Practical Estimation of Diversity from Abundance Data

Eric Marcon[1,2*]

Abstract
Measuring biodiversity requires empirical techniques to effectively estimate it from real data. The well-known underestimation of the number of species applies to low orders of diversity in general. I test nine estimators including three new ones on geometric and lognormal distributions that represent realistic, hyper-diverse communities. The best two estimators allow a good estimation of diversity of orders over 0.5, even when the sampling effort is low. I provide criteria to choose the estimator and the necessary code in the R package entropart.

Keywords
biodiversity, entropy, estimation

[1]AgroParisTech, AMAP, CIRAD, CNRS, INRAE, IRD, Univ Montpellier, Montpellier, France.

[2]UMR EcoFoG, AgroParisTech, CIRAD, CNRS, INRAE, Univ Antilles, Univ Guyane, Kourou, France.

***Corresponding author**: eric.marcon@agroparistech.fr,

Contents

1. Introduction

Measuring biodiversity requires both a robust theoretical framework (Patil and Taillie, 1982) and empirical techniques to effectively estimate the theoretical variables with real data (Beck and Schwanghart, 2010). In this paper I focus on species-neutral measures of diversity based on HCDT entropy (Havrda and Charvát, 1967; Daróczy, 1970; Tsallis, 1988) that fulfill the first requirement. Entropy measures the average surprise brought by observing individuals of a community. Surprise is a decreasing function of probability dropping to 0 when probability is 1. HCDT entropy uses a parameterized surprise function that is the deformed logarithm of order q of the reciprocal of probability (Marcon et al., 2014). Traditional measures of diversity, namely the number of species as well as Shannon's and Simpson's indices, are special cases of the HCDT entropy for values of q equal to 0, 1 and 2. HCDT entropy should be transformed into Hill numbers (Hill, 1973) for better interpretation of the value of diversity as an effective number of species (Jost, 2006). Hill numbers are simply the deformed exponential of HCDT entropy (Marcon et al., 2014). Rather than focusing on a single value of q, a profile of diversity, i.e. a plot of diversity against q, can be built (Tothmeresz, 1995). Low values of q (starting from 0) give much importance to rare species, whilst higher values (usually up to 2) focus on abundant species. Negative values of q are not used because of poor mathematical properties of their entropy (Beck, 2009), and values over 2 generally bring little more information. Ordering communities in terms of diversity requires that their profile do not cross (Tothmeresz, 1995); else, declaring a community more diverse than another only holds for a range of values of q reflecting the importance given to rare or frequent species (Lande et al., 2000).

To plot those profiles, diversity must be estimated from the data. Estimation bias (following the terminology of Dauby and Hardy, 2012) is a well-known issue (Marcon et al., 2014). Real data are almost always samples of larger communities, so some species may have been missed. The induced bias on the Simpson entropy is smaller than on the Shannon entropy because the former assigns lower weights to rare species,

i.e. the sampling bias is even more important when q decreases. Another estimation bias has been widely studied by physicists who generally consider that all species of a given community are known and their probabilities quantified. Their main issue is not at all missing species but the non-linearity of entropy measures (see Bonachela et al., 2008, for a short review). Estimating probabilities at power $q > 0$ by the power of their estimator is an important source of underestimation of entropy. The need for corrections has generated a considerable literature in ecological statistics and statistical physics.

This paper tests the performance of the state-of-the-art estimators when applied to the kind of data ecologists have to deal with. It starts with simulated distributions that have the advantage of being easily manipulated to generate various sampling intensities and evaluate the bias and root mean square error (RMSE) of the estimators. The classical models of the literature, namely the lognormal and the geometric distributions are addressed. The lognormal distribution describes, at least roughly, many hyper-diverse ecosystems even though the link between its statistical success and the underlying ecological mechanisms is poorly documented (Tokeshi, 1993). The geometric distribution is a far more difficult case because it is very uneven: the frequency of rare species is several orders of magnitude smaller than that of the frequent ones, making it impossible to observe with reasonable sampling effort (Haegeman et al., 2013). The best-known and best-performing estimators, including three new ones, are applied to those distributions and two actual forest data sets. The purpose of the paper is to provide recommendations about the estimation technique to choose when facing different types of data and draw general conclusions about the possible accuracy of diversity estimation.

Phylogenetic entropy is the sum of HCDT entropy along an ultrametric tree (Marcon and Hérault, 2015a) so estimating it reduces to estimating HCDT entropy. Phylogenetic diversity is then obtained as the deformed exponential of phyloentropy. In short, estimating phylodiversity relies on the methods presented here so the paper will focus on species-neutral diversity for clarity.

All analyses are made whith the package *entropart* (Marcon and Hérault, 2015b) for R (R Core Team, 2022). This paper comes with a Shiny application available on GitHub[1] to simulate communities and estimate their diversity.

2. Methods

Consider a community of species indexed by s. n_s is the number of individuals of species s sampled in the community, $n = \sum_s n_s$ the total number of sampled individuals. The (unknown) probability p_s for an individual to belong to species s is estimated by $\hat{p}_s = n_s/n$. The number of species represented by ν individuals in the sample of size n is $s_\nu^{(n)}$, so $s_0^{(n)}$ if the (unknown) number of unobserved species considering the sampling effort. $s_\nu^{(n)}$ is considered as a realization of the random variable $S_\nu^{(n)}$ so it is used to estimate its expectation $\mathbb{E}(S_\nu^{(n)})$.

π_ν is the sum of the probabilities p_s of species represented by ν individuals.

The deformed logarithm formalism (Tsallis, 1994) is very convenient to manipulate entropies. The deformed logarithm of order q is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1-q} \tag{1}$$

It converges to the natural logarithm when $q \to 1$.

The inverse function of $\ln_q x$ is the deformed exponential:

$$e_q^x = [1 + (1-q)x]^{\frac{1}{1-q}} \tag{2}$$

2.1 Sample coverage

The sample coverage (Good, 1953) is the probability for an individual in the community to belong to a species observed in the sample. It equals the sums of the probabilities of the observed species. It is an essential tool for diversity estimation because it is included in some estimators, e.g. Chao and Shen (2003), and it allows the evaluation of the completeness of the sampling (Chao and Jost, 2012). Its estimator given by Good is:

$$\hat{C} = 1 - \frac{s_1^{(n)}}{n} \tag{3}$$

It is biased (Zhang and Huang, 2007), because:

$$C = 1 - \frac{\mathbb{E}(S_1^{(n)}) - \pi_1}{n} \tag{4}$$

Good's estimator neglects the term π_1, the sum of the probabilities of singletons. It was built from Turing's frequency formula relating the average probability of species observed ν times to the number of species observed $\nu+1$ and ν times. This formula has been improved by Chao *et al.* (Chao and Shen, 2010; Chiu et al., 2014) to estimate π_1. Estimating the number of species by the Chao1 estimator (Chao, 1984), Chao and Shen (2010) obtained an improved estimator of the sample coverage:

$$\hat{C} = 1 - \frac{s_1^{(n)}}{n}\left[\frac{(n-1)\,s_1^{(n)}}{(n-1)\,s_1^{(n)} + 2s_2^{(n)}}\right] \tag{5}$$

This estimator has been further used by Chao and Jost (2015) to derive an estimator of entropy (see below).

[1] https://github.com/EricMarcon/estimation

An almost unbiased estimator has been derived using the information provided by the whole distribution (Chao et al., 1988; Zhang and Huang, 2007):

$$\hat{C} = 1 - \sum_{\nu=1}^{n} (-1)^{\nu+1} \binom{n}{\nu}^{-1} s_{\nu}^{(n)} \tag{6}$$

It is used in this paper.

2.2 Estimators of entropy

The existing estimators and the new ones proposed here can be classified into four main methods. The simplest one just consists of plugging the estimator of p_s, i.e. $\hat{p}_s = n_s/n$, into the definition of diversity to evaluate to obtain a so-called plug-in estimator, sometimes named naive estimator. The plug-in estimator of HCDT entropy of order q is:

$${}^{q}\hat{H} = \sum_{s} \hat{p}_s \ln_q \frac{1}{\hat{p}_s} \tag{7}$$

The plug-in estimator is useless in hyper-diverse communities because it severely underestimates diversity because of unobserved species and of the non-linearity of estimators.

Recent progress has been made in estimating the actual distribution of the probability of species by fitting a model of their distribution to the data. The distribution of the unobserved species can be added if their number is estimated and a distribution form is chosen. Chao et al. (2015) used a two-parameter model based on the estimation of the generalized sample coverage (not detailed here), estimated the total richness with the Chao1 estimator and modeled the unobserved species as a geometric distribution to unveil the complete rank-abundance distribution of an observed community. They applied the plug-in estimator this distribution: It will be called the "Chao-unveiled" here.

The Chao1 estimator was built according to the same theoretical approach as that of the unveiled rank-abundance distribution. It is a lower-bound estimator of the number of species. It has been improved by Chiu et al. (2014) who slightly reduced its negative bias with the iChao1 estimator, integrating species represented by 3 and 4 individuals. The "iChao-unveiled" estimator will be defined here as a variation on the "Chao-unveiled" estimator, where richness is estimated by the iChao1 estimator.

The jackknife estimator (Burnham and Overton, 1979) has shown good performances to estimate richness when the sampling effort is too low for the Chao1 estimator to perform well (Brose et al., 2003) even though it actually lacks theoretical support (Cormack, 1989). Estimating species richness with the jackknife estimator, whose order is selected according to the data, defines the "jackknife-unveiled" estimator. Using the jackknife estimator to unveil the tail of the abundance distribution was not the intention of Chao et al. (2015) because it is not consistent with their theoretical framework. It must be seen here as a merely empirical tool.

The second method relies on the Horvitz and Thompson (1952) estimator of the weighted sum of a function of its elements $x_1, x_2, ..., x_S$, say $\sum_s p_s f(x_s)$ when some of them are not observed. An unbiased estimator of the sum is obtained when each term is divided by its probability to be observed $1-(1-p_s)^n$. Chao and Shen (2003) proposed to combine it with the estimator of the sample coverage: conditionally to the set of observed species, an unbiased estimator (Ashbridge and Goudie, 2000) of p_s is $\tilde{p}_s = \hat{C}\hat{p}_s$. Chao and Shen estimated the Shannon entropy; the method has then been extended to HCDT entropy (Marcon et al., 2014) and similarity-based diversity (Marcon et al., 2014):

$${}^{q}\tilde{H} = \sum_{s} \frac{\hat{C}\hat{p}_s \ln_q \dfrac{1}{\hat{C}\hat{p}_s}}{1-\left(1-\hat{C}\hat{p}_s\right)^n} \tag{8}$$

A further progress can be done by replacing the conditional estimator of probabilities $\tilde{p}_s = \hat{C}\hat{p}_s$ by that of Chao et al. (2015). Since the improved probability estimator depends on the generalized sample coverage, the improved Chao-Shen estimator will be named the "generalized coverage" estimator.

The third method has been derived by Grassberger (1988) who gave a reduced-bias estimator of the value of an integer at power q. p_s^q is written as n_s^q/n^q and n_s^q is estimated (Marcon et al., 2014) as:

$$\tilde{n}_s^q = \frac{\Gamma(n_s+1)}{\Gamma(n_s-q+1)} + \frac{(-1)^n \Gamma(1+q)\sin\pi q}{\pi(n+1)} \tag{9}$$

The estimator of p_s^q is simply $\tilde{p}_s^q = \tilde{n}_s^q/n^q$. It is plugged into the formula of entropy to obtain the Grassberger estimator:

$${}^{q}\tilde{H} = \frac{1-\sum_s \tilde{p}_s^q}{q-1} \tag{10}$$

The last method has been the subject of an important literature in the last ten years. A review can be found in Chao et al. (2013), Appendix A. It relies on the estimation of $h_q = \sum_s p_s^q$. h_q can be written as the following sum:

$$h_q = \sum_{r=0}^{\infty} \binom{q-1}{r} (-1)^r \zeta_r \tag{11}$$

ζ_r is the generalized Simpson entropy $\sum_s p_s(1-p_s)^r$ defined by Zhang and Zhou (2010). The first n elements of the sum, denoted $\tilde{h}_q$, can be estimated with no bias (Zhang and Grabchak, 2016):

$$\tilde{h}_q = \sum_{s=1}^{S} \hat{p}_s \sum_{v=1}^{n-n_s} \left[\prod_{i=1}^{v} \frac{i-q}{i} \prod_{j=1}^{v} \left(1 - \frac{n_s-1}{n-j}\right) \right] \tag{12}$$

Zhang (2013) shows that the bias due to ignoring the remaining terms is asymptotically normal and decays exponentially fast. I'll call the Zhang and Grabchak (2016) estimator the one based on $\tilde{h}_q$:

$$ {}^q\tilde{H} = \frac{1-\tilde{h}_q}{q-1} \tag{13}$$

Some attempts have been made to estimate the remaining bias (Zhang and Grabchak, 2013). The most achieved one is that of Chao and Jost (2015), completing Chao et al. (2013). It relies on the estimation of the total number of species by the Chao1 estimator and a few approximations including that the actual probabilities of unobserved species can be assumed almost equal. A consequence is that the estimator of the average probability of species sampled once also equals the probability estimator of unobserved species. Its value is noted A. It is $2s_2^{(n)}/\left[(n-1)\,s_1^{(n)}+2s_2^{(n)}\right]$ if singletons and doubletons are present or $2/\left[(n-1)\left(s_1^{(n)}-1\right)+2\right]$ if doubletons are missing. The Chao-Jost estimator of HCDT entropy is:

$$ \begin{aligned} {}^q\tilde{H} = {} & \frac{1}{q-1}[1-\tilde{h}_q \\ & -\frac{s_1^{(n)}}{n}(1-A)^{1-n}\left(A^{q-1}-\sum_{r=0}^{n-1}\binom{q-1}{r}(A-1)^r\right)] \end{aligned} \tag{14}$$

In absence of singletons and doubletons, A is set to 1 and the estimator is identical to that of Zhang and Grabchak.

2.3 Confidence intervals

Two methods allow the evaluation of confidence intervals: asymptotic, closed forms are available for some estimators, or bootstrapping is required in the general case.

Esty (1983), completed by Zhang and Huang (2007), showed that the estimator of sample coverage (eq.(6) is asymptotically normal with the following confidence interval:

$$ C = \hat{C} \pm t_{1-\alpha/2}^{(n)} \frac{\sqrt{s_1^{(n)}\left(1-\frac{s_1^{(n)}}{n}\right)+2s_2^{(n)}}}{n} \tag{15}$$

Where $t_{1-\alpha/2}^{(n)}$ is the quantile of a Student distribution with n degrees of freedom at the risk threshold α, here 1.96 for all sample sizes and $\alpha = 5\%$.

The Zhang-Grabchak estimator is also asymptotically normal and comes with an asymptotic confidence interval (Zhang and Grabchak, 2016) implemented in the package *EntropyEstimation* (Cao and Grabchak, 2014).

The theoretical distribution of other estimators is unknown. They must be built by bootstrap techniques: the observed community is re-sampled, say 1000 times, and entropy is calculated each time. The $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of entropy are the bounds of the confidence interval. The issue of re-sampling a community is the same as that of sampling it: rare species are often eliminated, so the entropy is underestimated. Starting from the whole community, a first estimation bias is caused by sampling it. The estimators presented here aim at correcting it. When this observed community is re-sampled, a second estimation bias appears. Estimating the entropy of re-sampled communities with bias correction yields, on average, the entropy of the observed community estimated by the plug-in estimator (Marcon et al., 2012): if the estimator works well, it eliminates the second estimation bias but it cannot address the first one. The solution to this problem is simply to recenter the entropy distribution of re-sampled communities around the value of the entropy of the observed community (Marcon et al., 2012; Chao and Jost, 2015).

The re-sampling technique may just consist of drawing individuals in the observed community with replacement, or, equivalently, drawing a community in a multinomial distribution respecting the size and probability distribution of the observed community (Marcon et al., 2014). A more sophisticated technique has been proposed by Chao and Jost (2015). Given the sample size, the probability distribution of observed species can be estimated more accurately than by the estimator $\tilde{p}_s = \hat{C}\hat{p}_s$ which underestimates the probability of rare species (Chao et al., 2015). A better estimate of the probabilities is used (actually, a simplified version of that of the unveiled estimators above) and completed by an estimation of the number of unobserved species, whose probabilities are assumed identical. Despite these extra efforts, the distribution of the entropy of re-sampled community still has to be recentered.

The confidence interval of a biased estimator must be understood as the variability of its results. Since its bias is unknown, there is no guarantee that it contains the real value of diversity.

2.4 From entropy to diversity

All entropy estimations are finally transformed into diversity values to be interpretable (Jost, 2006). It is not correct to recenter the confidence interval of diversity estimations because of the non-linearity of the transformation of entropy into diversity (Marcon et al., 2012). The correct process consists of evaluating entropy with its confidence interval and make the final exponential transformation of all values into diversity.

2.5 Typical distributions

Comparing the performance of estimators requires simulations of realistic communities. The focus is put on two opposed models making sense in ecology. The log-normal distribution (Preston, 1948) fits well species-

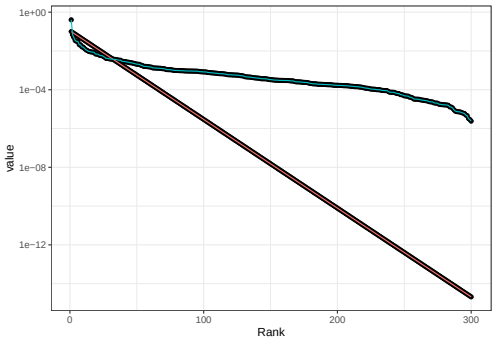

Figure 1. Rank-Abundance curves of 300 species following a lognormal (green, top curve) or a geometric distribution (red, straight line). The solid lines are the fitted models.

rich communities for several reasons, including populations dynamics (Engen and Lande, 1996), niche apportionment (Bulmer, 1974), or even statistical physics arguments (Pueyo et al., 2007; Dewar and Porté, 2008). It is often well fitted empirically (Tokeshi, 1990) even though it has been questioned theoretically (Williamson and Gaston, 2005). The local community distribution according to the neutral theory (Volkov et al., 2003) is not lognormal but departs from it very moderately. The logarithm of the species probabilities follows a Gaussian distribution.

The geometric series model (Motomura, 1932; Whittaker, 1972) generates far more uneven species distributions. In this model, the first species is represented by a part p of the total resources. The second one has the same part p of the remaining resources, and so on. Finally, probabilities are normalized to be proportional to the resources taken.

Artificial communities following those distributions are generated. Figure 1 presents a lognormal one, with log-standard-deviation equal to 2 (typical of the distribution of tree species in a rainforest) and a geometric distribution with parameter $p = 0.1$. Both contain 300 species.

2.6 Evaluation of the performance of estimators

The performance of each estimator was calculated as its average relative bias on all values of q (i.e. the average difference between the mean simulated entropy and its true value) and it Root Mean Square Error (RMSE, i.e. the square root of the sum of the squared bias and the variance, divided by the true value). The true entropy of each reference distribution was calculated with the known values of p_s. For each reference distribution, 1000 random samples of the chosen size were drawn in a multinomial distribution respecting the reference probabilities p_s. Entropy was calculated for q between 0 and 2. The average entropy and its first and last 2.5% quantiles were retained to build the profile and its confidence envelope (which is quite different from that of the estimation of real communities). Finally, entropy was transformed into diversity to be plotted.

3. Results

Simulations of independent sampling of individuals in real communities are multinomial samples of various sizes in the chosen species distributions. Sample sizes are between 200 and 5000 individuals to cover a range from obvious undersampling to a high-effort inventory: 5000 individuals correspond to 9 to 10 ha of forest.

3.1 Sample coverage

Performances of the estimator of sample coverage are evaluated first. 2 communities of each size between 200 and 5000 individuals were sampled in each typical distribution. The real and estimated sample coverages are compared on figure 2. The estimation of sample coverage is very efficient. A model II linear regression (Legendre, 2014) validated the accuracy of the estimation.

3.2 Entropy and diversity

The root mean square error of the estimators is shown on figure 3 for the lognormal and the geometric distributions with 300 species when 1000 individuals are sampled, a typical tropical forest inventory of trees.

Unsurprisingly, the plug-in estimator is severely biased and has the poorest results in the tests. The Chao-Jost estimator systematically outperforms the Zhang-Grabchak estimator (which actually performs little better than the plugin-estimator here) by construction. Its complementary estimation is not paid by increased variance. The Grassberger estimator is totally inefficient for low values of q as already noticed by Marcon et al. (2014). The generalized coverage estimator outperforms Chao-Shen because of its better estimation of conditional probabilities. The Chao-unveiled estimator is almost confused with the Chao-Jost estimator. Both are outperformed by the iChao-unveiled estimator because it improves the estimation of the number of species. The jackknife-unveiled estimator is more flexible than the previous ones to estimate the number of species. The order of the jackknife estimator it uses changes between simulations, causing an excessive variance for $q < 0.1$. It performs best for higher orders of diversity.

Results are consistent whatever the model. The general pattern is a poor estimation of low orders of diversity, and a quite accurate estimation of high orders, as previously shown by Haegeman et al. (2013). The RMSE varies a lot according to the model.

The discussion will focus on the best two estimators: Chao-Jost and jackknife-unveiled, ignoring the iChao-unveiled estimator which takes place between them but is too similar to the jackknife-unveiled to bring decisive arguments for the discussion. Figure 4 shows their profiles for a 1000-individual sample of a lognormal distribution of 300 species, with their confidence intervals.

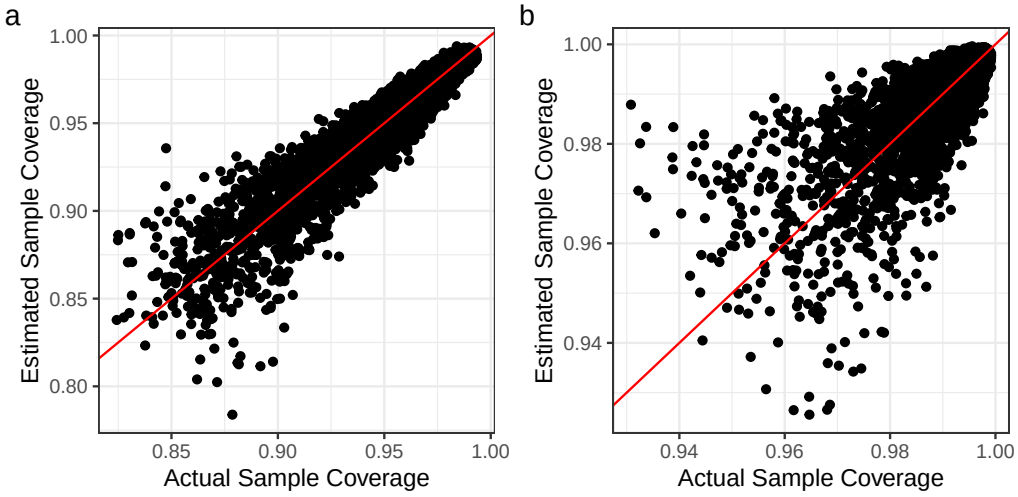

Figure 2. Estimated vs real sample coverage of simulated samples of a lognormal (a) or geometric (b) distribution of 300 species. Sample sizes are between 200 and 5000 individuals. The line represents the fit of a model II (Major Axis method) linear regression. a) Lognormal distribution: The estimated sample coverage is 0.991 times the real one plus 0.009, with an R^2 value around 94%. b) Geometric distribution: The estimated sample coverage is 1.046 times the real one minus 0.046, with an R^2 value around 69%.

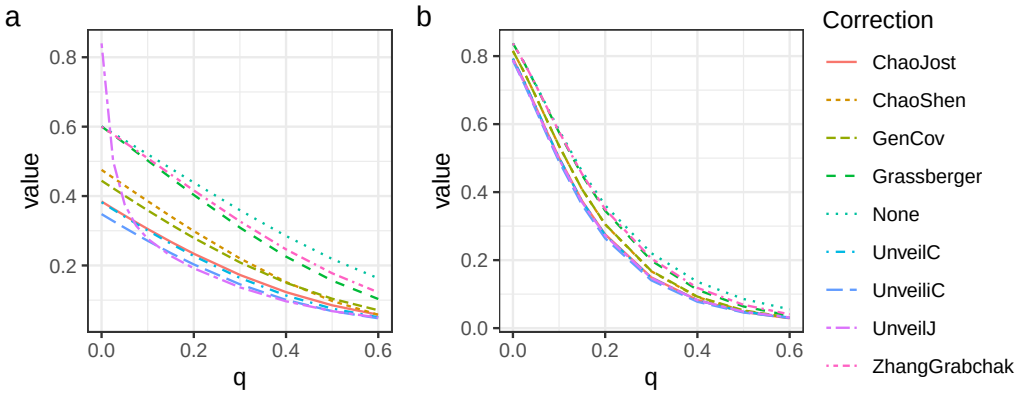

Figure 3. Estimated relative RMSE of the estimators of diversity based on 1000 samples of 1000 individuals of each typical distribution: lognormal (a) and geometric (b) of 300 species. The RMSE is normalized by the actual diversity. It is quite high for low orders of diversity, especially for the geometric distribution. Values of q over 0.6 are not shown because all estimators perform similarly well. The legend lists the estimators in the increasing order of RMSE for $q > 0.1$, where the estimator with the lowest RMSE is the jackknife-unveiled one, closely followed by the iChao-unveiled and Chao-Jost. Close to $q = 0$, the jackknife-unveiled estimator has a higher variance making it the least reliable estimator.

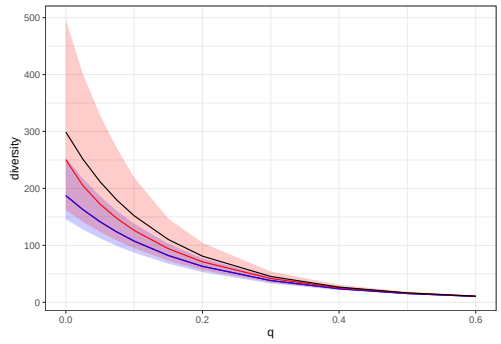

Figure 4. Diversity profiles estimated from 1000 random samples of 1000 individuals from a lognormal community of 300 species. The black line represents the real diversity (starting from ${}^0D \approx 300$). The jackknife-unveiled estimator is plotted by the red line (${}^0\hat{D} \approx 250$). Its confidence interval (red shade) is very wide. The Chao-Jost estimator (blue line: ${}^0\hat{D} \approx 200$) is more biased downward but its confidence interval (blue shade) is much smaller.

4. Discussion

The underlying distribution of species is the most important determinant of the success of diversity estimation: the estimation bias of heavy-tailed distributions decays more slowly when the sample size is increased (Zhang and Grabchak, 2013). Estimating the low-order diversity of a sample from a geometric distribution is all but impossible (Haegeman et al., 2013) but the low-order diversity of lognormal communities can be estimated meaningfully when the sample size is sufficient. Empirically, it is not possible to discriminate a severely-censured geometric distribution and a lognormal one (Tokeshi, 1993): both models fit well since most of the difference is contained by the unobserved tails of the distributions. So, theoretical, ecological arguments about the actual distribution of the community are necessary to decide whether an estimation of diversity is reliable.

Diversity of order over 0.5 is pretty well estimated in the context of this paper. Haegeman *et al.* showed that this remains true for $q \geq 1$, even when geometric communities of millions of species with parameter 0.5 (the most abundant species takes half the resources, the second one a quarter and so on) are addressed.

4.1 The sample coverage is not always the good indicator of the quality of estimation

The sample coverage can not be used as a proxy for how much an estimate of diversity can be relied upon. At the same sampling effort, the sample coverage appears to be higher for the geometric distribution: far more species are not sampled than in a lognormal distribution, but their total probability is smaller. For example, samples of 200 individuals drawn in 300-species geometric and lognormal communities yield an average estimation of 54 and 149 species by the jackknife-unveiled estimator, but the respective sample coverages are over 95% for the geometric distribution versus around 81% for the lognormal one. The estimation bias of the geometric distribution is thus much greater for low orders of diversity even though the sample coverage is higher.

Chao and Jost (2012) argue in favor of the sample coverage as a better measure of the sampling effort than the sample size: this must be understood as long as the underlying distribution of communities is the same. Then, standardizing the sampling effort by the sample coverage is pertinent.

4.2 Comparing the diversity of real communities with different distributions remains intractable

When the number of species of the theoretical distributions is doubled, everything else equal, the sampling bias increases. With the same sampling effort, the coverage of the lognormal distribution decreases. Doubling the effort brings both the sample coverage and the bias back to their previous level, with a reduced variance. These operations can be reproduced with the Shiny application provided with the paper.

This is a very simple and intuitive behavior, but it is completely different with the geometric distribution: the sample coverage does not change when richness is doubled because the probabilities of the 300 rarest species are negligible. Doubling the sample size does not restore the bias level. An extensive and rigorous analysis of the influence of the parameters of the theoretical distributions (beyond manipulating the number of species) is not the scope of this paper, but this simple example shows that no general and simple rules are available to compare the low-order diversity of communities of different nature.

4.3 Estimating the number of species is the critical step

The lower q, the more difficult the estimation is, but the estimation of the number of species has been long studied and simple rules of decision have been proposed (Burnham and Overton, 1979; Brose et al., 2003) to choose the most appropriate order of the jackknife estimator. Burnham and Overton derived a selection procedure to obtain the order allowing to minimize the RMSE of the estimation of the number of species. It is implemented in the package *SPECIES* [Wang (2011)} for R. Brose *et al.* showed (empirically) hat the first-order jackknife is selected when the sample completeness (terminology by Beck and Schwanghart, 2010), i.e. the proportion of observed species $(S - s_0^{(n)})/S$ is over 3/4 (precisely 74% in their paper). When it is less, higher orders have less bias but more variance. It is easy to estimate the number of species of an actual sample this way and compare it to the Chao1 estimator. If both coincide, the Chao-Jost estimator will perform well for the whole profile: its value at $q = 0$ is that of Chao1. Else, the jackknife-unveiled estimator will be the best choice since its value at $q = 0$ is the optimal-order jackknife. If one does not want to rely on the jackknife estimator for some reason, such

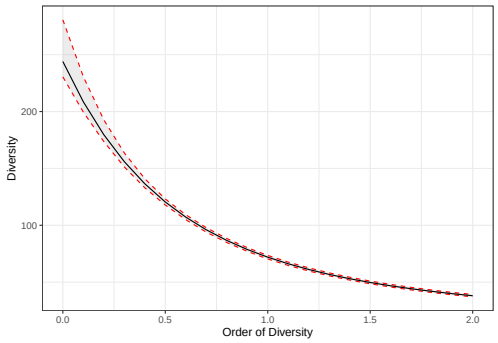

Figure 5. Estimated diversity profile of the tree species of the BCI 50-ha plot. The shaded area is the 95% confidence interval of the estimation.

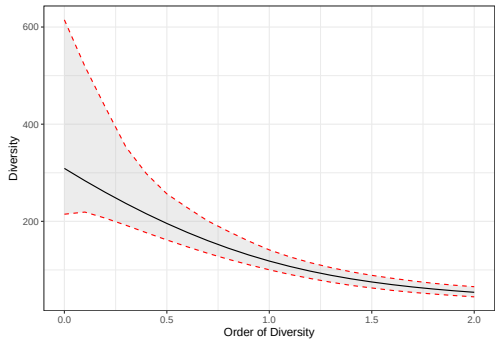

Figure 6. Estimated diversity profiles of the tree species of the Paracou 1-ha plot #18. The shaded zone is the 95% confidence interval of the estimation.

as its poor theoretical support, the iChao-unveiled estimator is a reasonable compromise as a lower bound estimation.

4.4 Better, but probably not much better, estimators may be derived

The most promising ways of research according to the present results are a better estimation of the remaining bias of the Zhang-Grabchak estimator and the improvement of the distribution modeling of the unveiled estimators. The first approach is that of the Chao-Jost estimator, which is limited by its estimation of the number of species (the lower bound, Chao1 estimator). The price for releasing this constraint is losing the elegant, closed form of the estimator allowed by appropriate approximations of the infinite sum of the unknown elements of eq.(11) for a numeric approximation.

The distribution of species is modeled with two parameters in the unveiled estimators. This can be refined by extending the technique presented by Chao et al. (2015) to higher orders of sample coverage. In both cases, better fitting the data to reduce the bias has its limits because the variance of estimation is likely to increase (Bonachela et al., 2008). So, the estimators presented here may not be far from the optimum trade-off (less bias with the jackknife-unveiled estimator, less variance with Chao-Jost).

5. Application to real data

The diversity of two real forest plots is estimated now. The first case is Barro Colorado Island's 50-ha plot of tropical forest, whose inventory data of trees over 10~cm diameter at breast height are available in the package *vegan* (Oksanen et al., 2012) for R. 225 species have been sampled, with a quite good fit to a log-normal distribution. The sample size is over 20000 individuals, the sample coverage is over 99.9%. Estimating the number of species with the Chao1 or the jackknife 1 estimators gives very similar results: 239 and 244 species. This is an unusually large dataset, whose diversity estimation (Figure 5) is quite easy.

The best estimator is Chao-Jost since the Chao1 estimator is appropriate for the number of species. The 95% confidence interval of the estimation is built by resampling according to the technique by Chao and Jost (2015). It is very small due to the abundance of data.

The second example takes place at the other extreme of sampling intensity. A 1-ha plot (plot 18) of tropical forest in the experimental forest of Paracou, French Guiana (Gourlet-Fleury et al., 2004), was inventoried. Data are available in the package *entropart* for R. Only 481 trees over 10 cm diameter at breast height have been sampled. They belong to 149 species. The sample coverage is $84.6 \pm 4.4\%$. The estimated number of species is 254 according to Chao1, but the appropriate Jackknife estimator (of order 3) returns 309 species. Clearly, the sampling effort is not sufficient for an accurate estimation: the sample coverage is too low and the estimation of the number of species too uncertain. With no doubt, the Chao-Jost estimator will severely underestimate diversity.

The jackknife-unveiled estimator is the best choice. Its confidence interval is very wide up to $q = 0.3$ (Figure 6). Over $q = 0.5$, the simulations above showed that the estimator has a very low variability, so the confidence interval is due to the uncertainty of the sampling only. At lower orders of diversity, the estimator's uncertainty amplifies it so the estimation is not reliable. In this case, the very little accuracy of the jackknife-unveiled estimator (the number of species is estimated between 237 and 439) is preferable to the far smaller confidence interval provided by a less variable but more biased estimator such as Chao-Jost that would probably not contain the actual values of low-order diversity (as in figure 4).

6. Conclusion

This paper tried to evaluate the performance of diversity estimation in real conditions with simulation studies covering a reasonable set of models. Unsurprisingly, estimating diversity is more difficult when the species distribution has a heavier tail and the number of species is greater. As of the state of the art, the

recommendation is to apply the Chao-Jost, the iChao-unveiled or the jackknife-unveiled estimator and consider diversity of order lower than 0.5 with caution.

When the sampling effort is high enough to allow a correct estimation of the number of species with the Chao1 estimator, the estimation by Chao-Jost is quite good down to $q = 0$. If this is not the case, the jackknife-unveiled estimator provides better results but with a higher variability. A conservative compromise for a first estimation of diversity, before choosing between Chao-Jost and jackknife-unveiled, is the iChao-unveiled estimator.

The *entropart* package for R allows computing species-neutral diversity and phylodiversity with all the estimators presented here.

7. Acknowledgments

This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01).

References

Ashbridge, J. and I. B. J. Goudie (2000). Coverage-adjusted estimators for mark-recapture in heterogeneous populations. *Communications in Statistics - Simulation and Computation 29*(4), 1215–1237.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics 50*(4), 495–510.

Beck, J. and W. Schwanghart (2010). Comparing measures of species diversity from incomplete inventories: An update. *Methods in Ecology and Evolution 1*(1), 38–44.

Bonachela, J. A., H. Hinrichsen, and M. A. Muñoz (2008). Entropy estimates of small data sets. *Journal of Physics A: Mathematical and Theoretical 41*(202001), 1–9.

Brose, U., N. D. Martinez, and R. J. Williams (2003). Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology 84*(9), 2364–2377.

Bulmer, M. G. (1974). On fitting the poisson lognormal distribution to species-abundance data. *Biometrics 30*(1), 101–110.

Burnham, K. P. and W. S. Overton (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology 60*(5), 927–936.

Cao, L. and M. Grabchak (2014). EntropyEstimation: Estimation of entropy and related quantities.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics 11*(4), 265–270.

Chao, A., T. C. Hsieh, R. L. Chazdon, R. K. Colwell, and N. J. Gotelli (2015). Unveiling the species-rank abundance distribution by generalizing good-turing sample coverage theory. *Ecology 96*(5), 1189–1201.

Chao, A. and L. Jost (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology 93*(12), 2533–2547.

Chao, A. and L. Jost (2015). Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution 6*(8), 873–882.

Chao, A., S.-M. Lee, and T.-C. Chen (1988). A generalized Good's nonparametric coverage estimator. *Chinese Journal of Mathematics 16*, 189–199.

Chao, A. and T.-J. Shen (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics 10*(4), 429–443.

Chao, A. and T.-J. Shen (2010). Program SPADE: Species prediction and diversity estimation. Program and user's guide.

Chao, A., Y.-T. Wang, and L. Jost (2013). Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution 4*(11), 1091–1100.

Chiu, C.-H., Y.-T. Wang, B. A. Walther, and A. Chao (2014). An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. *Biometrics 70*(3), 671–682.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics 45*(2), 395–413.

Daróczy, Z. (1970). Generalized information functions. *Information and Control 16*(1), 36–51.

Dauby, G. and O. J. Hardy (2012). Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. *Ecography 35*(7), 661–672.

Dewar, R. C. and A. Porté (2008). Statistical mechanics unifies different ecological patterns. *Journal of theoretical biology 251*(3), 389–403.

Engen, S. and R. Lande (1996). Population dynamic models generating the lognormal species abundance distribution. *Mathematical Biosciences 132*(2), 169–183.

Esty, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics 11*(3), 905–912.

Good, I. J. (1953). The population frequency of species and the estimation of population parameters. *Biometrika 40*(3/4), 237–264.

Gourlet-Fleury, S., J. M. Guehl, and O. Laroussinie (2004). *Ecology & Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Paris: Elsevier.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letters A 128*(6-7), 369–373.

Haegeman, B., J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, and J. S. Weitz (2013). Robust estimation of microbial diversity in theory and in practice. *The ISME journal 7*(6), 1092–101.

Havrda, J. and F. Charvát (1967). Quantification method of classification processes. Concept of structural alpha-entropy. *Kybernetika 3*(1), 30–35.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology 54*(2), 427–432.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Jost, L. (2006). Entropy and diversity. *Oikos 113*(2), 363–375.

Lande, R., P. J. DeVries, and T. R. Walla (2000). When species accumulation curves intersect: Implications for ranking diversity using small samples. *Oikos 89*(3), 601–605.

Legendre, P. (2014). Interpreting the replacement and richness difference components of beta diversity. *Global Ecology and Biogeography 23*(11), 1324–1334.

Marcon, E. and B. Hérault (2015a). Decomposing phylodiversity. *Methods in Ecology and Evolution 6*(3), 333–339.

Marcon, E. and B. Hérault (2015b). Entropart, an R package to measure and partition diversity. *Journal of Statistical Software 67*(8), 1–26.

Marcon, E., B. Hérault, C. Baraloto, and G. Lang (2012). The decomposition of Shannon's entropy and a confidence interval for *beta* diversity. *Oikos 121*(4), 516–522.

Marcon, E., I. Scotti, B. Hérault, V. Rossi, and G. Lang (2014). Generalization of the partitioning of Shannon diversity. *Plos One 9*(3), e90289.

Marcon, E., Z. Zhang, and B. Hérault (2014). The decomposition of similarity-based diversity and its bias correction. *HAL 00989454*(version 3).

Motomura, I. (1932). On the statistical treatment of communities. *Zoological Magazine 44*, 379–383.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner (2012). Vegan: Community ecology package.

Patil, G. P. and C. Taillie (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association 77*(379), 548–561.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology 29*(3), 254–283.

Pueyo, S., F. He, and T. Zillio (2007). The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecology letters 10*(11), 1017–28.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tokeshi, M. (1990). Niche apportionment or random assortment: Species abundance patterns revisited. *Journal of Animal Ecology 59*(3), 1129–1146.

Tokeshi, M. (1993). Species abundance patterns and community structure. *Advances in Ecological Research 24*, 111–186.

Tothmeresz, B. (1995). Comparison of different methods for diversity ordering. *Journal of Vegetation Science 6*(2), 283–290.

Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics 52*(1), 479–487.

Tsallis, C. (1994). What are the numbers that experiments provide? *Química Nova 17*(6), 468–471.

Volkov, I., J. R. Banavar, S. P. Hubbell, and A. Maritan (2003). Neutral theory and relative species abundance in ecology. *Nature 424*(6952), 1035–1037.

Wang, J.-P. (2011). SPECIES: An R package for species richness estimation. *Journal of Statistical Software 40*(9), 1–15.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon 21*(2/3), 213–251.

Williamson, M. and K. J. Gaston (2005). The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *Journal of Animal Ecology 74*(2001), 409–422.

Zhang, Z. (2013). Asymptotic normality of an entropy estimator with exponentially decaying bias. *IEEE Transactions on Information Theory 59*(1), 504–508.

Zhang, Z. and M. Grabchak (2013). Bias adjustment for a nonparametric entropy estimator. *Entropy 15*(6), 1999–2011.

Zhang, Z. and M. Grabchak (2016). Entropic representation and estimation of diversity indices. *Journal of Nonparametric Statistics 28*(3), 563–575.

Zhang, Z. and H. Huang (2007). Turing's formula revisited. *Journal of Quantitative Linguistics 14*(2-3), 222–241.

Zhang, Z. and J. Zhou (2010). Re-parameterization of multinomial distributions and diversity indices. *Journal of Statistical Planning and Inference 140*(7), 1731–1738.