Oikos 000: 000–000, 2011
doi: 10.1111/j.1600-0706.2011.19267.x

Subject Editor: Karin Johst. Accepted 20 June 2011

The decomposition of Shannon's entropy and a confidence interval for beta diversity

Eric Marcon, Bruno Hérault, Christopher Baraloto and Gabriel Lang

E. Marcon (eric.marcon@ecofog.gf), AgroParisTech, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – B. Hérault, Univ. des Antilles et de la Guyane, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – C. Baraloto, INRA, UMR EcoFoG, BP 709, FR-97310 Kourou, French Guiana. – G. Lang, AgroParisTech, UMR 518 Math. Info. Appli., 16 rue Claude Bernard, FR-75005 Paris, France.

Beta diversity is among the most employed theoretical concepts in ecology and biodiversity conservation. Up to date, a self-contained definition of it, with no reference to alpha and gamma diversity, has never been proposed. Using Kullback-Leibler divergence, we present the explicit formula of Shannon's β entropy, a bias correction for its estimator and a confidence interval. We also provide the mathematical framework to decompose Shannon diversity into several hierarchical nested levels. From botanical inventories of tropical forest plots in French Guiana, we estimate Shannon diversity at the plot, forest and regional level. We believe this is a complete and usefulness toolbox for ecologists interested in partitioning biodiversity.

Alpha, beta and gamma diversities are among the most employed theoretical concepts in ecology and biodiversity conservation. For most ecologists, alpha diversity traditionally reflects the within-habitat diversity (MacArthur 1965), whereas beta diversity is the component of 'total diversity' that is produced by differences in species composition among the sampling units, i.e. 'the extent of change of community composition' (Whittaker 1960). The need to partition diversity within and among habitats has both theoretical (e.g. gradient analyses) and applied (e.g. defining protected areas) consequences such that these concepts are widely employed in both ecology (Crist et al. 2003) and conservation biology (Steinitz et al. 2005).

The partitioning of diversity began with ecological niche studies (Allan 1975), but recent interest has focused 1) in partitioning biodiversity measures into independent components (Jost 2007, Pélissier and Couteron 2007, Jost et al. 2009) and 2) in analyzing patterns of diversity sampled from hierarchically scaled studies (Lande 1996, Loreau 2000). This recent interest largely builds on Lande's (1996) explanations for additive partitions of total diversity (gamma) into components within-samples (alpha) and among-samples (beta), following thus the original concepts of alpha, beta and gamma diversity (Whittaker 1972), even though Lande's framework was correct only for Shannon diversity (Jost 2006, 2007). Up to now, diversity partitioning approaches have been used over a wide range of ecosystems, including tropical (Condit et al. 2002) as well as temperate landscapes (Qian et al. 2005). Although diversity partitioning has deep conceptual meanings for ecologists, its usefulness to date has been impeded by 1) the lack of theoretical basis for interpreting the results (Jurasinski et al. 2009) and 2) the lack of statistical methods for testing null hypotheses (Crist et al. 2003).

As a good starting point, it has been acknowledged that most usual measures of diversity are particular cases of generalized entropy measures (Tsallis 1988). In this framework, the species richness is an entropy-based diversity measure of order 0, the Shannon diversity (Shannon 1948, Shannon and Weaver 1963) of order 1 and the Simpson (1949) of order 2 (Jost 2007). Decreasing the order of the diversity estimates is conceptually equivalent to enhancing the weight of rare species in the final diversity estimates (Keylock 2005). The Shannon estimates (order 1) are ecologically meaningful because all species are strictly weighted by their frequency. Furthermore, Jost (2007) shows that Shannon's estimate is the only common entropy measure that can be decomposed additively, so that the gamma entropy H_γ is the weighted sum of α entropy of partitions, H_α, and a between-partition entropy H_β, even when community weights are unequal. Transforming Shannon entropies into Hill numbers makes possible the derivation of the so-called 'true diversity' indices (Jost 2006, Tuomisto 2010), i.e. the number of equally-abundant elements needed to produce a given value of Shannon entropy. In the remainder, we will write 'entropy' for classical measures, and 'diversity' for their Hill numbers, to keep a consistent terminology.

Finally, Shannon measures have several intuitively expected properties of a diversity measure (Jost 2007): 1) alpha and beta components are mathematically independent, meaning that a high value of alpha does not force the beta component to be high and vice versa, 2) gamma

is completely determined by alpha and beta, and 3) alpha is never greater than gamma. These properties of Shannon measures, 1) and 2) shared by all Rényi's (1961) measures of entropy but not 3) when community weights are unequal, give them a privileged place as estimates of diversity. But, as far as we know, still lacking is an explicit mathematical formulation of H_β, whose value is always obtained by the difference $H_\gamma - H_\alpha$. Recently, Tuomisto (2010) provided an extremely detailed review of literature defining 'beta diversity as a function of alpha and gamma diversity' and, from this review, it emerges that a self-contained definition of H_β, with no reference to H_α and H_γ, has never been proposed.

The aim of this paper is to give the explicit mathematical formulation of a dataset measure of biodiversity (γ) into within- (α) and between-(β) partition diversities following Whittaker's concepts (1972).

The paper is organized as follows. First, we derive the decomposition of Shannon's entropy H using Kullback-Leibler divergence, yielding the analytic formulation of H_β. Then, we show how to compute its confidence interval and we derive its bias correction. We use simulated datasets to show properties of H_β when data are samples from the same community. We use a real dataset to show how Shannon's diversity can be decomposed and how biases can be corrected, including hierarchical nested levels of decomposition: after spliting gamma diversity into alpha and beta components, alpha diversity of each community can itself be further considered as a gamma diversity for sub-communities and decomposed.

Methods

Derivation of the decomposition

Kullback-Leibler divergence

A Kullback-Leibler (1951) divergence, now synonymous of 'relative entropy' although Kullback and Leibler did not mention entropy, measures how different two distributions of probabilities are. Given a model distribution **p**, actual frequencies **q** of a finite number of observations i, the Kullback-Leibler divergence T between **p** and **q** is:

$$T = \sum_i q_i \ln \frac{q_i}{p_i} \tag{1}$$

Economists know this measure T as Theil's (1967) dissimilarity index. Note that T is a measure of divergence *s.s.*, neither a dissimilarity nor a distance measure, because **p** and **q** do not play a symmetric role. We also note that, theoretically, the observed values in a given system are only estimates of the actual frequencies **q**. Unlike Mori et al. (2005), we interchange for simplicity **q** and $\hat{\mathbf{q}}$, without consequences for our purpose here.

Consider an ecological community partitioned into plots. The number of individuals of each species s in each plot i is denoted n_{si}. The number of individuals in plot i is $n_{+i} = \Sigma_s n_{si}$, the number of individuals of species s is $n_{s+} = \Sigma_i n_{si}$. The total number of individuals is n_{++}. The corresponding actual frequencies are $q_{si} = {}^{n_{si}}/_{n_{++}}$, with $\Sigma_i \Sigma_s q_{si} = 1$. The expected distribution will be $p_{si} = {}^{n_{+i}}/_{Sn_{++}}$ where S is the number of species. In other words, we expect that all species have the same frequency, and the number of individuals is proportional to the size of the plot.

Grouping rule

In this section, we derive a general Eq. 5 for grouping plots or species. Similar approaches are common in the economic literature (Bickenbach and Bode 2008), so we will follow its terminology.

Data are organized in a table where lines are species, indexed by s and columns are plots indexed by i. Consider any group of cells G: the contribution of the group to the total relative entropy is the sum of each cell's relative entropy. We denote it T_G^α:

$$T_G^\alpha = \sum_{g\in G} q_g \ln \frac{q_g}{p_g} \tag{2}$$

After grouping, a single cell remains. Its relative entropy is the between-group relative entropy, we denote it T_G^γ:

$$T_G^\gamma = q_G \ln \frac{q_G}{p_G} = \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g} \tag{3}$$

Proof: the probability for an individual to belong to the group is the sum of the probabilities that it belongs to any cell of the group.

The within-group relative entropy is:

$$T_G^\beta = \sum_{g\in G} \frac{q_g}{\sum_{g\in G} q_g} \ln \frac{\frac{q_g}{\sum_{g\in G} q_g}}{\frac{p_g}{\sum_{g\in G} p_g}} = \left(\sum_{g\in G} q_g\right)^{-1} \times \left[\sum_{g\in G} q_g \ln \frac{q_g}{p_g} - \left(\sum_{g\in G} q_g\right) \ln \frac{\sum_{g\in G} q_g}{\sum_{g\in G} p_g}\right] \tag{4}$$

Proof: within the group, the sum of probabilities is 1. Within-group probabilities are therefore normalized.

Finally, the total relative entropy of the group equals its between-group plus its within-group relative entropy:

$$T_G^\alpha = T_G^\gamma + \left(\sum_{g\in G} q_g\right) T_G^\beta \tag{5}$$

At this step, alpha, beta and gamma are purely conventional notations. They will be justified later.

Application to Shannon's index

We apply the previous result to Shannon's index of diversity. The expected probability for species s in plot i is 1/S (all species are expected to have the same frequency) multiplied by ${}^{n_{+i}}/_{n}$, the weight of plot i. The observed frequency is $q_{si} = {}^{n_{si}}/_{n_{++}}$. We group all the cells of species s. The relative entropy of the dataset for species s is:

$$T_s^{\alpha} = \sum_i q_{si} \ln \frac{q_{si}}{p_{si}} = \sum_i \frac{n_{si}}{n_{++}} \left(\ln \frac{n_{si}}{n_{+i}} + \ln S \right) \quad (6)$$

The gamma relative entropy of species s is:

$$T_s^{\gamma} = q_{s+} \ln \frac{q_{s+}}{p_{s+}} = \frac{n_{s+}}{n_{++}} \left(\ln \frac{n_{s+}}{n_{++}} + \ln S \right) \quad (7)$$

The between-plot relative entropy of species s is:

$$\left(\sum_i q_{si} \right) T_s^{\beta} = \left(\sum_i q_{si} \right) \sum_i \frac{q_{si}}{q_{s+}} \ln \frac{\frac{q_{si}}{q_{s+}}}{\frac{p_{si}}{p_{s+}}} = \frac{n_{s+}}{n_{++}} \sum_i \frac{n_{si}}{n_{s+}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} = \sum_i \frac{n_{si}}{n_{++}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} \quad (8)$$

We know Eq. 5 that $T_s^{\alpha} = T_s^{\gamma} + (\sum_i q_{si}) T_s^{\beta}$. This equality will be summed over all species to introduce diversity measures:

$$T_{\alpha} = \sum_s T_s^{\alpha} = \ln S + \sum_i \frac{n_{+i}}{n_{++}} \sum \frac{n_{+i}}{n_{++}} \ln \frac{n_{si}}{n_{+i}} = \ln S - \sum_i \frac{n_{+i}}{n_{++}} H_i^{\alpha} = \ln S - H_{\alpha} \quad (9)$$

H_i^{α} is the alpha diversity of plot i. It is computed according to local frequencies n_{si}/n_{+i}. H_{α} is the weighted sum of H_i^{α}. T_{α} is the Kullback-Leibler divergence between **p** and **q** for all plots and all species.

The gamma relative entropy sums to give the Kullback-Leibler divergence for the dataset:

$$T_{\gamma} = \sum_s T_s^{\gamma} = \ln S + \sum_s \frac{n_{s+}}{n_{++}} \ln \frac{n_{s+}}{n_{++}} = \ln S - H_{\gamma} \quad (10)$$

Finally, we sum between-plot relative entropy:

$$T_{\beta} = \sum_s \left(\sum_i q_{si} \right) T_s^{\beta} = \sum_i \sum_s \frac{n_{si}}{n_{++}} \ln \frac{\frac{n_{si}}{n_{s+}}}{\frac{n_{+i}}{n_{++}}} = \sum_i \frac{n_{+i}}{n_{++}} \sum_s \frac{n_{si}}{n_{+i}} \ln \frac{\frac{n_{si}}{n_{+i}}}{\frac{n_{s+}}{n_{++}}} \quad (11)$$

Combining Eq. 9, 10 and 11 and assuming $H_{\gamma} = H_{\alpha} + H_{\beta}$, we identify diversity:

$$H_{\beta} = \sum_i \frac{n_{+i}}{n_{++}} H_i^{\beta} = \sum_i \frac{n_{+i}}{n_{++}} \sum_s \frac{n_{si}}{n_{+i}} \frac{\frac{n_{si}}{n_{+i}}}{\frac{n_{s+}}{n_{++}}} \quad (12)$$

H_{β} is the weighted sum of contributions of plots i, H_i^{β}. These contributions are Kullback-Leibler divergences. The expected probabilities are $p_{si} = n_{s+}/n_{++}$. The probability to find an individual of species s in plot i is proportional to the frequency of the species in the dataset. All plots are expected to be identical. Observed frequencies are $q_{si} = n_{si}/n_{+i}$: actual frequencies differ from plot to plot. In agreement with intuition, diversity is the divergence between identical plots and real plots.

Hill (1973) numbers are the numbers of equiprobable species yielding the same measure of diversity as the actual data, also called the effective number of species. They allow one to transform non-intuitive values of Shannon diversity into easy-to-understand numbers. The Hill number for β diversity is the number of equally-weighted, completely distinct plots giving the same value of H_{β}, that is to say the effective number of plots.

Confidence intervals

We want to have a confidence interval of H_{β}: plot data are samples of wider communities so observed values of H_{β} may vary due to sampling stochasticity. The confidence interval is computed by Monte-Carlo simulations assuming the species distribution of plots and resampling them. First, we draw each value of n_{si} in a binomial law $B(n_{+i}, n_{si}/n_{+i})$ and we calculate H_{β}. We then repeat the simulation a large number of times, (e.g. 10 000) and eliminate extreme values according to the chosen risk level α. For α = 5%, the confidence interval of the null hypothesis is between the 251st and the 9750th simulated values of H_{β}.

Sampling bias

Sampling bias occurs because the community under study contains an unknown number of species denoted $\tilde{S}$. Only S species have been observed in plots: some rare species have not been sampled, introducing a downward bias of H equal to $(\tilde{S}-1)/2n$ plus a negligible term, derived by Basharin (1959), that remains intractable. Chao and Shen (2003) built an unbiased estimator for H_{α} in plot i:

$$\tilde{H}_i^{\alpha} = \sum_s \frac{C_i \frac{n_{si}}{n_{+1}} \ln C_i \frac{n_{si}}{n_{+1}}}{1 - \left(1 - C_i \frac{n_{si}}{n_{+1}} \right)^n} \quad (13)$$

C_i is the estimator of the sample coverage (Good 1953), that is to say the proportion of observed species in terms of probabilities in plot i. We denote C the sample coverage for the whole dataset. A C equal to 90% means that unobserved species represent 10% of individuals of the community. The sample coverage is estimated by $1 - S_1/n$ where S_1 is the number of species observed only once (we will call them singletons) in the sample. Shannon's alpha or gamma entropy can be estimated easily without bias by simply taking into account singletons and the sample size.

We follow Chao et Shen to derive an unbiased estimator for $\tilde{H}_{\beta}$. The denominator was introduced by Horvitz and

Thompson (1952) to correct for unobserved species: it is equal to the probability to not sample each species s. Probability estimators are observed frequencies multiplied by the sample coverage so that they sum to 1 including unobserved species. These corrections are applied to H_i^{β}. The Horvitz-Thompson correction is the same as for $\tilde{H}_i^{\alpha}$. Frequencies in each plot are multiplied by the plot's coverage, while those of the whole dataset are multiplied by the whole dataset's coverage. We get:

$$\tilde{H}_i^{\beta} = \sum_s \frac{C_i \frac{n_{si}}{n_{+i}} \ln \frac{C_i \frac{n_{si}}{n_{+i}}}{C \frac{n_{s+}}{n++}}}{1-\left(1-C_i \frac{n_{si}}{n_{+i}}\right)^n} \quad (14)$$

In turn, when resampling plots (by drawing binomial laws as we do, or by bootstrapping), some rare species among the S observed are eliminated. Alpha and gamma diversities of simulated plots are thus systematically lower than those of the original ones. Alpha diversity is more biased than gamma diversity due to sample sizes. Beta diversity is thus overestimated. This can be corrected by applying Chao and Shen's bias correction, but remains incomplete because the unobserved number of species samples are drawn from S, not $\tilde{S}$. Finally, the unbiased simulated values are distributed around the biased actual ones. We will call them semi-unbiased values. No technique is available for a nested bias correction that would correct for the unobserved species among $\tilde{S}$.

A complete correction of the bias of simulations can be done numerically, for H_β as well as H_α and H_γ. For a given plot, the bias is constant as it only depends on the unobserved number of species and the sample size. As a result, the simulated variance of H is not affected: the biased simulations can be simply re-centered around the actual value of H.

Examples

We first used a simulated dataset to illustrate that the minimum value of H_β depends on both the number of species in the community and the sampling effort. We simulated two frequency distributions of respectively 20 and 40 species in plots. Simulated frequencies follow a uniform law. Then we drew a pair of plots 10 000 times from these communities, with an expectancy of 500 individuals or 5000 individuals. We computed H_β for each pair of plots to construct a frequency histogram of H_β, smoothed as a density function. No bias correction is needed here. H_β is not zero due to stochasticity although plots are samples of the same community.

We also provide a real example to show how deal with actual data. We measured Shannon diversity for tree communities in four 1-ha plots of tropical rain forest at the Nouragues and Paracou field stations in French Guiana. Both sites are seasonal lowland forests receiving about 3 m of annual precipitation, with tree composition dominated by *Fabaceae*, *Chrysobalanaceae*, *Lecythidaceae*, *Sapotaceae* and *Burseraceae* (Bongers et al. 2001, Gourlet-Fleury et al. 2005). The two plots within each site were chosen to represent the most common contrasting environments found for hilltop terra firme forest at each site. At Paracou, the two example plots occur on migmatite associated with the Bonidoro geological formation, with one plot exhibiting blocked vertical drainage (P06) and the other with strong vertical drainage and incipient podzolization (P18). At Nouragues, the two example plots occur on weathered granite with sandy soils (NH20) and metavolcanic rock of the Paramaca formation with clay-rich laterite soils. In all four plots all trees were sampled in 2008 by professional climbers to obtain herbarium vouchers, each of which was identified to distinct morphospecies at the Cayenne regional herbarium (Baraloto et al. 2010). We assume for simplicity here that an acceptable sample of each forest site is obtained when its two plots are united.

Results

The simulations exemplified how H_β values depend on the number of individuals sampled. H_β does not change if all numbers of individuals are multiplied by 10, while maintaining actual frequencies the same. But frequencies vary randomly and, as a divergence, H_β accumulates these fluctuations. Its expected value is not 0 even when plots are from the same community. When more individuals are drawn, species frequencies converge to their probability due to the law of large numbers, so H_β converges to 0. In a 20-species community, an observed value $H_\beta = 0.005$ shows a significant difference between plots if it is obtained from two 5000-tree plots (Fig. 1, solid curve on the left). If the plots

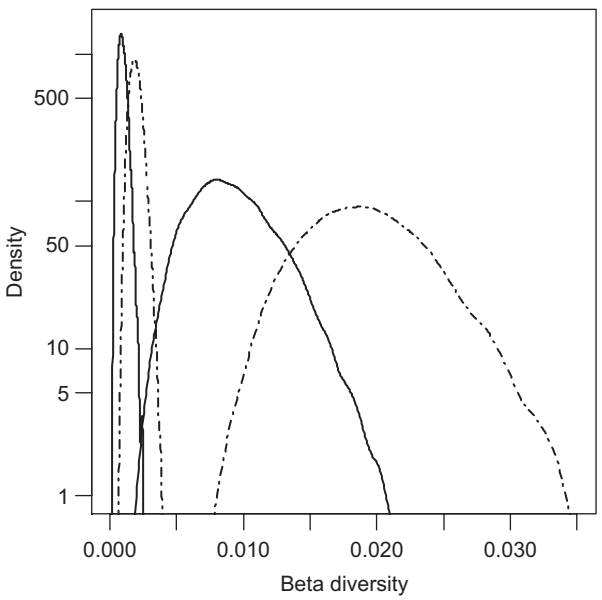

Figure 1. Probability densities of H_β obtained from 10 000 simulations of the model described in details in the text. Two plots are drawn from the same community. H_β is not zero because of stochastic differences between the plots. The first two curves on the left concern plots around 5000 individuals, the right ones plots around 500 individuals. Dotted lines are for 40-species plots, solid lines for 20 species. Everything else equal, expected H_βs decrease with the number of individuals and increase with the number of species.

Table 1. An example of hierarchical decomposition of Shannon entropy (H) for tropical tree communities. Trees with diameter at breast height > 10 cm were inventoried in four 1-ha tropical rain forest plots in French Guiana. The first two plots (NH20, NL11) are from the Nouragues forest station, the last two from the Paracou (P006, P018) forest station. Within forests, the weighted sum of alpha ($\widetilde{H}_{plots}$) and beta entropies ($\widetilde{H}_{\beta}^{plots}$) equals the within forest gamma entropy ($\widetilde{H}_{forest}$). This within forest gamma entropy can be considered as the alpha entropy at the between-forest level. In this way, adding $\widetilde{H}_{forest}$ to the beta entropy between forests ($\widetilde{H}_{\beta}^{forest}$) gives the total entropy. Hill numbers are the numbers of equiprobable species or completely different plots or forests yielding the same measure of diversity as the actual data. Beta diversities are given with their 95% confidence interval between square brackets. All values are bias corrected.

	NH20	NL11	P006	P018
No. of trees	558	515	643	481
No. of observed species S	203	182	147	149
Total no. of species $\widetilde{S}$ estimated by Jacknife1	321	279	215	223
$\widetilde{H}_{plots}$	5.01	4.92	4.40	4.67
Hill no. (true plot alpha diversity)	151	137	82	107
$\widetilde{H}_{\beta}^{plots}$ with 95% CI	0.33 [0.30;0.36]		0.36 [0.33;0.39]	
Hill no. plots (true beta diversity)	1.39 [1.35;1.43]		1.44 [1.39;1.48]	
$\widetilde{H}_{forest}$	5.31		4.90	
Hill no. (true forest diversity)	201		134	
$\widetilde{H}_{\beta}^{forest}$ with 95% CI	0.33 [0.31;0.35]			
Hill no. forests (true beta diversity)	1.39 [1.36;1.41]			
$\widetilde{H}_{total}$	5.44			
Hill no. total (true gamma diversity)	230			

contain only 500 trees (Fig. 1, solid curve on the right), the same value is no longer significant. H_{β} also tends to be higher when the number of species increases for the same sample size: less individuals per species mean more stochasticity because relative entropy is calculated per species (Eq. 8) and summed.

The second example illustrates how Shannon diversity can be hierarchically partitioned (Table 1) and how to deal with biases. We first followed Beck and Schwanghart (2010) to validate the possibility to correctly estimate diversity. The total number of species $\widetilde{S}$ was estimated according to Brose et al.'s (2003) framework to evaluate sample completeness, that is to say the proportion in numbers of observed species ($S/\widetilde{S}$), which is lower than sample coverage, the proportion in probabilities. Jacknife1 appeared to be the appropriate estimator, and completeness was around 2/3 in all plots. This is enough to validate bias correction of Shannon indices according to Beck and Schwanghart's empirical findings.

The first result is that plots at the Nouragues site are more diverse than those at Paracou. Hill numbers offer an intuitive representation of the level of diversity. For example, the Nouragues NH20 plot is as diverse as one of the same size with 151 equally frequent species, almost twice the value obtained at Paracou P006. Next, plots can be grouped into forests, i.e. Nouragues and Paracou. The value of H_{forest}, that is to say the γ diversity of the forest, is the sum of the weighted H_{α} of plots plus the H_{β} between plots. In turn, H_{forest} can be treated as an α diversity and one can reiterate the same procedure. Successive values of H are given in Table 1.

The confidence interval of H_{β} is shown. For example, concerning the Nouragues plots, H_{β} is estimated at 0.33, corresponding to a Hill number equal to 1.39 plots (95% confidence interval between 1.35 and 1.43). Note that theoretical Hill values for this example of two plots are between 1 (perfect equality of distribution, $H_{\beta} = 0$) and 2 (equal number of trees with no species in common, $H_{\beta} = \ln 2 \approx 0.7$). We can see that diversity within forests is roughly the same as that between forests (all values of Hill Numbers are around 1.4) and all values are highly significant (the probability to have $H_{\beta} = 0$ is so low that this can be considered as impossible).

We could have chosen to group the four sampled plots directly. In this case, H_{β} between all plots would be 0.67. The corresponding Hill number would be 1.95 (confidence interval between 1.89 and 2.01) meaning that the four plots are almost equivalent to two completely different ones.

Discussion

Previous works

In this paper, we propose a self-contained definition of H_{β}, with no reference to H_{α} and H_{γ}. The form of H_{β} we provide has already been derived by Ricotta and Avena (2003), but they did not relate it with H_{α} and H_{γ}. Also, Ludovisi and Taticchi (2006) decomposed a Kullback-Leibler divergence with a different approach, in order to develop new measures of β diversity.

Interpretation and properties of H_{β}

Kullback-Leibler divergences provide the necessary framework to decompose Shannon diversity. Shannon's α diversity is the difference between the logarithm of the number of species and Theil's relative entropy, that is to say the Kullback-Leibler divergence between a distribution where all species have the same frequency and actual data. Shannon's γ diversity has the same definition after grouping plots. Shannon's β diversity is the Kullback-Leibler divergence between actual plots and identical ones.

In the proposed diversity partitioning framework, H_{β} is very different from H_{α} and H_{γ} because it is a measure of divergence, not of diversity itself. More similar species relative abundances result in higher H_{α} and H_{γ}, but H_{β} increases when plots are less similar. This is in agreement with the original definition of diversity expressed by Whittaker (1960). Converting Shannon's entropy to Hill numbers allows a unified definition of diversity as a number of effective objects

(species or plots), or 'true diversity' measures (thoroughly discussed by Tuomisto 2010, p. 8).

The maximum theoretical value of H_β is lnS when all plots have an α diversity equal to zero (i.e. they contain a single species that is different among plots); and γ diversity also has its maximum value, equal to lnS. This is possible only if the number of samples equals the number of species and the number of individuals in every sample is the same. A more realistic situation is H_β equal to the logarithm of the number of samples. In this case, samples contain the same number of individuals (equal weight) but have no species in common. This is a special case of the maximum value derived by Jost (2007, Eq. 21), equal to the Shannon entropy of the weights of plots when they have no species in common. This is why observed β diversity makes sense only when compared to the weighted number of plots (Jost 2007 proposed to normalize it to the unit interval): β diversity between Nouragues and Paracou in our example is 1.39 for two samples, interpretable as a marked difference between communities. If it were 1.39 for 10 samples, it would mean that they are almost similar.

The minimum value is 0 when all plots are completely identical in species relative abundances. This never happens if they are random samples of the same community (Fig. 1). So the confidence interval cannot be used as a test for community equality because it never contains 0. We do not provide such a test, following Jones and Matloff (1986) for example, because increasing sample size always allows to make it significant as, unlike models of our theoretical examples, real communities are never exactly identical.

The estimator $\widetilde{H}_\beta$ may be negative if bias correction is erroneous. The simplest example is given by two or more exactly identical plots with singletons. Biased H_β is 0 so $\widetilde{H}_\beta < 0$. Bias correction assumes that data are random samples so that singletons allow estimating unobserved species, but this artificial example violates these assumptions.

A user's guide

In summary, we propose a complete procedure to analyze data. The R (R Development Core Team 2010) code we wrote and used can be found as a supplementary material for use in further studies to compute unbiased values of H and confidence intervals.

The first step consists in evaluating the completeness of each plot in the sense of Beck and Schwanghart (2010) following Brose et al.'s (2003, Fig. 6) framework. Note that the term 'coverage' was replaced by 'completeness' by Beck and Schwanghart (2010) to avoid confusion with coverage defined by Good (1953). If completeness is above 50%, bias correction is very efficient such that Shannon's diversity can be estimated as follows:

- Unbiased $\widetilde{H}_i^\alpha$ are computed according to Eq. (13) in each plot i. $\widetilde{H}_\alpha$ is the weighted sum of plot diversities: $\widetilde{H}_\alpha = \sum_i n_{+i}/n_{++}\widetilde{H}_i^\alpha$.
- Unbiased $\widetilde{H}_\beta$ is obtained the same way, from Eq. (14).
- The confidence interval of $\widetilde{H}_\beta$ is computed by simulating plots, calculating their semi-unbiased beta diversity and re-centering the distribution around $\widetilde{H}_\beta$.
- $\widetilde{H}_\gamma$ is calculated according to Eq. (13) after grouping data.
- Finally, all entropy values should be transformed into true diversities (their exponential) to allow interpretation.

Conclusion

In this paper, we provided the explicit formula of Shannon's β diversity and the mathematical framework to justify it. We showed that Shannon's β diversity is the Kullback-Leibler divergence between actual plots and the average plot. We also explained how to calculate the interval confidence of H_β. As real data are almost always incomplete (i.e. some rare species have not been sampled), we provided bias correction for the estimator of H_β. Finally, we showed how to decompose Shannon diversity into several nested levels. All results are interpretable intuitively after transformation into Hill Numbers.

This diversity partitioning is flexible enough to analyze any a priori determinant of species diversity. We believe that the use of explicit diversity partitioning will help both ecologists to understand the factors shaping the spatial and temporal distribution of biodiversity and nature practitioners to design effective strategies for protecting biodiversity (Veech et al. 2002).

Acknowledgements – We wish to sincerely thank Lou Jost whose suggestions and demands made us improve the quality of the paper very significantly.

References

Allan, J. D. 1975. Components of diversity. – Oecologia 18: 359–367.

Baraloto, C. et al. 2010. Functional trait variation and sampling strategies in species rich plant communities. – Funct. Ecol. 24: 208–216.

Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. – Theor. Probabil. Appl. 4: 333–336.

Beck, J. and Schwanghart, W. 2010. Comparing measures of species diversity from incomplete inventories: an update. – Meth. Ecol. Evol. 1: 38–44.

Bickenbach, F. and Bode, E. 2008. Disproportionality measures of concentration, specialization, and localization. – Int. Regional Sci. Rev. 31: 359–388.

Bongers, F. et al. (eds) 2001. Nouragues: dynamics and plant-animal interactions in a neotropical rainforest. – Kluwer.

Brose, U. et al. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. – Ecology 84: 2364–2377.

Chao, A. and Shen, T. J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. – Environ. Ecol. Stat. 10: 429–443.

Condit, R. et al. 2002. Beta-diversity in tropical forest trees. – Science 295: 666–669.

Crist, T. O. et al. 2003. Partitioning species diversity across landscapes and regions: a hierarchical analysis of alpha, beta, and gamma diversity. – Am. Nat. 162: 734–743.

Good, I. J. 1953. On the population frequency of species and the estimation of population parameters. – Biometrika 40: 237–264.

Gourlet-Fleury, S. et al. 2005. Using models to predict recovery and assess tree species vulnerability in logged tropical forests: a case study from French Guiana. – For. Ecol. Manage. 209: 69–86.
Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – Ecology 54: 427–432.
Horvitz, D. G. and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. – J. Am. Stat. Ass. 47: 663–685.
Jones, D. and Matloff, N. 1986. Statistical hypothesis testing in biology: a contradiction in terms. – J. Econ. Entomol. 79: 1156–1160.
Jost, L. 2006. Entropy and diversity. – Oikos 113: 363–375.
Jost, L. 2007. Partitioning diversity into independent alpha and beta components. – Ecology 88: 2427–2439.
Jost, L. et al. 2009. Partitioning diversity for conservation analyses. – Divers. Distrib. 16: 65–76.
Jurasinski, G. et al. 2009. Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. – Oecologia 159: 15–26.
Keylock, C. J. 2005. Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. – Oikos 109: 203–207.
Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. – Ann. Math. Stat. 22: 79–85.
Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – Oikos 76: 5–13.
Loreau, M. 2000. Are communities saturated? On the relationship between alpha, beta and gamma diversity. – Ecol. Lett. 3: 73–76.
Ludovisi, A. and Taticchi, M. I. 2006. Investigating beta diversity by Kullback-Leibler information measures. – Ecol. Modell. 192: 299–313.
MacArthur, R. H. 1965. Patterns of species diversity. – Biol. Rev. 40: 510–533.
Mori, T. et al. 2005. A divergence statistic for industrial localization. – Rev. Econ. Stat. 87: 635–651.
Pélissier, R. and Couteron, P. 2007. An operational, additive framework for species diversity partitioning and beta-diversity analysis. – J. Ecol. 95: 294–300.
Qian, H. et al. 2005. Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. – Ecol. Lett. 8: 15–22.
Rényi, A. 1961. On masures of entropy and information. – In: Neyman, J. (ed.), 4th Berkeley Symp. Math. Stat. Probabil. Univ. of California Press, pp. 547–561.
Ricotta, C. and Avena, G. 2003. An information-theoretical measure of β-diversity. – Plant Biosyst. 137: 57–61.
Shannon, C. E. 1948. A mathematical theory of communication. – Bell System Tech. J. 27: 379–423, 623–656.
Shannon, C. E. and Weaver, W. 1963. The mathematical theory of communication. – Univ. of Illinois Press.
Simpson, E. H. 1949. Measurement of diversity. – Nature 163: 688.
Steinitz, O. et al. 2005. Predicting regional patterns of similarity in species composition for conservation planning. – Conserv. Biol. 19: 1978–1988.
Theil, H. 1967. Economics and information theory. – Rand McNally and Co.
Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. – J. Stat. Phys. 52: 479–487.
Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. – Ecography 33: 2–22.
Veech, J. A. et al. 2002. The additive partitioning of species diversity: recent revival of an old idea. – Oikos 99: 3–9.
Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. – Ecol. Monogr. 30: 279–338.
Whittaker, R. H. 1972. Evolution and measurement of species diversity. – Taxon 21: 213–251.