included in the sample. Also, the quadrat size is usually defined by practical reasons rather than by statistical criteria, and such arbitrary choice can have a number of consequences to the estimation of diversity (He and Legendre, 2002; Heltshe and Forrester, 1983, 1985; Pielou, 1975; Zahl, 1977).

Another factor – which is perhaps the most important in determining the performance of an estimator of diversity – is the proportion of rare species. Rare species are more likely to be missed in samples of limited size, so communities composed by many of them will have their diversity underestimated. This effect is expected to be exacerbated if we use indices that emphasize the species richness component. A good way to objectively account for changes in the proportion of rare species is the use of species-abundance models to describe the ecological assemblage on which a diversity index is based. These models are reviewed in Ferreira and Petrere (2008) and here we shall consider only a few models—the log-series (Fisher et al., 1943), the truncated lognormal (Bulmer, 1974; Preston, 1948), the geometric distribution (Pielou, 1975) and the Broken-Stick model (MacArthur, 1957). From the first to the last model, the proportion of rare species in the community is decreased. Note that we used the geometric *distribution* as in Pielou (1975) and not the geometric *series* (see Motomura, 1932 for that approach), which are considerably different models regarding the proportion of rare species.

Finally, a generalized variance for the maximum likelihood estimator of Eq. (1) is not currently available because the closed forms of such variances are highly dependent on the choice of q. There are only estimators of the variance for specific values of q: Basharin (1959) and Bowman et al. (1971) calculated the variance for H' ($q \to 1$), and Simpson (1949) calculated it for D ($q = 2$). To overcome this issue, we must consider the use of resampling methods, such as the Jackknife and the non-parametric bootstrap, in order to produce confidence intervals of the estimates and to correct for bias. These methods have been most extensively used to estimate species richness (Colwell, 2013; Gotelli and Colwell, 2011), but they are equally applicable to diversity indices in general (Magurran, 2004; Manly, 2007). The first-order Jackknife is a systematic resampling method, in which the sampling units (the individual itself when using random individual sampling or each quadrat when using quadrat sampling) in a sample size of n are discarded one at a time and, for each cycle, the index of interest is re-estimated by means of a pseudo-value. The Jackknife is then the average of the n pseudo-values (Efron and Tibshirani, 1993; Manly, 2007; Miller, 1974). The non-parametric bootstrap is a stochastic resampling method, characterized by drawing, with replacement and with equal probability, a sample of the same size as the original to be used for estimating the desired index, and the procedure is repeated a large number of times (Efron, 1979; Efron and Tibshirani, 1993; Manly, 2007; Smith and Van Belle, 1984). By allowing the discard of one or a few sampling units at a time, these resampling methods are expected to mimic the process of sampling itself. The more rare species are present in the original sample, the higher is expected to be the number of species that have been missed in the first place and, accordingly, the larger will be the re-adjustment in the diversity estimates by Jackknife or bootstrap.