The Generalized Simpson's Entropy is a Measure of Biodiversity

Michael Grabchak, Eric Marcon, Gabriel Lang, Zhiyi Zhang

The Generalized Simpson's Entropy is a Measure of Biodiversity

Michael Grabchak[1], Eric Marcon[2]*, Gabriel Lang[3], Zhiyi Zhang[1]

Abstract
Modern measures of diversity satisfy reasonable axioms, are parameterized to produce diversity profiles, come with an effective number of species to simplify their interpretation, and estimators to apply them to real data. Generalized Simpson's entropy has all of these features and can be used as a measure of biodiversity. Moreover, unlike most commonly used diversity indices, it has unbiased estimator, which allows for robust estimation of the diversity of poorly sampled, rich communities.

Keywords
diversity, estimators, rare species.

[1]*Department of Mathematics and Statistics, University of North Carolina at Charlotte. Charlotte, NC 28223*
[2]*AgroParisTech, UMR EcoFoG, CNRS, CIRAD, INRA, Université des Antilles, Université de Guyane. BP 709, F-97310 Kourou, French Guiana.*
[3]*UMR 518 Mia, AgroParisTech, INRA, Université Paris-Saclay. F-75015 Paris, France.*
***Corresponding author**: Eric.Marcon@ecofog.gf

Contents

1 Introduction **1**
2 Methods **2**
2.1 Generalized Simpson's Entropy . . . 2
2.2 Axioms for a measure of diversity . . . 2
2.3 The generalized Simpson's entropy is a measure of diversity 3
2.4 Estimation . . . 3
2.5 Comparing distributions . . . 4
2.6 Effective number of species . . . 4
3 Example Data and Results **4**
4 Discussion **5**
4.1 Interpretation . . . 5
4.2 HCDT entropy . . . 6
4.3 Hurlbert's diversity . . . 7
5 Conclusion **7**
6 Appendix 1: Proofs **8**
6.1 Proof of Proposition 1 . . . 8
6.2 Gradient . . . 8
6.3 Hessian . . . 9
6.4 Extremum when all probabilities are equal . . . 9
6.5 Maximum of the function . . . 9
6.6 Confidence intervals . . . 10
7 Appendix 2: R code **11**

1. Introduction

Many indices of biodiversity have been proposed based on different definitions of diversity and different visions of the biological aspects to address (Ricotta, 2005). Indeed, measuring diversity requires both a robust theoretical framework (Patil and Taillie, 1982) and empirical techniques to effectively estimate it (Beck and Schwanghart, 2010). We focus on species-neutral diversity, i.e. the diversity of the distribution of species, ignoring their features. Classical measures of this type of diversity include richness (the number of species), Shannon's entropy (Shannon, 1948), and Simpson's index (Simpson, 1949).

Since one index is generally insufficent to fully capture the diversity of a community, modern measures of diversity are parameterizable, allowing the user to give more or less relative importance to rare versus frequent species (Rényi, 1961). Further, they can be expressed as an effective number of species (Hill, 1973), which allows for an easy interpretation of their value (Jost, 2006). Among the most popular of these are HCDT entropy (Havrda and Charvát, 1967; Daróczy, 1970; Tsallis, 1988), which includes richness, Simpson's index, and Shannon's entropy as special cases, Rényi's entropy (Rényi, 1961), and the less-used Hurlbert's index (Hurlbert, 1971). These indices can be used to estimate the diversity of a community and then to plot its value against the parameter that controls the weight of rare species to obtain a diversity profile (Hill, 1973). The profiles of two communities can be compared to provide a partial order of their diversity. If the profiles do not cross, one community can be declared more diverse than the other (Tothmeresz, 1995).

HCDT entropy has many desirable properties (Jost, 2006; Marcon *et al.*, 2014) but, despite recent progress (Chao and Jost, 2015), it cannot be accurately estimated when the communities are insufficiently sampled (Marcon, 2015). Rényi's entropy is related to HCDT entropy by a

straightforward transformation: the natural logarithm of the deformed exponential (Marcon *et al.*, 2014). Its properties are very similar and, hence, it will not be treated here. Hurlbert's index has a simple practical interpretation and can be estimated with no bias, but only up to when its parameter is less than the sample size.

We introduce generalized Simpson's entropy as a measure of biodiversity for its particular performance when it is used to estimate the diversity of small samples of hyper-diverse communities. The generalized Simpson's entropy ζ_r is parameterized: increasing its parameter r gives more importance to rare species. It has a simple interpretation, specifically, in a species accumulation curve, ζ_r is the probability that the individual sampled at rank $r+1$ belongs to a new species. We show that ζ_r is a valid measure of diversity, satisfying the axioms established in the literature (Rényi, 1961; Patil and Taillie, 1982). We then show how to estimate ζ_r with no bias and how to construct confidence intervals, which can be used to compare the diversities of different communities. Next, we derive a simple formula for the corresponding effective number of species and discuss its estimation. Finally, we compare it to HCDT entropy and Hurlbert's index on a real-world example of under-sampled tropical forest to illustrate its decisive advantage when applied to this type of data.

2. Methods

2.1 Generalized Simpson's Entropy

Let $\ell_1, \ell_2, \ldots, \ell_S$ be the species in a community, and let p_s be the proportion of individuals belonging to species ℓ_s. Necessarily, $0 \leq p_s \leq 1$ and $\sum_{s=1}^{S} p_s = 1$. We can interpret p_s as the probability of seeing an individual of species ℓ_s when sampling one individual from this community. Generalized Simpson's entropy is a family of diversity indices defined by

$$\zeta_r = \sum_{k=1}^{S} p_k (1-p_k)^r, \quad r = 1, 2, \ldots. \tag{1}$$

The parameter r is called the order of ζ_r. Note that, as r increases, ζ_r gives more relative weight to rare species than to more common ones. Note further that $0 \leq \zeta_r \leq 1$. In fact ζ_r is the probability that the $(r+1)$st observation will be of a species that has not been observed before.

Generalized Simpson's entropy was introduced as part of a larger class in Zhang and Zhou (2010) and was further studied in Zhang and Grabchak (2014). The name comes from the fact that $1-\zeta_1$ corresponds to Simpson's index as defined in Simpson (1949). A major advantage to working with this family is that there exists an unbiased estimator of ζ_r whenever r is strictly less than the sample size. While a similar result holds for Hurlbert's index, this is not the case with most popular diversity indices including HCDT entropy and Rényi's entropy, which do not have unbiased estimators. Now, we turn to the question of when and why generalized Simpson's entropy is a good measure of diversity.

2.2 Axioms for a measure of diversity

Historically, measures of diversity have been defined as functions mapping the proportions $p_1, p_2, \ldots, p_S$ into the real line, and satisfying certain axioms. We write $H(p_1, p_2, \ldots, p_S)$ to denote a generic function of this type. We begin with three of the most commonly assumed axioms. The first two are from Rényi (1961) after Faddeev (1956).

Axiom 1 (Symmetry)**.** *$H(p_1, p_2, \ldots, p_S)$ must be a symmetric function of its variables.*

This means that no species can have a particular role in the measure.

Axiom 2 (Continuity)**.** *$H(p_1, p_2, \ldots, p_S)$ must be a continuous function of the vector $(p_1, p_2, \ldots, p_S)$.*

This ensures that a small change in probabilities yields a small change in the measure. In particular, two communities differing by a species with a probability very close to 0 have almost the same diversity.

Axiom 3 (Evenness)**.** *For a fixed number of species S, the maximum diversity is achieved when all species probabilities are equal, i.e.,*

$$H(p_1, p_2, \ldots, p_S) \leq H(1/S, 1/S, \ldots, 1/S).$$

This axiom was called evenness by Gregorius (2014). It means that the most diverse community is the one where all species have the same proportions.

We will give a more restrictive version of this axiom. Toward this end, following Patil and Taillie (1982), we define a *transfer of probability*. This is an operation that consists of taking two species with $p_s < p_t$ and modifying these probabilities to increase p_s by $h > 0$ and decrease p_t by h, such that we still have $p_s + h \leq p_t - h$. In other words, some individuals of a more common species are replaced by ones of a less common species, but in such a way that the order of the two species does not change.

Axiom 4 (Principle of transfers)**.** *Any transfer of probability must increase diversity.*

The principle of transfers comes from the literature of inequality (Dalton, 1920). It is clear that this axiom is stronger than the axiom of evenness: if any transfer increases diversity, then, necessarily, the maximum value is reached when no more transfer is possible, i.e. when all proportions are equal.

Generalized Simpson's entropy belongs to an important class of diversity indices, which are called trace-form entropies in statistical physics and dichotomous diversity

indices in Patil and Taillie (1982). This class consists of indices of the form $H(p_1, p_2, \ldots, p_S) = \sum_{s=1}^{S} p_s I(p_s)$, where $I(p)$ is called the information function. Indices of this type were studied extensively in Patil and Taillie (1982) and Gregorius (2014). $I(p)$ defines the amount of information (Shannon, 1948), or uncertainty (Rényi, 1961), or surprise (Marcon and Hérault, 2015). All of these terms can be taken as synonyms; they get at the idea that $I(p)$ measures the rarity of individuals from a species with proportion p (Patil and Taillie, 1982). This discussion leads to the following axiom.

Axiom 5 (Decreasing information)**.** *$I(p)$ must be a decreasing function of p on the interval $(0,1]$ and $I(1) = 0$.*

This can be interpreted to mean that observing an individual from an abundant species brings less information than observing one from a rare species, and if an individual is observed from a species that has probability 1, then this observation brings no information at all.

Patil and Taillie (1982) showed that Axiom 5 ensures that adding a new species increases diversity. They also showed that both the principle of transfers and the axiom of decreasing information are satisfied if the function $g(p) = pI(p)$ is concave on the interval $[0,1]$. However, for generalized Simpson's entropy,

$$g(p) = p(1-p)^r, \quad p \in [0,1] \tag{2}$$

is not a concave function of p if $r > 1$. In fact, for $r > 1$ generalized Simpson's entropy does not satisfy the principle of transfers. For this reason Gregorius (2014), in a study of many different entropies, did not retain it. However, we will show that generalized Simpson's entropies satisfy a weaker version of the principle of transfers, and are, nevertheless, useful measures of diversity.

2.3 The generalized Simpson's entropy is a measure of diversity

It is easy to see that generalized Simpson's entropy always satisfies Axioms 1, 2 and 5, but, as we have discussed, it does not satisfy Axiom 4. However, we will show that it satisfies a weak version of it and that it satisfies Axiom 3 for a limited, but wide range of orders r.

Axiom 6 (Weak principle of transfers)**.** *Any transfer of probability must increase diversity as long as the sum of the probabilities of the concerned species is below a certain threshold, i.e., the principle of transfers holds so long as*

$$p_s + p_t \leq T \text{ for some } T \in (0,1].$$

We now give our results about the properties of generalized Simpson's entropy. The proofs are in Appendix 1.

Proposition 1. *Generalized Simpson's entropy of order r respects the weak principle of transfers with $T = \frac{2}{r+1}$.*

Proposition 2. *Generalized Simpson's entropy of order r respects the evenness axiom if $r \leq S-1$.*

In light of Proposition 2, we will limit the order to $r = 1, 2, \ldots, (S-1)$. In this case, generalized Simpson's entropy satisfies Axioms 1-3, and can be regarded as a measure of diversity. Moreover, it satisfies Axiom 5 and the weak principle of transfers up to $T = \frac{2}{r+1} \geq \frac{2}{S}$. Thus, a transfer of probability increases diversity, except between very abundant species.

2.4 Estimation

In practice, the proportions, $(p_1, p_2, \ldots, p_S)$, are unknown and, hence, the value of generalized Simpson's entropy as well as any other diversity index is unknown and can only be estimated from data. Toward this end, assume that we have a random sample of n individuals from a given community. Let n_s be the number of individuals sampled from species ℓ_s, and note that $n = \sum_{s=1}^{S} n_s$. We can estimate p_s by $\hat{p}_s = n_s/n$.

A naive estimator of ζ_r is given by the so-called "plug-in" estimator $\sum_{s=1}^{S} \hat{p}_s(1-\hat{p}_s)^r$. Unfortunately, this may have quite a bit of bias. However, for $1 \leq r \leq (n-1)$, an unbiased estimator of ζ_r exists and is given by

$$Z_r = \frac{n^{r+1}[n-r-1]!}{n!} \sum_{s=1}^{S} \hat{p}_s \prod_{j=0}^{r-1} \left(1 - \hat{p}_s - \frac{j}{n}\right), \tag{3}$$

see Zhang and Zhou (2010). There it is shown that Z_r is a uniformly minimum variance unbiased estimator (umvue) for ζ_r when $1 \leq r \leq (n-1)$.

Note that the sum in (3) ranges over all of the species in the community. This may appear impractical since we generally do not know the value of S. However, for any species ℓ_s that is not observed in our sample we have $\hat{p}_s = 0$, and we do not need to include it in the sum. Assume that we have observed $K \leq S$ different species in the sample and that these species are $\ell'_1, \ell'_2, \ldots, \ell'_K$. For each $s = 1, 2, \ldots, K$, let n'_s be the number of individuals from species ℓ'_s sampled, and let $\hat{p}'_s = n'_s/n$ be the estimated proportion of species ℓ'_s. In this case we can write

$$Z_r = \frac{n^{r+1}[n-r-1]!}{n!} \sum_{s=1}^{K} \hat{p}'_s \prod_{j=0}^{r-1} \left(1 - \hat{p}'_s - \frac{j}{n}\right). \tag{4}$$

With a few simple algebraic steps, we can rewrite this in the form

$$Z_r = \sum_{s=1}^{K} \hat{p}'_s \prod_{j=1}^{r} \left(1 - \frac{n'_s - 1}{n-j}\right), \tag{5}$$

which we have found to be more tractable for computational purposes.

In Zhang and Zhou (2010) and Zhang and Grabchak (2014) it is shown that Z_r is consistent and asymptotically normal. These facts can be used to construct asymptotic

confidence intervals. Toward this end define the $(K-1)\times(K-1)$ dimensional matrix given by

$$\hat{\Sigma} = \begin{pmatrix} \hat{p}'_1(1-\hat{p}'_1) & -\hat{p}'_1\hat{p}'_2 & \cdots & -\hat{p}'_1\hat{p}'_{K-1} \\ -\hat{p}'_2\hat{p}'_1 & \hat{p}'_2(1-\hat{p}'_2) & \cdots & -\hat{p}'_2\hat{p}'_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ -\hat{p}'_{K-1}\hat{p}'_1 & -\hat{p}'_{K-1}p'_2 & \cdots & \hat{p}'_{K-1}(1-\hat{p}'_{K-1}) \end{pmatrix}$$

and the $(K-1)$ dimensional column vector $\hat{h}_r$, where for each $j = 1,\cdots,(K-1)$ the jth component of $\hat{h}_r$ is given by

$$\left(1-\hat{p}'_j\right)^r + r\hat{p}'_j\left(1-\hat{p}'_j\right)^{r-1} - \left(1-\hat{p}'_K\right)^r - r\hat{p}'_K\left(1-\hat{p}'_K\right)^{r-1}.$$

When $r \le (S-1)$ and there exists at least one s with $p_s \neq 1/S$ (i.e. we do not have a uniform distribution) then an asymptotic $(1-\alpha)100\%$ confidence interval for ζ_r is given by

$$Z_r \pm z_{\alpha/2}\frac{\hat{\sigma}_r}{\sqrt{n}},$$

where

$$\hat{\sigma}_r = \sqrt{\hat{h}_r^T\hat{\Sigma}\hat{h}_r} \tag{6}$$

is the estimated standard deviation, $\hat{h}_r^T$ is the transpose of $\hat{h}_r$, and $z_{\alpha/2}$ is a number satisfying $P(Z > z_{\alpha/2}) = \alpha/2$ where $Z \sim N(0,1)$ is a standard normal random variable. Methods for evaluating Z_r and $\hat{\sigma}_r$ are available in the package *EntropyEstimation* (Cao and Grabchak, 2014) for R (R Development Core Team, 2016). For details about the confidence interval see Appendix 1.

2.5 Comparing distributions

In many situations it is important not only to estimate the diversity of one community, but to compare the diversities of two different communities. To do this we discuss the construction of confidence intervals for the difference between the generalized Simpson's entropies of two communities.

Fix an order r and let $\zeta_r^{(1)}$ and $\zeta_r^{(2)}$ be the generalized Simpson's entropies of the first and second community respectively. To estimate these assume that we have a random sample of size n_1 from the first community and a random sample of size n_2 from the second community. Assume further that these two samples are independent of each other and that $r \le (\min\{n_1,n_2\}-1)$, where $\min\{n_1,n_2\}$ is the minimum of n_1 and n_2. If both communities satisfy the conditions given in Section 2.4, an asymptotic $(1-\alpha)100\%$ confidence interval for the difference $\zeta_r^{(1)} - \zeta_r^{(2)}$ is given by

$$\left[Z_r^{(1)} - Z_r^{(2)}\right] \pm z_{\alpha/2}\sqrt{\frac{\left[\hat{\sigma}_r^{(1)}\right]^2}{n_1} + \frac{\left[\hat{\sigma}_r^{(2)}\right]^2}{n_2}},$$

where $Z_r^{(1)}$ and $Z_r^{(2)}$ are the estimates of $\zeta_r^{(1)}$ and $\zeta_r^{(2)}$ and $\hat{\sigma}_r^{(1)}$ and $\hat{\sigma}_r^{(2)}$ are the estimated standard deviations as in (6).

In practice, it is often not enough to look at only one diversity index. For this reason we may want to look at an entire profile of generalized Simpson's entropies. This can be done as follows. Fix any positive integer $v \le (\min\{n_1,n_2\}-1)$. For each $r = 1,2,\ldots,v$ we can estimate $Z_r^{(1)}$, $Z_r^{(2)}$, and the corresponding confidence interval. Looking at these for all values of r gives a pointwise confidence envelop. We can now see if the two communities have statistically significant differences in the amount of diversity by seeing if zero is in the envelop or not. If it is generally in the envelop then the differences are not significant, and if it is generally outside of the envelop then the differences are significant.

2.6 Effective number of species

The effective number of species (Hill, 1973) is the number of equiprobable species that would yield the same diversity as the data (Gregorius, 1991). It is a measure of diversity *sensu sticto* (Jost, 2006). We will write *entropy* for ζ_r and *diversity* for its effective number, which we denote by ${}^rD^\zeta$. To derive ${}^rD^\zeta$ we assume

$$\zeta_r = \sum_{s=1}^{{}^rD^\zeta}\frac{1}{{}^rD^\zeta}\left(1-\frac{1}{{}^rD^\zeta}\right)^r, \tag{7}$$

and then simple algebra yields

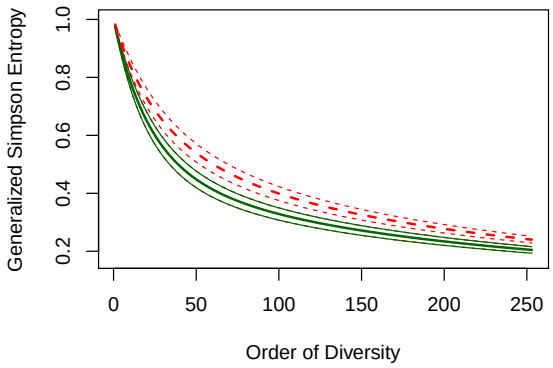
$${}^rD^\zeta = \frac{1}{1-\zeta_r^{\frac{1}{r}}}. \tag{8}$$

Note that (7) assumes that ${}^rD^\zeta$ is an integer, while in (8) it is generally not an integer. This is not an issue because (7) is just a formalism used to derive (8). A more developed argumentation can be found in Gregorius (2014), Appendix B.

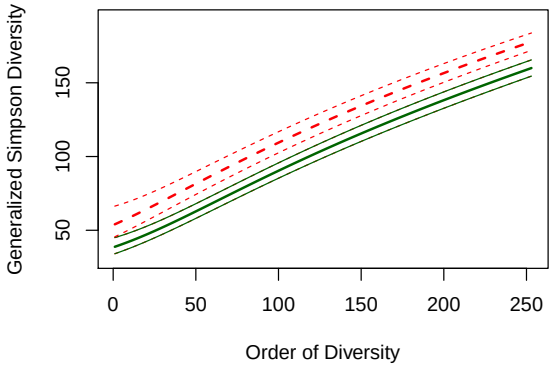
Since the function $f(t) = 1/(1-t^{1/r})$, $t \in [0,1]$ is monotonely increasing, we can transforms confidence intervals for ζ_r into confidence intervals for ${}^rD^\zeta$ as follows. If (L,U) is a $(1-\alpha)100\%$ confidence interval for ζ_r then $(f(L),f(U))$ is a $(1-\alpha)100\%$ confidence interval for ${}^rD^\zeta$. It is important to note that any inference based on such confidence intervals for ${}^rD^\zeta$ is equivalent to inference based on the original confidence interval for ζ_r.

3. Example Data and Results

In this section we apply our methodology to estimate and compare the diversities of two 1-ha plots (#6 and #18) of tropical forest in the experimental forest of Paracou, French Guiana (Gourlet-Fleury *et al.*, 2004). Respectively 641 and 483 trees with diameter at breast height over 10 cm were inventoried. The data is available in the *entropart* package for R.



(a) Generalized Simpson's entropy profile.



(b) Generalized Simpson's diversity profile.

Figure 1. Generalized Simpson's (a) entropy and (b) diversity profiles of Paracou plots 6 (solid, green lines) and 18 (dotted, red lines). The bold lines represent the estimated values, surrounded by their 95% confidence envelopes.

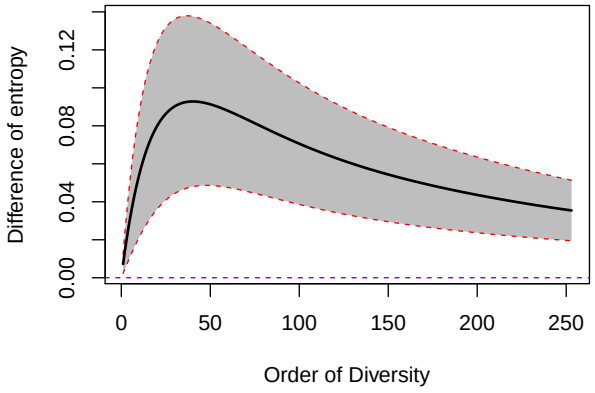


Figure 2. Difference between the generalized Simpson's entropy of plots 6 and 18 with their 95% confidence envelope. The horizontal dotted line represents the null hypothesis of identical diversity. Since it is always outside of the confidence envelope, identical diversity is rejected.

In the data we observe 147 and 149 species from plots 6 and 18 respectively. However, both plots are poorly sampled and we must adjust these values. The best estimators of richness are jackknives (Burnham and Overton, 1979). We use a jackknife of order 2 for plot 6 and one of order 3 for plot 18: the choice of the optimal order follows both Burnham and Overton (1979) and Brose *et al.* (2003). The estimated richness is, respectively, 254 and 309 species. For this reason we estimate generalized Simpson's entropy up to order $r = 253$. This, along with a 95% confidence envelope is given in Figure 1a.

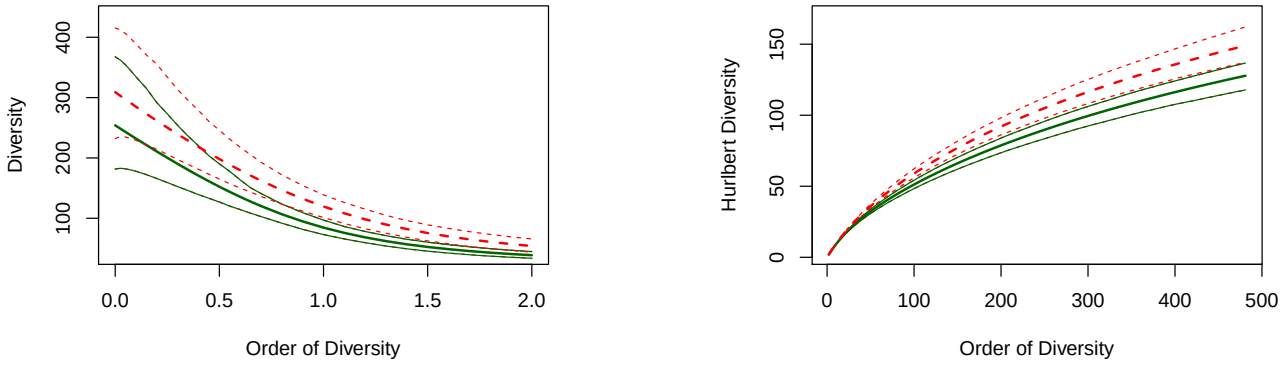
The generalized Simpson's diversity profiles along with a 95% confidence envelope are given in Figure 1b. These give more intuitive information since they represent the effective numbers of species. Their values at $r = 1$ are given, respectively, by 39 and 46 species. Increasing values of r give more importance to rare species, which leads to the increase in the effective number of species seen in the graph.

Plot 18 is undoubtedly more diverse than plot 6, with a fairly stable difference of between 15 and 19 effective species. In Figure 2 the difference between the entropies is plotted with its 95% confidence envelope to test it against the null hypothesis of zero difference. Since zero is never in this envelope, we concude that plot 18 is significantly more diverse than plot 6.

4. Discussion

4.1 Interpretation

Generalized Simpson's entropy of order r can be interpreted as the average information brought by the observation of an individual. Its information function $I(p) = (1-p)^r$ represents the probability of not observing



(a) HCDT diversity profile.

(b) Hurlbert's diversity profile.

Figure 3. (a) HCDT and (b) Hurlbert's diversity profiles of Paracou plots 6 (solid, green lines) and 18 (dotted, red lines). The bold lines represent the estimated values, surrounded by their 95% confidence envelope (obtained by 1000 bootstraps)

a single individual of a species with proportion p in a sample of size r. Thus I is an intuitive measure of rarity.

Chao *et al.* (2013) interpreted ζ_r as the probability that the individual sampled at rank $(r+1)$ belongs to a previously unobserved species in a species accumulation curve. A related interpretation is as follows. If X is the number of species observed exactly once in a sample of size $(r+1)$, then $\zeta_r = \mathrm{E}[X]/(r+1)$.

These interpretations are not limited to orders $r < S$. However, when $r \geq S$, ζ_r is no longer a reasonable measure of diversity. Further, in this case, it may not be maximized at the uniform distribution, which could lead the effective number of species, ${}^{r}D^{\zeta}$, to be greater than the actual number of species.

4.2 HCDT entropy

In this section we compare our results to those based on the more standard HCDT entropy, which is given by

$$ {}^{q}T = \frac{\sum_{s=1}^{S} p_s^q - 1}{1-q}, \; q \geq 0, $$

where for $q = 1$ this is interpreted by its limiting value as ${}^{1}T = -\sum_{s=1}^{S} p_s \log p_s$. The effective number of species for HCDT entropy was derived in Hill (1973). It is given by

$$ {}^{q}D^{T} = \left(\sum_{s=1}^{S} p_s^q \right)^{1/(1-q)}, \; q \geq 0, $$

where for $q = 1$ this is interpreted by its limiting value as ${}^{q}D^{T} = e^{{}^{1}T}$. We call this quantity HCDT diversity, although in the literature it is often called Hill's diversity number. For our data, plots of ${}^{q}D^{T}$ for $q \in [0,2]$ along with a 95% confidence envelope are given in Figure 3a. Here ${}^{q}D^{T}$ was estimated using the jackknife-unveiled estimator of Marcon (2015) and the confidence envelope was estimated using bootstrap.

It is easy to see that the importance of rare species increases for HCDT entropy as q *decreases*. In comparison the importance of rare species for generalized Simpson's entropy increases as r *increases*. Note that ${}^{2}T = \zeta_1$. To see what values of q in HCDT entropy correspond to other values of r for generalized Simpson's entropy, we can find when ${}^{r}D^{\zeta} = {}^{q}D^{T}$. Since we can only use ζ_r up to $r = S-1$ it is of interest to find which value of q corresponds to this value. For our data we find that in plot 6 $q = 0.5$ corresponds to $r = 253$ and in plot 18 $q = 0.55$ corresponds to $r = 308$.

The main difficulty in working with HCDT entropy is that its estimators have quite a bit of bias especially for smaller values of q (Marcon, 2015). This is illustrated in Figure 3a, where we see that the confidence intervals of the estimated values of the HCDT diversity of plots 6 and 18 have significant overlap up to $q = 0.75$.

Bias is not an issue with generalized Simpson's entropy, which can be estimated with no bias, regardless of the sample size. However, the issue with generalized Simpson's entropy is that it can only be considered for orders $r \leq S-1$, and larger values of r correspond to smaller values of q for HCDT entropy. In our example, the generalized Simpson's diversity profile can be compared to the part of the HCDT diversity profile between $q = 0.5$ and $q = 2$. Focusing more on rare species is not possible. HCDT diversity allows that theoretically, but is seriously limited by its estimation issues: the profile has a wide confidence envelope and is not conclusive below

$q = 0.75$.

On the whole, generalized Simpson's entropy allows a more comprehensive comparison of diversity profiles. If richness were greater, higher orders of generalized Simpson's diversity could be used and estimated with no bias, while low-order HCDT estimation would get more uncertain (Marcon, 2015).

4.3 Hurlbert's diversity

Another measure of diversity, which is related to generalized Simpson's entropy, was introduced in Hurlbert (1971). It is given by

$$ {}^{k}H = \sum_{s=1}^{S} \left[1 - (1 - p_s)^k \right], \; k = 1, 2, \ldots, $$

and corresponds to the expected number of species found in a sample of size k. It is easily verified that ${}^{1}H = 1$ and that ${}^{2}H = 1 + \zeta_1$. The higher the value of k, the greater the importance given to rare species. While there is no simple formula for the corresponding effective number of species, an iterative procedure for finding it was developed in Dauby and Hardy (2012).

Hurlbert (1971) developed an unbiased estimator of ${}^{k}H$ for all k smaller than the sample size. This is similar to what is needed to estimate generalized Simpson's entropy, although, generalized Simpson's entropy also needs $k < S$ for it to be a measure of diversity. We estimate Hurlbert's index for the two plots, convert them into effective numbers of species, and use bootstrap to get a 95% confidence envelop. The results are given in Figure 3b. We see that the maximum effective numbers of species are well below those of the generalized Simpson's diversity. Thus Hurlbert's diversity finds fewer rare species, making it a less interesting alternative for our purpose.

5. Conclusion

Generalized Simpson's entropy is a measure of diversity respecting the classical axioms. Further, there is a simple formula to transform it into an effective number of species. It faces issues that limit its use: namely, for it to respect the axiom of evenness, it order must be smaller that the number of species in the population. On the other hand, it has a decisive advantage over other metrics: it has an easy to calculate uniformly minimum variance unbiased estimator, which is consistent and asymptotically normal. These properties make it a useful tool for estimating diversity and allows us to robustly compare hyper-diverse, poorly sampled communities.

R code to reproduce the examples in the paper, based on the packages *EntropyEstimation* and *entropart* (Marcon and Hérault, 2015), is given in Appendix 2.

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

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

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

Burnham KP, Overton WS (1979). "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals." *Ecology*, **60**(5), 927–936.

Cao L, Grabchak M (2014). "EntropyEstimation: Estimation of Entropy and Related Quantities." URL http://cran.r-project.org/package=EntropyEstimation.

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

Chao A, Wang YT, Jost L (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.

Dalton H (1920). "The measurement of the inequality of incomes." *The Economic Journal*, **30**(119), 348–361.

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

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

Faddeev DK (1956). "On the concept of entropy of a finite probabilistic scheme." *Uspekhi Mat. Nauk*, **1**(67), 227–231.

Gourlet-Fleury S, Guehl JM, Laroussinie O (2004). *Ecology & Management of a Neotropical Rainforest. Lessons Drawnfrom Paracou, a Long-Term Experimental Research Site in French Guiana*. Elsevier, Paris, France.

Gregorius HR (1991). "On the concept of effective number." *Theoretical population biology*, **40**(2), 269–83.

Gregorius HR (2014). "Partitioning of diversity : the "within communities" component." *Web Ecology*, **14**, 51–60.

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

Hill MO (1973). "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology*, **54**(2), 427–432.

Hurlbert SH (1971). "The Nonconcept of Species Diversity: A Critique and Alternative Parameters." *Ecology*, **52**(4), 577–586.

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

Marcon E (2015). "Practical Estimation of Diversity from Abundance Data." *HAL*, **01212435**(version 2).

Marcon E, Hérault B (2015). "entropart, an R Package to Partition Diversity." *Journal of Statistical Software*, **67**(8), 1–26.

Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014). "Generalization of the Partitioning of Shannon Diversity." *Plos One*, **9**(3), e90289. doi:doi:10.1371/journal.pone.0090289.

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

R Development Core Team (2016). "R: A Language and Environment for Statistical Computing." URL http://www.r-project.org.

Rényi A (1961). "On Measures of Entropy and Information." In J Neyman (ed.), *4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561. University of California Press, Berkeley, USA.

Ricotta C (2005). "Through the jungle of biological diversity." *Acta Biotheoretica*, **53**(1), 29–38.

Shannon CE (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, **27**, 379–423, 623–656.

Simpson EH (1949). "Measurement of diversity." *Nature*, **163**(4148), 688.

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.

Zhang Z, Grabchak M (2014). "Entropic Representation and Estimation of Diversity Indices." *arXiv*, **1403.3031**(v. 2), 1–12.

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

6. Appendix 1: Proofs

The proof of Proposition 1 is given first. The gradient and Hessian of the generalized Simpson's entropy are then calculated and the proof of the satisfaction of the evenness axiom is given. Finally, we explain where the confidence intervals come from.

Several of the proofs require understanding the properties of the function $g(p) = p(1-p)^r$ and its first derivative. For $p \in [0,1)$ the first two derivatives of g are

$$g'(p) = (1-p)^r - rp(1-p)^{r-1} \tag{9}$$

and

$$g''(p) = r\left[(r+1)\,p - 2\right](1-p)^{r-2}. \tag{10}$$

Lemma 1. *1. For $p \in [0,1]$ we have $g'(p) \geq 0$ if and only if $p \in \left[0, \frac{1}{r+1}\right]$.*
2. The function g' is strictly decreasing for $p \in \left[0, \frac{2}{r+1}\right)$ and strictly increasing for $p \in \left(\frac{2}{r+1}, 1\right)$.
3. The function g' is nonincreasing for $p \in \left[0, \frac{2}{r+1}\right]$.

The information in this Lemma is summarized in Table 1.

Proof. The first part follows from the fact that $g'(p) \geq 0$ holds if and only if $(1-p)^r \geq rp(1-p)^{r-1}$, which holds if and only if $p \in \left[0, \frac{1}{r+1}\right]$. For the second part we need to characterize when $g''(p)$ is positive and when it is negative. Since, for $p \in [0,1)$, $r(1-p)^{r-2} > 0$, it follows that $g'(p)$ is strictly decreasing when $[(r+1)\,p - 2] < 0$, which holds if and only if $p \in \left[0, \frac{2}{r+1}\right)$. Similarly it is strictly increasing if and only if $[(r+1)\,p - 2] > 0$, which holds when $p \in \left(\frac{2}{r+1}, 1\right)$. The proof of the third part is similar to that of the second part. □

6.1 Proof of Proposition 1

Proof. A differentiable trace-form entropy satisfies the principle of transfers so long as $g'(p)$ is decreasing (Patil and Taillie, 1982, Theorem 4.2, with a typo: read $V'(\pi_j) \geq V'(\pi_i)$). From here the result follows by Lemma 1. □

6.2 Gradient

Generalized Simpson's entropy is given by

$$\zeta_r = \sum_{s=1}^{S} p_s(1-p_s)^r, \quad r = 1,2,\ldots. \tag{11}$$

Table 1. Variation table of the function g

p	0	$\frac{1}{S}$	$\frac{1}{r+1}$	$\frac{2}{r+1}$		1
$g''(p)$		$-$	0	$+$		
$g'(p)$	1	$\searrow$	0	$\nearrow$		0
			$-\left(\frac{r-1}{r+1}\right)^{r-1}$			

Since $\sum_{s=1}^{S} p_s = 1$, it can be written as a function of all probabilities but the last as

$$f(p_1, p_2, \ldots, p_{S-1}) = \sum_{s=1}^{S-1} p_s (1-p_s)^r + \left(1 - \sum_{s=1}^{S-1} p_s\right) \left(\sum_{s=1}^{S-1} p_s\right)^r.$$

The gradient of f is the vector $\left(\frac{\partial f}{\partial p_1}, \frac{\partial f}{\partial p_2}, \ldots, \frac{\partial f}{\partial p_{S-1}}\right)$, where for $u = 1, 2, \ldots, (S-1)$

$$\frac{\partial f}{\partial p_u} = (1-p_u)^r - p_u r (1-p_u)^{r-1} - \left(\sum_{s=1}^{S-1} p_s\right)^r + \left(1 - \sum_{s=1}^{S-1} p_s\right) r \left(\sum_{s=1}^{S-1} p_s\right)^{r-1}. \quad (12)$$

6.3 Hessian

The Hessian of f is the $(S-1) \times (S-1)$ matrix with $\frac{\partial^2 f}{\partial p_v \partial p_u}$ in position (u, v), where for $v \neq u$

$$\frac{\partial^2 f}{\partial p_v \partial p_u} = -2r \left(\sum_{s=1}^{S-1} p_s\right)^{r-1} + \left(1 - \sum_{s=1}^{S-1} p_s\right) r(r-1) \left(\sum_{s=1}^{S-1} p_s\right)^{r-2}$$

and

$$\frac{\partial^2 f}{\partial p_u^2} = -2r(1-p_u)^{r-1} + p_u r(r-1)(1-p_s)^{r-2} - 2r \left(\sum_{s=1}^{S-1} p_s\right)^{r-1} + \left(1 - \sum_{s=1}^{S-1} p_s\right) r(r-1) \left(\sum_{s=1}^{S-1} p_s\right)^{r-2}.$$

6.4 Extremum when all probabilities are equal

Proposition 3. *When all probabilities are equal, the generalized Simpson's entropy reaches a local maximum if $r+1 < 2S$ and a local minimum if $r+1 > 2S$.*

Proof. When $p_s = \frac{1}{S}$ for each $s = 1, 2, \ldots, S$

$$\frac{\partial f}{\partial p_s} = \left(\frac{S-1}{S}\right)^r - \frac{1}{S} r \left(\frac{S-1}{S}\right)^{r-1} - \left(\frac{S-1}{S}\right)^r + \frac{1}{S} r \left(\frac{S-1}{S}\right)^{r-1} = 0,$$

which means that the gradient is zero and this is a critical point. At this point the Hessian contains terms

$$\frac{\partial^2 f}{\partial p_u^2} = \frac{r}{S} \left(\frac{S-1}{S}\right)^{r-2} 2[r - 2S + 1] \quad (13)$$

and for $v \neq u$

$$\frac{\partial^2 f}{\partial p_v \partial p_u} = -2r \left(\frac{S-1}{S}\right)^{r-1} + \left(\frac{1}{S}\right) r(r-1) \left(\frac{S-1}{S}\right)^{r-2} = \frac{r}{S} \left(\frac{S-1}{S}\right)^{r-2} [r - 2S + 1]. \quad (14)$$

Denote $h(S, r) = \frac{r}{S} \left(\frac{S-1}{S}\right)^{r-2} 2[r - 2S + 1]$. The Hessian matrix is

$$\mathbf{H} = h(S, r) \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{pmatrix}. \quad (15)$$

It is easy to check that the matrx $\mathbf{H}/h(S, r)$ is positive definite. Thus $\mathbf{H}$ is positive definite if $h(S, r) > 0$ and negative definite if $h(S, r) < 0$. The sign of $h(S, r)$ is that of $r - 2S + 1$. Thus, by the second derivative test, when all probabilities are equal

- f reaches a local maximum if $r + 1 < 2S$;
- f reaches a local minimum if $r + 1 > 2S$.

This completes the proof. □

6.5 Maximum of the function

The proof of Proposition 2 follows immediately from the following.

Proposition 4. *Let $r \leq S - 1$.*
1. The global maximum of ζ_r is reached when all proportions are equal.
2. There are no other local maxima.

Proof. From (12) it follows that

$$\frac{\partial f}{\partial p_u} = g'(p_u) - g'\left(1 - \sum_{s=1}^{S-1} p_s\right). \quad (16)$$

Thus, when the gradient equals zero it means that, for every $u = 1, 2, \ldots, (S-1)$, we have

$$g'(p_u) = g'\left(1 - \sum_{s=1}^{S-1} p_s\right),$$

which implies that

$$g'(p_1) = g'(p_2) = \cdots = g'(p_S). \tag{17}$$

To guarantee that $\sum_{s=1}^{S} p_s = 1$ there must be at least one $u \in \{1, 2, \ldots, S\}$ with $p_u \leq \frac{1}{S}$. The assumption that $r \leq (S-1)$ implies that $p_u \leq \frac{1}{r+1}$. Combining this with Lemma 1 implies that $g'(p_u) \geq 0$. Combining this with (17) implies that $g'(p_s) \geq 0$ for each s. By Lemma 1 this means that $p_s \leq \frac{1}{r+1}$ for each s. Since, by Lemma 1, g' is strictly decreasing on $\left(0, \frac{2}{r+1}\right)$ it follows that $p_1 = p_2 = \cdots = p_S = 1/S$. Thus the only critical point is at the uniform distribution. By Proposition 3 this is a local maximum, hence it is the global maximum as well. □

Remark 1. *In summary, Propositions 3 and 4 imply that*

- *When $r \leq S-1$, ζ_r has a global maximum at the uniform distribution.*
- *When $S \leq r \leq 2S-1$, ζ_r has a local maximum at the uniform distribution.*
- *When $r \geq 2S$, ζ_r has a local minimum at the uniform distribution.*

This leaves the question of whether, in the case $S \leq r \leq 2S-1$, the local maximum at the uniform distribution needs to be a global maximum. In general it does not. To illustrate this we consider the simple case where $p_1 = p$ and $p_u = \frac{1-p}{S-1}$ for $u = 2, 3, \ldots, S$. In this case

$$\zeta_r = u(p) = p(1-p)^r + (1-p)\left(1 - \frac{1-p}{S-1}\right)^r.$$

In Figure 4 a plot of u is given for $S = 10$ and $r = 15 \in [S, 2S-1]$. The plot show that, in this case, the global maximum is not at the uniform distribution.

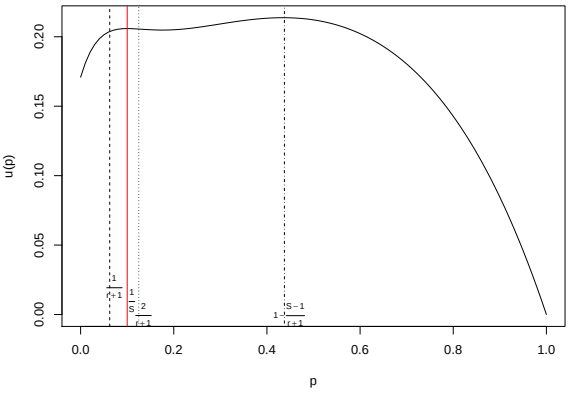


Figure 4. Plot of $u(p)$ for $S = 10$ and $r = 15$

6.6 Confidence intervals

The confidence intervals in Sections 2.4 and 2.5 follow immediately from the following result.

Proposition 5. *If $r \leq S-1$ and there exists an s with $p_s \neq 1/S$ then*

$$\sqrt{n}\frac{Z_v - \zeta_v}{\hat{\sigma}_v} \xrightarrow{L} N(0,1) \text{ as } n \to \infty. \tag{18}$$

Proof. In Zhang and Grabchak (2014) it was shown that (18) holds for any r and any $(p_1, p_2, \ldots, p_S)$ for which the gradient is not zero. Proposition 3 and the proof of Proposition 4 imply that this always holds under the given conditions. □

7. Appendix 2: R code

```
library("entropart")
library("EntropyEstimation")

############## General settings ########################
# Generalized Simpson function
q.seq <- c(seq(0, .1, .025), seq(.15, .65, .05), seq(.7, 2, .1))
NumberOfSimulations <- 1000
Alpha <- 0.05
########################################################



############## Plot P6 and P18, data ########################
# Get data
data(Paracou618)
NsP6 <- as.AbdVector(Paracou618.MC$Nsi[, 1])
NsP18 <- as.AbdVector(Paracou618.MC$Nsi[, 2])
# Richness
(Richness(NsP6, Correction = "None"))
(S6 <- Richness(NsP6, Correction="Jackknife"))
(Richness(NsP18, Correction = "None"))
(S18 <- Richness(NsP18, Correction="Jackknife"))
S <- min(S6, S18)



############## Plot P6 and P18, HCDT ########################
# Calculate HCDT diversity profiles
D6 <- CommunityProfile(Diversity, NsP6, NumberOfSimulations=NumberOfSimulations,
        q.seq=q.seq, Correction="UnveilJ")
D18 <- CommunityProfile(Diversity, NsP18, NumberOfSimulations=NumberOfSimulations,
        q.seq=q.seq, Correction="UnveilJ")

# Plot both profiles
plot(D18$x, D18$y, ylim=c(min(D18$low)*.9, max(D18$high)*1.05), main="",
     xlab = "Order of Diversity", ylab = "Diversity", xlim=range(q.seq), type="n")
CEnvelope(D6, LineWidth=2, main="", xlim=range(q.seq), ShadeColor=NA, col="darkgreen")
lines(D6$x, D6$high, col="darkgreen", lty=1)
lines(D6$x, D6$low, col="darkgreen", lty=1)
CEnvelope(D18, lty=2, LineWidth=2, col="red", BorderColor="red", ShadeColor="NA")


############## Plot P6 and P18, zeta ########################
# Gen Simpson profile, plot 6
zeta6 <- CommunityProfile(GenSimpson, NsP6, 1:(S-1))
sigma6 <- sapply(1:(S-1), function(r) GenSimp.sd(NsP6,r))
ic6 <- qnorm(1-Alpha/2)*sigma6/sqrt(sum(NsP6))
zeta6$low <- zeta6$y - ic6
zeta6$high <- zeta6$y + ic6
# Gen Simpson profile, plot 18
zeta18 <- CommunityProfile(GenSimpson, NsP18, 1:(S-1))
sigma18 <- sapply(1:(S-1), function(r) GenSimp.sd(NsP18,r))
ic18 <- qnorm(1-Alpha/2)*sigma18/sqrt(sum(NsP18))
```

```
zeta18$low <- zeta18$y - ic18
zeta18$high <- zeta18$y + ic18

# Plot both profiles
plot(zeta18$x, zeta18$y, ylim=c(min(zeta6$low)*.9, 1), main="", xlab = "Order of Diversity",
     ylab = "Generalized Simpson Entropy", xlim=c(1, S-1), type="n")
CEnvelope(zeta6, LineWidth=2, main="", xlim=c(1, S-1), ShadeColor=NA, col="darkgreen")
lines(zeta6$x, zeta6$high, col="darkgreen", lty=1)
lines(zeta6$x, zeta6$low, col="darkgreen", lty=1)
CEnvelope(zeta18, lty=2, LineWidth=2, col="red", BorderColor="red", ShadeColor="NA")


############## Plot P6 and P18, zeta Diversity ###########
# Transform entropy into diversity
zeta6D <- zeta6
zeta6D$y <- 1/(1-(zeta6$y)^(1/zeta6$x))
zeta6D$low <- 1/(1-(zeta6$low)^(1/zeta6$x))
zeta6D$high <- 1/(1-(zeta6$high)^(1/zeta6$x))
zeta18D <- zeta18
zeta18D$y <- 1/(1-(zeta18$y)^(1/zeta18$x))
zeta18D$low <- 1/(1-(zeta18$low)^(1/zeta18$x))
zeta18D$high <- 1/(1-(zeta18$high)^(1/zeta18$x))
# Plot both profiles
plot(zeta18D$x, zeta18D$y, ylim=c(min(zeta6D$low)*.9, max(zeta18D$high)*1.05), main="",
     xlab = "Order of Diversity", ylab = "Generalized Simpson Diversity",
     xlim=c(1, S-1), type="n")
CEnvelope(zeta6D, LineWidth=2, main="", xlim=c(1, S-1), ShadeColor=NA, col="darkgreen")
lines(zeta6D$x, zeta6D$high, col="darkgreen", lty=1)
lines(zeta6D$x, zeta6D$low, col="darkgreen", lty=1)
CEnvelope(zeta18D, lty=2, LineWidth=2, col="red", BorderColor="red", ShadeColor="NA")


############## Plot difference of entropy ########################
# Calculate P18-P6 with CI
Difference <- list(x=zeta18$x, y=zeta18$y-zeta6$y)
class(Difference) <- class(zeta18)
icDifference <- qnorm(1-Alpha/2)*sqrt(sigma6^2/sum(NsP6) + sigma18^2/sum(NsP18))
Difference$low <- Difference$y - icDifference
Difference$high <- Difference$y + icDifference
# Plot
plot(Difference, ylab="Difference of entropy")
abline(h=0, col="blue", lty=2)


############## Hurlbert ########################
# Hurlbert
N6 <- sum(NsP6)
N18 <- sum(NsP18)
N <- min(N6, N18)
Dk6 <- CommunityProfile(Hurlbert, NsP6, NumberOfSimulations=NumberOfSimulations, q.seq=2:N)
Dk18 <- CommunityProfile(Hurlbert, NsP18, NumberOfSimulations=NumberOfSimulations, q.seq=2:N)

# Plot both profiles
plot(Dk18$x, Dk18$y, ylim=c(min(Dk18$low)*.9, max(Dk18$high)*1.05), main="",
     xlab = "Order of Diversity", ylab = "Hurlbert Diversity", xlim=c(2,N), type="n")
```

```
CEnvelope(Dk6, LineWidth=2, main="", xlim=c(2,N), ShadeColor=NA, col="darkgreen")
lines(Dk6$x, Dk6$high, col="darkgreen", lty=1)
lines(Dk6$x, Dk6$low, col="darkgreen", lty=1)
CEnvelope(Dk18, lty=2, LineWidth=2, col="red", BorderColor="red", ShadeColor="NA")


############## Equivalence of orders ###########
# Min q value at max r
(qMin6 <- D6$x[which.min(D6$y > max(zeta6D$y))])
qOK6 <- which(D6$y < max(zeta6D$y))
(qMin18 <- D18$x[which.min(D18$y > max(zeta18D$y))])
qOK18 <- which(D18$y < max(zeta18D$y))

# Extend P18 zeta
zeta18e <- CommunityProfile(GenSimpson, NsP18, 1:(S18-1))
sigma18e <- sapply(1:(S18-1), function(r) GenSimp.sd(NsP18,r))
ic18e <- qnorm(1-Alpha/2)*sigma18e/sqrt(sum(NsP18))
zeta18e$low <- zeta18e$y - ic18e
zeta18e$high <- zeta18e$y + ic18e
# diversity
zeta18eD <- zeta18e
zeta18eD$y <- 1/(1-(zeta18e$y)^(1/zeta18e$x))
zeta18eD$low <- 1/(1-(zeta18e$low)^(1/zeta18e$x))
zeta18eD$high <- 1/(1-(zeta18e$high)^(1/zeta18e$x))
plot(zeta18eD)
# Min q value at max r
(qMin18e <- D18$x[which.min(D18$y > max(zeta18eD$y))])
qOK18e <- which(D18$y < max(zeta18eD$y))



############## Appendix proofs ##########################
# univariate function u
u <- function(p, S, r) {
  return( (1-p)*(1-(1-p)/(S-1))^r + p*(1-p)^r )
}
# graphical parameters
rshift <- .01
tcex <- .7
# parameters
S <- 10
r <- 15
curve(u(x, S=S, r=r), from=0, to=1, xlab="p", ylab="u(p)")
abline(v=1/S, col="red")
text(x = 1/S +rshift, y = 0.01, labels=expression(frac(1,S)), cex=tcex)
abline(v=1/(r+1), lty=2)
text(x = 1/(r+1) +rshift, y = 0.02, labels=expression(frac(1,r+1)), cex=tcex)
abline(v=2/(r+1), lty=3)
text(x = 2/(r+1) +rshift, y = 0, labels=expression(frac(2,r+1)), cex=tcex)
abline(v=1-(S-1)/(r+1), lty=4)
text(x = 1-(S-1)/(r+1) +rshift, y = 0, labels=expression(1-frac(S-1,r+1)), cex=tcex)
```