Ecology, 98(5), 2017, pp. 1444–1454

Estimating species richness in hyper-diverse large tree communities

Hans ter Steege,[1,2,3,7] Daniel Sabatier,[4] Sylvia Mota de Oliveira,[1] William E. Magnusson,[5] Jean-François Molino,[4] Vitor F. Gomes,[2] Edwin T. Pos,[1,6] and Rafael P. Salomão[2]

[1] *Naturalis Biodiversity Center, Leiden, The Netherlands*
[2] *Museu Paraense Emílio Goeldi, Belem, Para, Brazil*
[3] *Systems Ecology, Free University Amsterdam, Amsterdam, The Netherlands*
[4] *AMAP, IRD, Cirad, CNRS, INRA, Université de Montpellier, Montpellier, France*
[5] *Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brazil*
[6] *Ecology & Biodiversity Group, Department of Biology, Utrecht University, Utrecht, The Netherlands*

Abstract. Species richness estimation is one of the most widely used analyses carried out by ecologists, and nonparametric estimators are probably the most used techniques to carry out such estimations. We tested the assumptions and results of nonparametric estimators and those of a logseries approach to species richness estimation for simulated tropical forests and five data sets from the field. We conclude that nonparametric estimators are not suitable to estimate species richness in tropical forests, where sampling intensity is usually low and richness is high, because the assumptions of the methods do not meet the sampling strategy used in most studies. The logseries, while also requiring substantial sampling, is much more effective in estimating species richness than commonly used nonparametric estimators, and its assumptions better match the way field data is being collected.

Key words: Amazon; logseries; nonparametric estimators; species estimation; species richness; tropical forests.

Introduction

Species-richness estimation is one of the most widely used analyses carried out by ecologists, either to compare samples obtained with different efforts, or by extrapolation, to predict the number of species present in an area larger than the one sampled. Extrapolation methods are frequently used for geographically large areas, where coverage of the complete range is out of reach, too labor intensive, or too expensive.

Parametric species-richness estimation is based on parameter inference for either one of the two main relationships describing assemblages: the number of individuals (N) in a community or the area (A) this community occupies. In these cases, the number of species (S) only depends on the relative or rank abundance distribution of the species (RAD; Izsák and Pavoine 2012) or the species–area relationship (SAR; Rosenzweig 1995). As a general rule of thumb, in any number of random samples of an area, the number of species that remain undetected will increase with increased S and A (Gotelli and Colwell 2001), precluding any attempt to directly quantify the RAD or the SAR from samples. This clearly poses a problem in tropical forests that are generally both large and rich.

There has been a long argument as to whether the logseries (Fisher et al. 1943), the log-normal (Preston 1948), or alternative distributions (McGill et al. 2007) give the best fit for rank abundance distributions (RADs), how much the fit is dependent on scale or sampling completeness, and to which extent the best fitting model reflects the biological processes underlying the distribution. The use of nonparametric estimators of species richness, such as Chao, ICE (incidence-based coverage estimator of species richness), and jackknifing, has been proposed as a way of dealing with this uncertainty, because they do not assume any underlying distribution. It would be wrong, however, to suppose that they are less sensitive to other assumptions than parametric methods or that they do not suffer from other drawbacks. Brose et al. (2003) noted that sampling-theoretical methods of estimation require high sampling intensity to avoid what Wang and Lindsay (2005) call the "severe under-estimation observed from popular nonparametric estimators due to the interplay of inadequate sampling effort, large heterogeneity and skewness." Xu et al. (2012) also reported that nonparametric methods severely underestimate richness and emphasized that these methods should not be used across heterogeneous landscapes. This is largely because nonparametric estimators based on a sampling estimate of the rare tail of the SAR are very sensitive to the shape of the abundance distribution. As underlined by Harte and Kitzes (2015), "The rare tail is emphasized because the shape of the species–area relationship is especially influenced by the numbers of rare species." Although the performance of estimators has been frequently compared (Brose et al. 2003, Chiarucci et al. 2003, Walther and Moore 2005, Hortal et al. 2006, Xu et al. 2012), much less of the ecological literature critically evaluates their assumptions and caveats.

Manuscript received 13 April 2016; revised 24 January 2017; accepted 3 March 2017. Corresponding Editor: María Uriarte.

[7] E-mail: hans.tersteege@naturalis.nl

Perhaps the most commonly used estimator for species richness is the Chao1 nonparametric estimator (Chao 1989, Chao et al. 2009), which estimates the number of species as:

$$S_{\text{estimated}} = S_{\text{observed}} + \frac{f_1^2}{2f_2}$$

where f_1 is the number of species with 1 individual in the sample (singletons) and f_2 is the number of species with 2 individuals in the sample (doubletons). The Chao1 estimator and other nonparametric estimators make no assumptions about the underlying species-abundance distribution, but do assume that sampling is random with replacement across the whole area. When $f_1 = 0$, it is assumed that all species have been collected and $S_{\text{estimated}} = S_{\text{observed}}$ (Chao et al. 2009).

Chao Bunge (Chao and Bunge 2002), Chao Lee ACE, and Chao Lee ACEI (Chao and Lee 1992), and jackknife (Burnham and Overton 1979), which are variations on the original Chao1 estimator, are also dependent on the fractions of the rare or infrequent species, and require "a sufficiently high overlap fraction [..] to produce a reliable estimate of the species" (Chao and Bunge 2002), and all are based on the capture–recapture principle that requires sampling with replacement.

In contrast, the logseries is not based on a capture–recapture principle and was among the first attempts to mathematically describe the relationship between the number of species and number of individuals in a biological context (Fisher et al. 1943). It is given by

$$\Phi_n = \frac{\alpha x^n}{n}$$

where: Φ_n is the number of species with n individuals; α is Fisher's α; and $x = N/(N + \alpha)$ (N being the number of individuals in the total sample; x being asymptotically equal to 1 with large sample sizes). Hence, we expect α from samples to quickly approach α of the total landscape, after which it will be practically independent of sample size. Fisher's alpha can be calculated from the number of individuals (N) and species (S) in a sample by iteratively solving

$$\alpha = \frac{S}{\ln(1 + N/\alpha)}.$$

The logseries is essentially a geometric summation, which builds up from the first term (Φ_1), the singletons. The number of singletons is thus predictable in a logseries ($\Phi_1 = \alpha x$) and is always the largest class. As x is very close to 1 for reasonably large samples, $\Phi_1 \approx \alpha$ in such samples. Similarly, the number of doubletons is: $\Phi_2 = \alpha\, x^2/2 \approx \alpha/2$. When we assume that RADs of communities follow the logseries, this has implications for the nonparametric Chao1 estimator. For large samples, the Chao1 estimator (note that $f_1^2/[2f_2] = \Phi_1^2/[2\Phi_2]$) will simply become $S_{\text{estimated}} = S_{\text{observed}} + \alpha^2/[2(\alpha/2)] = S_{\text{observed}} + \alpha$. Consequently, we predict that, for reasonably large samples, for which α is constant, Chao1 always estimates the number of unseen species as α, regardless of the size of the samples.

Hubbell's neutral theory was the first ecological theory deriving the logseries from the basic biological processes of birth rate (b) and death rate (d) (Hubbell 2001, 2015). It can be shown that, in this model, $x = b/d$. Neutral theory (NT) derives a distribution, the zero sum multinomial (ZSM), which, for large communities with little drift, approaches a logseries. For small local communities (limited immigration and drift), the ZSM approaches a lognormal (Hubbell 2001).

Here we compare commonly used nonparametric estimators of species richness to one parametric estimator based on the logseries for the purpose of estimating species richness in large areas of tropical forest. We specifically chose the logseries as we are trying to estimate richness in very large areas where the ZSM approaches this distribution. We show by simulations and comparisons with empirical data that the assumptions of the parametric estimator are less sensitive to deviations than those of the nonparametric estimators.

Methods

Simulations

We modeled forest communities of 10,000 1-ha plots (a 100-km^2 area), each plot with 500 individuals. We initially filled each of the 10,000 ha with a random sample of 500 individuals from a metacommunity (MC). The MC was constructed using a logseries of 15 million individuals and a Fisher's α of 300, which is roughly equivalent to a rich central Amazonian rainforest (see *Field data*). We used a logseries as this conforms to the structure expected (Hubbell 2001) and found in tropical forests (Hubbell et al. 2008, ter Steege et al. 2013, Hubbell 2015). After filling the plots randomly from the MC, the mean Fisher's α of all plots and that of the virtual forest initially is, as expected, equivalent to that of the MC.

During the simulations, trees were randomly selected to be removed (1 per plot per time step) and new recruitment could come from dispersal (m) from four sources:

1. Recruitment from dispersal inside the plot (m_{plot}), equivalent to local recruitment. Local recruitment is random within the plot, i.e., we assume no spatial structure inside the plots.
2. Recruitment from dispersal from the surrounding eight plots. Dispersal probability based on dispersal distance was based on the model of Chisholm and Lichstein (2009), modified by Pos et al. (*in press*). The dispersal probability from the adjacent plots (m_{adjacent}) is computed from dispersal distance as (Pos et al., *in press*):

$$m_{\mathrm{adjacent}} = 0.3 \times \frac{A - (l - 2 \times d)^2}{A}$$

where: A is the area of the plot (10,000 m^2), l = length of the plot (100 m), and d = the average dispersal distance. Assuming an average dispersal range of 10–40 m m_{adjacent} is in the range of 0.108–0.288.

3. Recruitment from dispersal from the surrounding forest (10,000 ha), comparable to long-distance dispersal. Individuals for replacement were drawn randomly from the 10,000 ha. This assumes that long-distance dispersal is not spatially driven. We used a probability of $m_{\mathrm{forest}} = 0.1 \times m_{\mathrm{adjacent}}$.
4. Recruitment from dispersal from the MC, this is comparable to infrequent very long-distance dispersal, also termed vagrancy. The individuals were drawn randomly from the MC, assuming that very long-distance dispersal too is not spatially driven. We used a probability of $m_{\mathrm{MC}} = 0.01 \times m_{\mathrm{adjacent}}$.
5. Speciation (v) as defined in the Unified Neutral Theory of Biodiversity and Biogeography (Hubbell 2001):

$$v = \frac{\theta}{2 \times J} = \frac{250}{2 \times 10{,}000 \times 500} = 2.5e^{-5}$$

where θ is the biodiversity number, asymptotically equivalent to Fisher's alpha, and J is the size of the community.

Parameters 2–4 were calculated first. Local recruitment (1) was then calculated as $m_{\mathrm{plot}} = 1 - m_{\mathrm{adjacent}} - m_{\mathrm{forest}} - m_{\mathrm{MC}} - v$.

We ran 30,000 time steps for each model with mean dispersal distances of 10, 15, 20, 25, 30, and 40 m. At each time step, one individual per plot was randomly selected to be replaced by another individual based on the five probabilities above. Thus, 10,000 individuals were replaced at each time step.

After each simulation, we plotted the RAD with a fit of the logseries and lognormal, the Species Area Curve with Chao1 estimator, the Fisher's α to area curve, the exact richness of the simulated community, and the predicted richness based on Fisher's α and the Chao1 estimator. All curves were based on the average of 50 draws from 1 to all 10,000 plots. We also plotted the results for the average of 50 random draws of 100 plots from our virtual forest.

We also ran the simulation model for a sample of 49 ha of forest (7 × 7 ha), using the field data of BCI (Table 1). We simulated a forest area of 49 plots, using a MC of 15 km^2 (the size of BCI), an alpha of 50, a density of 429 individuals/ha, and a dispersal distance of 40 m (Chisholm and Lichstein 2009) for $m_{\mathrm{adj}} = 0.288$, and $v = 0.00119$.

Simulations and calculations were carried out with custom-made scripts in R (R Development Core Team 2011).

Field data

We used field data from the following four sites: (1) Barro Colorado Island (BCI), a 50-ha plot in old growth forest (Condit et al. 2002; this well-known data set was also used in Chao et al. (2009); (2) Reserva Ducke (RD; Appendix S1: Fig. S1), a forest reserve of 100 km^2 in central Amazonia, just north of Manaus (Castilho 2004); (3) Piste de St Elie (PSE; Appendix S1: Fig. S2), mixed forest in northern French Guiana (Sabatier et al. 1997); (4) the Monte Branco Plateau (MBP; Appendix S1: Fig. S3), a large bauxite plateau of 3750 ha in Para, Brazil (Salomão 2015).

BCI tree data was extracted from vegan (Oksanen et al. 2008), tree data for RD and PSE are integrated in the ATDN database (ter Steege et al. 2013) and extracted from that source, MBP tree data (R. P. Salomão, *unpublished data*) was taxonomically harmonized with the ATDN database.

We extrapolated the species richness for an area in which the plots were located; for RD for 7.2 million individuals (the area of the full 100-km^2 reserve); for PSE an imaginary 1500-ha forest area encompassing the plots; for MBP the 3750 ha that comprises the complete plateau (Table 1). The plots are well spread across these areas. For BCI we estimated richness for the 50-ha plot.

For each of the plot data sets, we carried out the following analyses:

1. Plotted the RAD of the data set with the exact logseries and lognormal for the number of individuals (N) and species (S) in the field sample;

TABLE 1. Botanical inventories used for the analysis.

Locality	No. plots	Plot area (ha)	N	S	Target area	Target individuals	Source
BCI	50	1.00	21,457	225	50 ha	21,457	1
RD	72	0.50	25,066	1,233	100 km^2	7,200,000	2
PSE	20	1.00	1,450	574	1,500 ha	933,750	3
MBP	301	0.25	36,546	703	3,750 ha	1,821,229	4

Note: Localities are Barro Colorado Island (BCI), Reserva Ducke (RD), Piste de St Elie (PSE), and Monte Branco Plateau (MBP). Variables are number of plots sampled, plot area, number of individuals sampled (N), number of species recorded (S), the target area for which estimates were made, and number of individuals in the target area based on average density. Data sources are 1, Condit et al. (2002); 2, Castilho (2004); 3, Sabatier et al. (1997); 4, Salomão (2015).

2. Constructed a curve of the mean species richness by area, based on 50 randomizations of the field data;
3. Constructed a curve of the mean of Fisher's α by area, based on the same 50 randomizations of the field data;
4. Estimated species richness in the target area for all subsamples of the 50 randomizations based on Fisher's α of the sub-samples as follows: $S = \alpha \times \ln(1 + N/\alpha)$ (Fisher et al. 1943); where α = Fisher's α, and N is the number of trees in the subsample and the variance of S as given by Fisher et al. (1943): $\mathrm{var}_S = \alpha \ln([2N + \alpha]/[N + \alpha]) - \alpha^2 N/(N + \alpha)^2$;
5. Estimated species richness in the target area for all sub-samples of the 50 randomizations, based on Chao1: $S_{\mathrm{est}} = S_{\mathrm{obs}} + f_1^2/(2f_2)$;
6. Estimated the species richness for the field data set for a number of nonparametric estimators (Chao 1984, Chao Bunge, Chao Lee ACE, Chao Lee ACEI, jackknife), as provided in the R package SPECIES (Wang 2011).

The 50 randomizations of the plot data were produced without replacement from one plot to the number of plots in the field data set.

Results

Simulations

The simulations of our virtual forest with mean dispersal distance of 20 m produced a RAD that is close to a logseries, but not fully identical (Fig. 1A). Species richness calculated with the Chao1 estimator, as predicted, becomes S_{observed} plus ~Fisher's α for larger samples (Fig. 1B). While Fisher's α and species richness calculated with Fisher's α tend to asymptotically approach the community value, species richness calculated with the Chao1 estimator follows the shape of the species–area curve and finally overestimates the richness of the total sample by approximately Fisher's α.

All simulations (d = 10–40 m) show similar results (Appendix S1: Figs. S4, S6, S8, S10, S12, S14, S16; Data S1: SPAR samples.csv). With increasing mean dispersal distance and, hence, stronger input from the adjacent plots, Fisher's α tends to be overestimated slightly before it reaches the value of the total virtual forest, and the number of species estimated to be in the full virtual forest increases from 2071 to 2098. The calculations for 50 samples of 100 plots suggest that, although Fisher's α

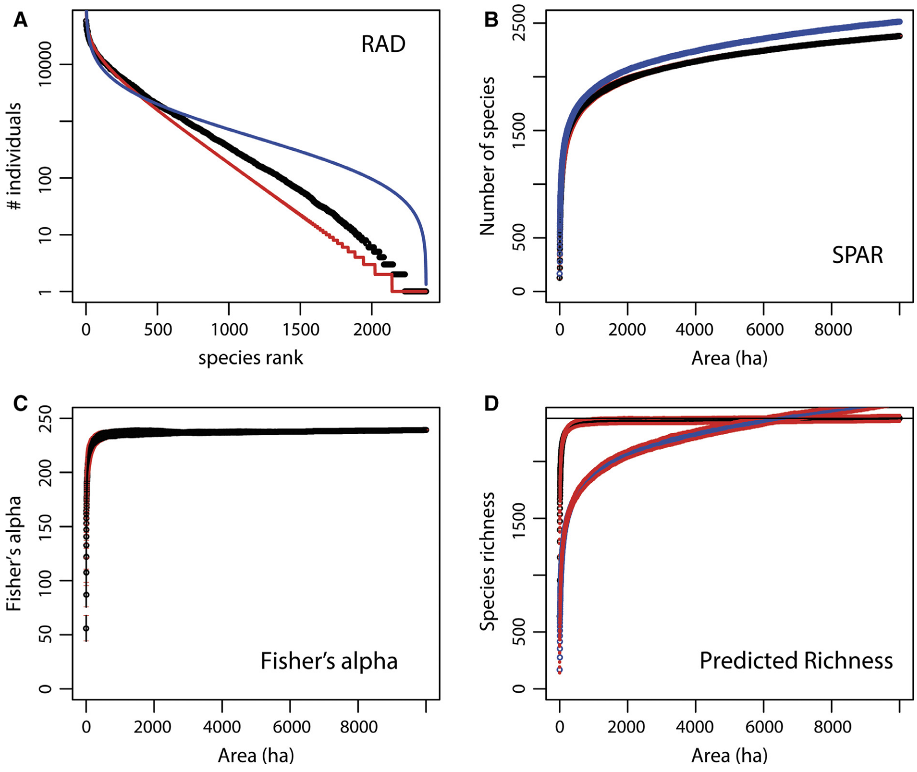

Fig. 1. Simulation of a 10,000-ha virtual forest with mean dispersal distance of 20 m. Parameters used are $m_{\mathrm{plot}} = 0.78688$; $m_{\mathrm{adjacent}} = 0.192$; $m_{\mathrm{forest}} = 0.0192$; $m_{\mathrm{MC}} = 0.00192$; $v = 10^{-4}$. (A) Rank abundance distribution (RAD) of the total virtual (black) with logseries fit (red) and lognormal fit (blue). (B) Species area (SPAR) curve for the total virtual forest and estimated richness ($S_{\mathrm{estimated}}$) based on Chao1 (blue). (C) Fisher's α area curve for the virtual forest. (D) Species richness estimated with Fisher's α (black), Chao1 (blue), each with 95% CI (red), and actual species richness of the simulated community (horizontal line). m_{plot} = local recruitment; m_{adjacent} = recruitment from adjacent plots; m_{forest} = recruitment from total forest; m_{MC} = recruitment from metacommunity; v = speciation.

predicts a richness closer to the known richness for the virtual forest, it is still an underestimate of 3–17% (Appendix S1: Figs. S5, S7, S9, S11, S13, S15; Data S1: sample by nr of plots.csv). For a similar sample size, the Chao1 estimator provides an underestimate of 43–51%, depending on the dispersal distance chosen (Data S1).

Simulations of 49 ha of BCI

Simulations of a 49-ha virtual plot based on the BCI data produced a RAD (Appendix S1: Fig. S18) very similar to that of the forest in the real 50-ha BCI plot (Fig. 2). Fisher's α was very close to the final value for the simulated forest after 10 plots. Consequently, species richness was also close to its simulated richness after sampling 10 plots. Species richness calculated with Chao1 is, as predicted, the species area curve plus Fisher's α of the sample. Thus, even when all individuals have been sampled, Chao1 still predicts unobserved species with a magnitude of Fisher's α. This is because, as in real forests, the virtual forest of 49 ha still contains singletons.

Field data

In all cases: BCI (Fig. 2), RD (Fig. 3), PSE (Fig. 4), and MBP (Fig. 5), the RAD showed a hollow curve with few common and many rare species and, except for BCI, the logseries provided a reasonable fit. In all cases, Fisher's α was very close to that of the full sample with less than 20 plots sampled. For small samples, Chao1 provided a severe underestimate for the richness in the sample, and even for the final sample, $S_{\text{estimated}}$ was almost equivalent to S_{observed} + Fisher's α.

Species estimates for the target area made with Fisher's α were much larger than those made with the asymptotic Chao1 estimator, which were close to S_{observed} + Fisher's α of the measured data (Figs. 2–5). All other nonparametric estimators also predict much lower values for richness, comparable to the Chao1 estimator (Table 2). Only for BCI, where the area for which richness was to be estimated was similar to the actual sample, did the nonparametric estimators approach the estimate based on Fisher's α.

For the BCI and MBP data, and simulations with higher mean dispersal distances, Fisher's α peaked before it leveled off to its final value similar to the simulations; i.e., it showed a hump (see Figs. 2, 5). Fisher's α, however, rose regularly for PSE, RD, and for simulations with lower mean dispersal distances (Figs. 1, 2, 4, 5).

Discussion

Based on our simulations with a spatially semi-explicit model, Fisher's α provides a more accurate prediction of

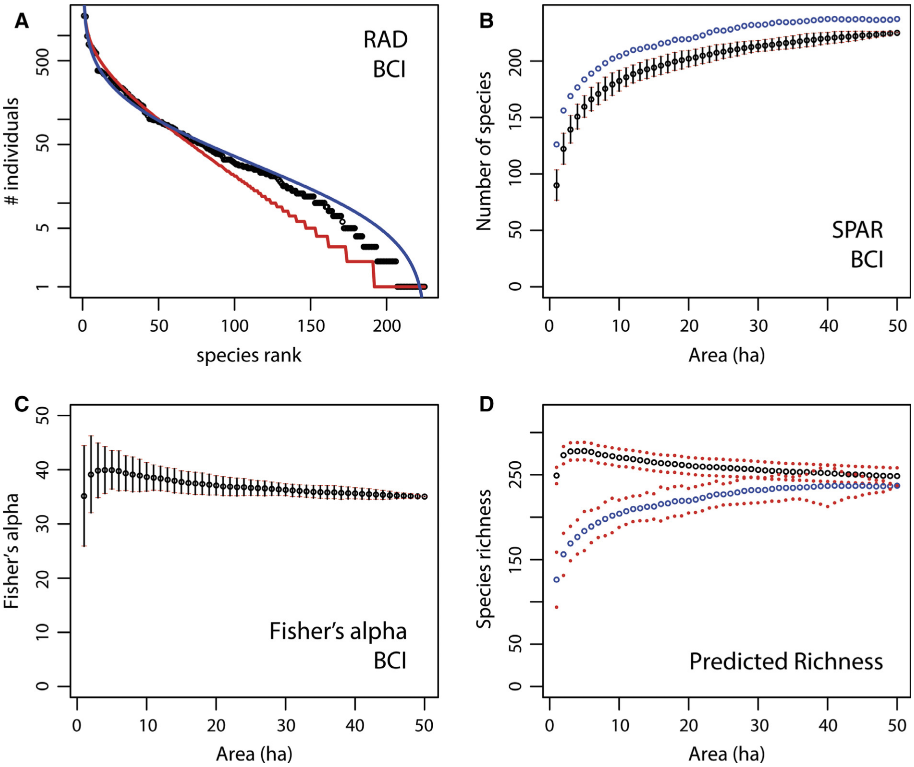

FIG. 2. Barro Colorado Island field data (BCI). (A) Rank abundance distribution (RAD) of BCI with logseries fit (red) and lognormal fit (blue). (B) Species area curve for BCI and estimated richness ($S_{\text{estimated}}$) based on Chao1 (blue). (C) Fisher's α area curve for BCI. (D) Species richness estimated for a 100-ha area on BCI with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

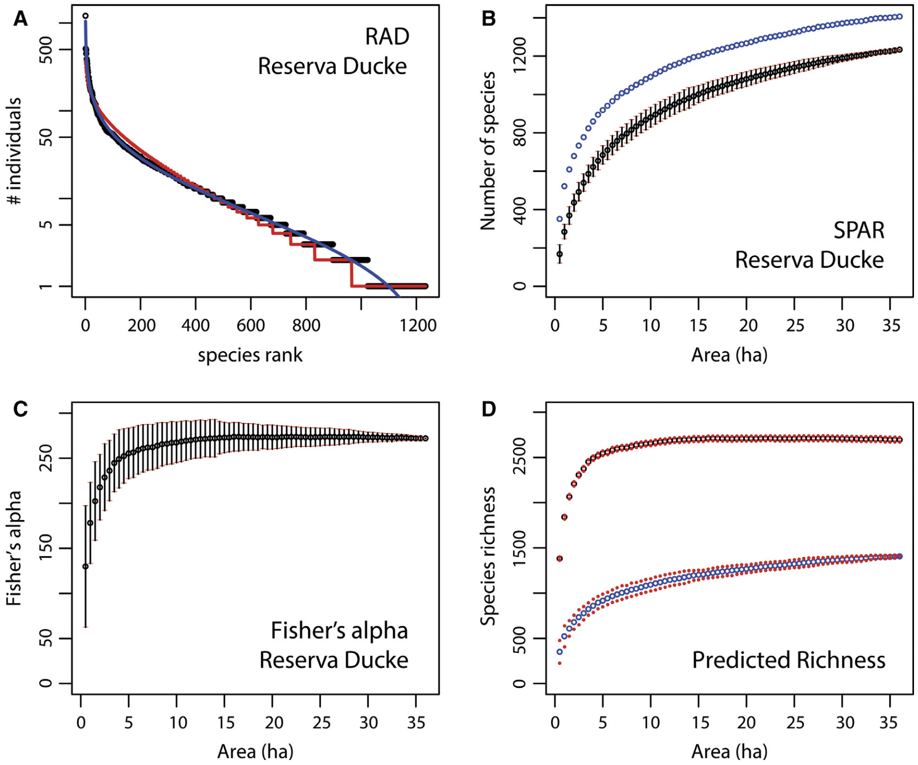

FIG. 3. Reserva Ducke field data (RD). (A) Rank abundance distribution (RAD) of RD with logseries fit (red) and lognormal fit (blue). (B) Species area curve for RD and estimated richness ($S_{\mathrm{estimated}}$) based on Chao1 (blue). (C) Fisher's α area curve for RD. (D) Species richness estimated for the total 100-km^2 RD area with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

species richness in the virtual-forest communities than does the nonparametric Chao1 and other nonparametric methods, especially if sample intensity is low. We believe that the failure of nonparametric methods to estimate diversity is mainly due to the resampling approach with its need of high sampling effort and its expected loss of singletons, and the lack of definition of the target area. We elaborate on this below.

Based on resampling the BCI plot data, Chao et al. (2009) found that, to detect 90% of the species, a median sample size of 80% of the area is necessary. Also Chiarucci et al. (2003), using modeled vegetation, found that nonparametric estimators need at least 15–30% of the area to be sampled for reasonable estimates of the species richness of the whole area. Using these methods with low sampling effort leads to serious underestimation, as Brose et al. (2003) and our models clearly show. In real life, even though trees are not removed by our sampling (and resampling is thus statistically possible), the chances of resampling the same plot are negligible. In the Amazon with a sample of 1170 1-ha plots in an area of over 5 million km^2 (ter Steege et al. 2013), that chance would be just 2×10^{-9}. At the intensities at which tropical forests are sampled (0.0002% for the Amazon) nonparametric methods simply cannot accurately estimate the number of species in the whole area. Also, when locations of previous studies are known, researchers are unlikely to resample a plot.

With capture and recapture techniques and the nonparametric estimators tested, sampling is considered complete when there are no singletons in the data (Chao et al. 2009). In tree plots, the disappearance of singletons would be the result of sampling the data many times over with replacement (Chao et al. 2009). This resampling results in the estimated richness asymptotically approaching true richness when the number of singletons is zero, as the total number of species cannot be larger than those observed in the total data set (Chao et al. 2009). We argued above that, in the case of research in tropical forests, plots are probably never sampled with replacement. Thus, the number of species is expected to increase with sample size as predicted by the "First Law of Biodiversity" (Rosenzweig 1995; "larger samples yield more species") and many other theories of biodiversity (MacArthur and Wilson 1967, Kimura 1985, Hubbell 2001, Harte et al. 2008, Harte 2011). In addition, singletons will remain (often close in number to Fisher's alpha). In the above theories, singletons are the representatives of the biological processes of immigration, extinction, or speciation. Singletons might be species on their way to extinction or new species coming in by speciation or migration. The latter are hence necessary to maintain

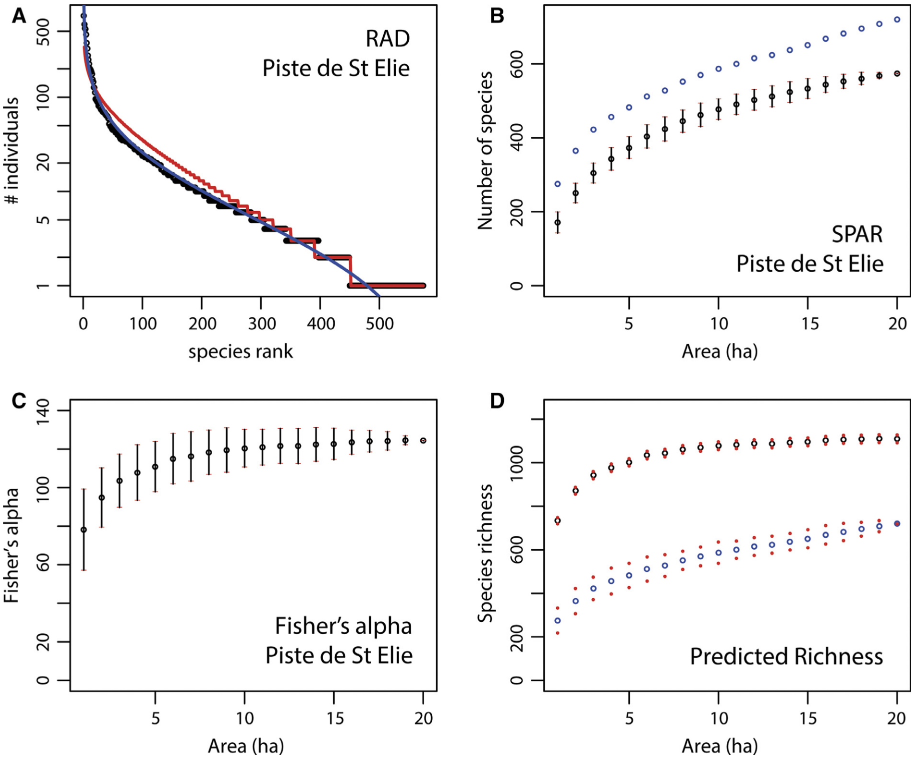

FIG. 4. Piste de Saint Elie field data (PSE). (A) Rank abundance distribution (RAD) of RD with logseries fit (red) and lognormal fit (blue). (B) Species area curve for RD and estimated richness ($S_{\text{estimated}}$) based on Chao1 (blue). (C) Fisher's α area curve for RD. (D) Species richness estimated for the total 15 km^2 area surrounding the plots with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

richness. Without these processes, fixation will occur due to ecological drift, analogous to genetic drift from population genetics. Thus, when sampling without replacement, the lack of singletons in these systems would suggest incomplete rather than complete sampling.

Finally, as most tropical tree field data conforms to the logseries (see references in *Introduction*), the Chao1 index becomes scale invariant, always estimating the same number of missing species to exactly the amount of Fisher's α. This was shown mathematically in the introduction for Chao1 and empirically in our simulations. While we did not show this mathematically for the other nonparametric estimators, they are derived from the same theoretical framework of capture-recapture and estimate similar richness (Appendix S1: Fig. S19; Table 2) and thus also provide severe underestimates with low sampling intensities.

For the full Amazon area (~5.5 million km^2), ter Steege et al. (2013) estimated ~16,000 tree species based on a sample of 1170 plots of 1 ha. They applied at least 18 different extrapolation methods from software packages SPECIES (Wang 2011), and CatchAll (Bunge et al. 2012) to their plot data (ter Steege et al. 2013). Almost all were rejected, as they predicted the total number of Amazonian tree species to fall in the range 4015–6412, a demonstrably severe underestimation of the true species richness (Fine 2001). A new estimator, implemented in CatchAll (WLRM_UnTransf; Rocchetti et al. 2011, Bunge et al. 2012) gave an estimated total richness above 11,000, closer to that calculated by ter Steege et al. (2013) with their logseries extrapolation, but was not selected by the program as the best estimator. The ACE1_Max tau estimator gave a result greatly exceeding the estimate with the log-series but its tau was much higher (9048) than the recommended value (tau < 10). The failure of these models to fit the Amazonian data is not surprising. These estimators performed poorly because at least one of their assumptions, high sampling intensity, was not met: a condition unlikely to be met in any large forested area. Recently, an extensive search in several data providers and herbaria showed that nearly 12,000 tree species have actually been collected in Amazonia, with a collecting density as low as 10 collections per 100 km^2 (ter Steege et al. 2016). The authors concluded that the estimate of 16,000 is entirely plausible. Importantly, even if this was an overestimate of the total number of species, the number of species already recorded is almost twice that estimated with most nonparametric methods.

Using different methods to estimate or extrapolate the SAR, such as maximum entropy inference (Harte 2011, Harte and Kitzes 2015) or a power-law based fitting

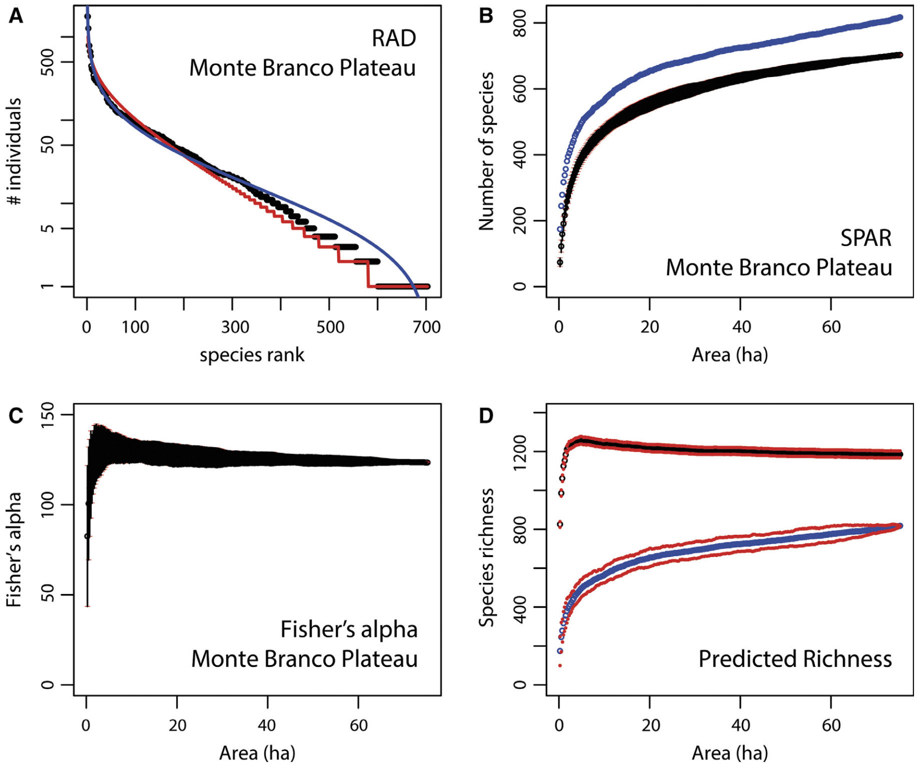

FIG. 5. Monte Branco Plateau field data (MBP). (A) Rank abundance distribution (RAD) of MBP with logseries fit (red) and lognormal fit (blue). (B) Species area curve for MBP and estimated richness ($S_{\text{estimated}}$) based on Chao1 (blue). (C) Fisher's α area curve for MBP. (D) Species richness estimated for the total 37.5 km^2 MBP area with Fisher's α (black) and Chao1 (blue), each with 95% CI (red).

from multi-scale sampling (Plotkin et al. 2000; Krishnamani et al. (2004), also showed that regional scale diversity of trees was estimated acceptably from small plots samples. Interestingly, the abundance distribution model arising from the MaxEnt approach is most often a logseries (Harte and Kitzes (2015).

Using the logseries is not without assumptions, however. Our virtual forest is neutral with regard to the environment; i.e., demographic probabilities for each individual, regardless of species identity, are equal. Hence, the only cause of aggregation is limited dispersal of individuals, but, given enough time, even ranges of very dispersal-limited species can become large. In real life, species will segregate the environment based on ecological preferences as well. Hence, beta-diversity in real forests is higher than in our virtual-forest stand and a peak of Fisher's α is expected when a large heterogeneous area is sampled over a range of sampling intensities.

TABLE 2. Species estimates based on plot samples in BCI, RD, PSE, and MBP.

Variable	BCI	SE	RD	SE	PSE	SE	MBP	SE
Number of plots	50		72		20		301	
Number of individuals	21,457		25,066		12,450		36,546	
Number of species	225		1,233		574		703	
Target area	50 ha		100 km^2		1,500 ha		3,750 ha	
Target individuals	21,457		6,960,000		933,750		1,821,229	
$S_{\text{estimated}}$ with								
Fisher's α	225		2,759		1,110		1,185	
Chao 1984	239	8.3	1,408	32	724	36	821	31
Chao Bunge	243	9.6	1,423	32	715	34	823	31
Chao Lee ACE	238	6.1	1,375	20	669	18	738	16
Chao Lee ACEI	241	8	1,405	26	694	25	805	23
Jackknife	244	6.1	1,591	59	1,066	124	920	40

BCI is known to have clear segregation of species based on soil moisture (Hubbell and Foster 1983) and the relationship between Fisher's α and area peaks at a relatively low number of plots. We also expect the species on MBP to be similarly clumped because of the clear peak in Fisher's α at low sample sizes. At MBP, plot size may also influence the peaking of Fisher's α. As the plots are smaller (0.25 ha), the recruitment to the plots will be more affected by the adjacent plots as m_{adjacent} is very much dependent on the ratio between the plot boundary and mean dispersal distance (Chisholm and Lichstein 2009). The modeled and observed peaks can be explained by a relationship between beta diversity and alpha diversity. At low migration rates, recruits mostly come from within plots; hence beta diversity is maximized but alpha diversity is not because each plot is practically isolated and losing species due to ecological drift. This means that, for just sampling one plot, Fisher's α will be much lower than the average of the whole forest. Continuous sampling, however, will gradually result in the average Fisher's α. There will be no peak because the probability for each plot bringing new species to the whole is the same and thus the increase will be gradual until Fisher's α is equal to that of the virtual forest. When migration increases, however, plots close by exchange more species and beta and local alpha diversities increase simultaneously. In this case, sampling a few plots randomly will likely initially overestimate Fisher's α, because each sample includes new species in the total sample due to the combined higher beta and alpha diversity, creating a fast rise in Fisher's α. Continued sampling adds more individuals to the total sample and, at some point, species will be resampled, lowering Fisher's α again. When dispersal is so high as to be similar across the complete virtual forest, composition would essentially be very similar for all plots with very high local alpha- and low beta-diversities and Fisher's α would not peak but increase quickly to its virtual-forest value, as in the virtual 49-ha BCI.

Is estimating species richness still a long way off?

Chiarucci (2012) suggested that "estimating species richness is still a long way off!" Nonparametric estimators underestimate richness ([Figs. 3, 4, 5; Table 2] and Xu et al. 2012), while area-based estimators tended to overestimate richness (Xu et al. 2012). Xu et al. (2012) concluded that Maxent greatly overestimated richness. However, their perceived overestimate was based on the richness they expected, which was based on a list of species found in their area. We believe that many of us do not fully comprehend the consequences of the logseries model. One of us was also surprised when we estimated the expected species for RD, which was much more than was expected based on extensive fieldwork for the flora of the area (Riberiro et al. 1999) and ecological fieldwork. However, with an Fisher's α of 271 for the plots of RD, assuming that this is close to the correct Fisher's α for the area, we expect 271 species with only 1 individual, 135 with 2 individuals, 90 with 3 individuals, 68 with 4 individuals and nearly 800 species would have 10 individuals or less. RD covers 100 km^2, with an average tree density of 696 trees/ha (ter Steege et al. 2013). That indicates a total of 6.96 million individuals. The chance of such rare species there with feasible sampling intensity is thus very, very small. This is the consequence of using this theoretical framework (see also Hubbell (2015).

Because many researchers using nonparametric estimators assume that sampling is complete when the samples contain no singletons, an assumption that does not agree with ecological theory or with most ecological sampling, they are likely to severely underestimate richness when sampling level is low. Therefore, we suggest that the use of nonparametric estimators should be discouraged in studies with low sampling intensity in large remote areas. If the data can reasonably be assumed to follow a logseries, species estimation by means of Fisher's α is likely a better option. Other methods that produce abundance distributions with many singletons, matching most observational data, such as various parametric methods (Bunge and Barger 2008) or phenomenological theories, such as Maximum Entropy (Harte 2011), are probably also good alternatives.

Fisher's paradox

The term Fisher's paradox was coined by Hubbell (2015): "The logseries is an infinite series that mathematically goes on forever. But the world's forests are finite in size. So what happens to estimates of species abundance when the entire world is your sample? [. . .] The paradox would seem to run even deeper, because Fisher's logseries predicts that many more of the world's tropical tree species are hyper-rare. [. . .] The truth is, we still have inadequate data to definitively answer the "how many tropical tree species?" question. Ecologists at present are forced to make huge extrapolations from existing inventory plot data to the entire world."

Hubbell (2015) believes hyper-rare species do exist, as do we and in the case of areas smaller than the world, so do singletons. What then are those singletons? For an area like the Amazon, a huge and open system, singletons are most likely the result of species (locally) going extinct or new immigrants. ter Steege et al. (2016; Appendix S1: Fig. S7) showed that several singleton species are in fact species found only once in the Amazon but common in the Cerrado, Andes, and even Atlantic forest; "vagrants" in the terminology of Magurran and Henderson (2003). However, this may suggest that singletons or other hyper rare species are found mainly on the edges of an area. In the Amazon, they were not, and include such iconic species as *Asteranthos brasiliensis* Desf. (endemic to the middle and upper Rio Negro) and *Duckeodendron cestroides* Kuhlm. (endemic to an area around Manaus, central Amazon).

We believe that even if all individuals of the Amazon forest could be measured and identified, the biological processes of extinction and immigration would lead to the presence of at least ~750 singleton species, based on the Fisher's α found for the area (ter Steege et al. 2013) and a huge number of hyper-rare species, some of which may have small contracted ranges, but some of which may be spread over large areas (Zizka et al., *in press*). One of the most important merits of NT is to emphasize the role of migration in building and maintaining assemblage structures. However, the underlying mathematical model is based on a discretization down to the individual level, where a random process is supposed to play and can be expressed as per capita probabilities. In a complex system, such as tropical forests, clearly not only chance acts upon birth, death, dispersal, and migration. This could result from acquiring a new competitive advantage, losing a competitor because of a pest or losing a pest because a super-pest develops. Myriad combinations are possible. The processes involved at local scale are not exclusively random, but from local to global their combined effects on species abundances may sometimes appear to be.

Conclusion

To evaluate diversity of a rich, complex, large, open system, a parametric approach based on a probabilistic model, such as Fisher's logseries, seems to be more applicable than a non-parametric one, because such a system is driven by the random walk resulting from an infinity of processes that vary among scales, and where chance affects many biological processes, and not just the random sampling context considered by nonparametric methods.

Acknowledgments

We thank Carolina V. Castilho for the use of the RD plot data. H. ter Steege, V. F. Gomes, and R. Salomão are supported by grant 407232/2013-3–PVE–MEC/MCTI/CAPES/CNPq/FAPs; D. Sabatier and J. F. Molino are supported by a BGF grant for the project DynForDiv from the French Ministry of Ecology (MEDDE) and an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025), S. M. de Oliveira is supported by the "Alberta Mennega Stichting." W. E. Magnusson was supported by through INCT-CENBAM and PPBio-AmOc. The Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) provided logistical support to INCT-CENBAM.

Literature Cited

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

Bunge, J., and K. Barger. 2008. Parametric models for estimating the number of classes. Biometrical Journal 50:971–982.

Bunge, J., L. Woodard, D. Böhning, J. A. Foster, S. Connolly, and H. K. Allen. 2012. Estimating population diversity with CatchAll. Bioinformatics 28:1045–1047.

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

Castilho, C. V. 2004. Variação espacial e temporal da biomassa arbórea viva em 64 km2 de floresta de terra-firme na Amazônia central. Dissertation. INPA, Manaus, Brazil.

Chao, A. 1989. Estimating population size for sparse data in capture-recapture experiments. Biometrics 45:427–438.

Chao, A., and J. Bunge. 2002. Estimating the number of species in a stochastic abundance model. Biometrics 58:531–539.

Chao, A., R. K. Colwell, C.-W. Lin, and N. J. Gotelli. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125–1133.

Chao, A., and S.-M. Lee. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210–217.

Chiarucci, A. 2012. Estimating species richness: still a long way off!. Journal of Vegetation Science 23:1003–1005.

Chiarucci, A., N. J. Enright, G. L. W. Perry, B. P. Miller, and B. B. Lamont. 2003. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions 9:283–295.

Chisholm, R. A., and J. W. Lichstein. 2009. Linking dispersal, immigration and scale in the neutral theory of biodiversity. Ecology Letters 12:1385–1393.

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

Fine, P. V. A. 2001. An evaluation of the geographic area hypothesis using the latitudinal gradient in North American tree diversity. Evolutionary Ecology Research 3:413–428.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42–58.

Gotelli, N. J., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379–391.

Harte, J. 2011. Maximum entropy and ecology. A theory of abundance, distribution, and energetics. Oxford University Press, Oxford, UK.

Harte, J., and J. Kitzes. 2015. Inferring regional-scale species diversity from small-plot censuses. PLoS ONE 10:e0117527.

Harte, J., T. Zillio, E. Conlisk, and A. B. Smith. 2008. Maximum entropy and the state-variable approach to macroecology. Ecology 89:2700–2711.

Hortal, J., P. A. V. Borges, and C. Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75:274–287.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Hubbell, S. 2015. Estimating the global number of tropical tree species, and Fisher's paradox. Proceedings of the National Academy of Sciences USA 112:7343–7344.

Hubbell, S. P., and R. B. Foster. 1983. Diversity of canopy trees in a neotropical forest and implications for conservation. Pages 25–41 *in* S. J. Sutton, T. C. Whitmore, and A. C. Chadwick, editors. Tropical rain forest: ecology and management. Blackwell Science, Oxford, UK.

Hubbell, S. P., F. He, R. Condit, L. Borda de-Água, J. Kellner, and H. ter Steege. 2008. How many tree species are there in the Amazon and how many of them will go extinct? Proceedings of the National Academy of Sciences USA 105: 11498–11504.

Izsák, J., and S. Pavoine. 2012. Links between the species abundance distribution and the shape of the corresponding rank abundance curve. Ecological Indicators 14:1–6.

Kimura, M. 1985. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK.

Krishnamani, R., A. Kumar, and J. Harte. 2004. Estimating species richness at large spatial scales using data from small discrete plots. Ecography 27:637–642.

MacArthur, R. R., and E. O. Wilson. 1967. The theory of island biogeography. Princeton University Press, Princeton, New Jersey, USA.

Magurran, A. E., and P. A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422:714–716.

McGill, B. J., R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas, B. J. Enquist, J. L. Green, and F. He. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters 10:995–1015.

Oksanen, J., R. Kindt, P. Legendre, B. O'Hara, G. L. Simpson, M. Henry, H. Stevens, and H. Wagner. 2008. The vegan package. CRAN network, http://vegan.r-forge.r-project.org/

Plotkin, J. B., et al. 2000. Predicting species diversity in tropical forests. Proceedings of the National Academy of Sciences of the United States of America 97:10850–10854.

Pos, E., et al. *In press*. Estimating migration probability of spatially implicit versus explicit neutral models – a comparison between slow and fast growing Amazonian forests. Ecology and Evolution 2017:1–21.

Preston, F. W. 1948. The commonness, and rarity of species. Ecology 29:254–283.

R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org

Riberiro, J. E. L. S., M. J. G. Hopkins, A. Vicentini, C. A. Sothers, M. A. S. Costa, J. M. Brito, M. A. D. Souza, L. H. P. Martins, L. G. Lohmann, and P. A. C. L. Assunção. 1999. Flora da Reserva Ducke: Guia de Identificação das Plantas Vasculares de uma Floresta de Terra Firme na Amazônica Central. DFID, INPA, Manaus, Brazil.

Rocchetti, I., J. Bunge, and D. Bohning. 2011. Population size estimation based upon ratios of recapture probabilities. Annals of Applied Statistics 5:1512–1533.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge University Press, Cambridge, UK.

Sabatier, D., M. Grimaldi, M. F. Prevost, J. Guillaume, M. Godron, M. Dosso, and P. Curmi. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology 131:81–108.

Salomão, R. P. 2015. Restauraçao florestal de precisão: dinâmica e espécies estruturantes. Evolução de áreas restauradas em uma Unidade de Conservação na Amazônia – Porto Trombetas, Pará. OmniScriptum GmbH & Co. KG, Saarbrücken, Germany.

ter Steege, H., et al. 2013. Hyperdominance in the Amazonian tree flora. Science 342:1243092.

ter Steege, H., R. W. Vaessen, D. Cardenas, D. Sabatier, A. Antonelli, S. Mota de Oliveira, N. Pitman, J. M. Jørgensen, and R. P. Salomão. 2016. The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. Scientific Reports 6:29549.

Walther, B. A., and J. L. Moore. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815–829.

Wang, J.-P. 2011. SPECIES: an R package for species richness estimation. Journal of Statistical Software 40:1–15.

Wang, J.-P. Z., and B. G. Lindsay. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. Journal of the American Statistical Association 100:942–959.

Xu, H., S. Liu, Y. Li, R. Zang, and F. He. 2012. Assessing nonparametric and area-based methods for estimating regional species richness. Journal of Vegetation Science 23: 1006–1012.

Zizka, A., H. T. Steege, M. D. C. R. Pessoa, and A. Antonelli. *In press*. Finding needles in the haystack: Where to look for rare species in the American tropics. Ecography 40:1–12.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.1813/suppinfo