Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/3/10/e1701438/DC1

Supplementary Materials for

Upscaling species richness and abundances in tropical forests

Anna Tovo, Samir Suweis, Marco Formentin, Marco Favretti, Igor Volkov, Jayanth R. Banavar, Sandro Azaele, Amos Maritan

Published 18 October 2017, *Sci. Adv.* **3**, e1701438 (2017)
DOI: 10.1126/sciadv.1701438

This PDF file includes:

- fig. S8. Plot, in logarithmic scale, of the percentage p_{pred}% that one ought to sample to have a precision estimate of around 5% for the predicted percentage of hyper-rare species, that is, species with fewer than 1000 individuals at the global scale.
- table S1. Predicted number of singletons in the whole area of each tropical forest obtained by applying our method (NB method).
- table S2. Prediction of the total number of species obtained by applying both NB and LS methods to the forest generated according to an NB and distributed in 8900 × 8900 units according to two different modified Thomas processes with the same density of clusters $\rho = 6 \times 10^{-5}$ and different clump sizes $\sigma = 15$ and 200.
- table S3. Summary table of the most popular biodiversity estimators.
- table S4. Comparison between NB, LS, $Chao_{wor}$, and the Harte methods on empirical data.
- table S5. Comparison between the NB, LS, $Chao_{wor}$, and Harte methods on BCI empirical data.
- table S6. Number of species and singletons in 15 forests in our data set with the percentage p% (last column of the table) of surveyed area.
- Reference (*54*)

Supplementary Materials.
Upscaling Species Richness and Abundances in Tropical Forests

section S1. Upscaling biodiversity

In this section, we describe in detail the new framework used for upscaling biodiversity. We first describe the new approach based on the negative binomial distribution (NB) for the relative species abundance (RSA) and then the method based on Fisher's log-series (LS)[25]. Both approaches assume a scale independent form of the RSA. In order to apply the LS method[25], one needs to estimate the total abundance for the whole forest, whereas in the new approach this information is not needed. See Figure 2 of the main text for a schematic presentation of the methods.

NB method As explained in the Materials and Methods, the SAD is postulated to have a NB functional form, $\mathcal{P}(n|r,\xi\,)$, for non-zero populations, with parameters $(r,\xi\,)$ (r is known as the clustering coefficient)

$$P(n|1) = c(r,\xi\,)\mathcal{P}(n|r,\xi\,) \qquad \text{with} \qquad \mathcal{P}(n|r,\xi\,) = \binom{n+r-1}{n}\xi^n(1-\xi)^r, \qquad c(r,\xi\,) = \frac{1}{1-(1-\xi)^r}, \tag{S1}$$

where c is the normalization constant. The constant c is determined by imposing $\sum_{n=1}^{\infty} P(n|1) = 1$, where the sum starts from $n = 1$ because species with zero abundance at the scale of the whole forest will also be absent in the sub-plots. Note that $\mathcal{P}(n|r,\xi\,)$ is normalized for $n \geq 0$. This is because, in the sub-plots, there is a non-zero probability of species, present in the whole forest,

having $n = 0$ individuals, thereby accounting for the number of missing species in the sub-plots.

Let us now consider a sub-sample of area a of the whole forest and define $p = a/A$ the scale of the sample, that is the fraction of the sampled forest. The first step is to compute the RSA in the sub-sample.

We will assume that the sub-sample RSA is not affected by spatial correlations due to both interspecific and intraspecific interactions. This hypothesis is well satisfied as we will show in Section S.4 using *in silico* generated forests with various degrees of spatial correlations. Under this hypothesis, the conditional probability that a species has k individuals in the smaller area, $a = pA$, given that it has total abundance n in the whole region of area A is given by the binomial distribution

$$\mathcal{P}_{binom}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, \dots, n \tag{S2}$$

and $\mathcal{P}_{binom}(k|n,p) = 0$ if $k > n$. Now we want to prove that the sub-sample RSA, $P(k|p)$, is again a NB, for $k \geq 1$, with rescaled parameter ξ and the same r. Indeed, the probability, $\mathcal{P}_{sub}(k|p)$, of finding a species with population $k \geq 0$ in the sub-plot of area $a = pA$ is

$$\begin{aligned}
\mathcal{P}_{sub}(k|p) &= \sum_{n \geq k} \mathcal{P}_{binom}(k|n,p) P(n|1) = \sum_{n \geq k} \binom{n}{k} p^k (1-p)^{n-k} \cdot c(r,\xi\,) \binom{n+r-1}{n} \xi^n (1-\xi)^r \\
&= c(r,\xi\,) \binom{k+r-1}{k} \Big(\frac{p\xi}{1-\xi(1-p)} \Big)^k \Big(\frac{1-\xi}{1-\xi(1-p)} \Big)^r \\
&= c(r,\xi\,) \binom{k+r-1}{k} \hat{\xi}_p^k (1-\hat{\xi}_p)^r = c(r,\xi\,) \cdot \mathcal{P}(k|r,\hat{\xi}_p) \qquad k \geq 1,
\end{aligned} \tag{S3}$$

$$\begin{aligned}
\mathcal{P}_{sub}(0|p) &= \sum_{n \geq 1} \mathcal{P}_{binom}(k=0|n,p) P(n|1) = \sum_{n=1}^{\infty} (1-p)^n \cdot c(r,\xi\,) \binom{n+r-1}{n} \xi^n (1-\xi)^r \\
&= \frac{1}{1-(1-\xi)^{-r}} \Big(1 - \frac{1}{(1-\xi(1-p))^r} \Big) = 1 - \sum_{k \geq 1} \mathcal{P}_{sub}(k|p) \qquad k = 0,
\end{aligned} \tag{S4}$$

where we inserted the following explicit relation for $\hat{\xi}_p$

$$\hat{\xi}_p = \frac{p\xi}{1-\xi(1-p)} \tag{S5}$$

Recall that our method uses *only* the information we can infer from a sub-sample at some scale p^*. Therefore, we only have information on the abundances of the $S^*(\leq S)$ species present in the surveyed area. By denoting the number of species of abundance k at scale p^* by $S^*(k)$, we get

$$\frac{S^*(k)}{S^*} \equiv P(k|p^*) = \frac{\mathcal{P}_{sub}(k|p^*)}{\sum_{k'\geq 1}\mathcal{P}_{sub}(k'|p^*)} = \frac{\mathcal{P}(k|r,\hat{\xi}_{p^*})}{\sum_{k'\geq 1}\mathcal{P}(k'|r,\hat{\xi}_{p^*})} = c(r,\hat{\xi}_{p^*})\cdot\mathcal{P}(k|r,\hat{\xi}_{p^*}) \qquad k\geq 1 \tag{S6}$$

which, due to Eq.(S1), is a NB normalized for $k \geq 1$, whereas $\mathcal{P}(k|r,\hat{\xi}_{p^*})$ is normalized for $k \geq 0$. We have therefore obtained the key result that starting with a NB distribution for the RSA at the global scale, the RSA at smaller scales is also distributed according to a negative binomial with the same clustering coefficient r and a rescaled parameter $\hat{\xi}_{p^*}$ depending on both ξ and p^*. A RSA with the property of having the same functional form at different scales is said to be form-invariant.

By fitting the RSA of the data at the scale p^* we can thus find both the parameters r and $\hat{\xi}_{p^*}$ and, by inverting Eq.(S5) we can get ξ

$$\xi = \frac{\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(1-p^*)} \tag{S7}$$

Using Eq.(S5) to eliminate ξ from the last equation, one gets the following relation for the parameter ξ at the two scales p and p^* referred in the main text

$$\hat{\xi}_p = \frac{p\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(p-p^*)} \equiv U(p,p^*|\hat{\xi}_{p^*}). \tag{S8}$$

from which, of course, one can recover both Eqs.(S5) and (S7) where one has to use that $\xi \equiv \hat{\xi}_{p=1}$.

We now wish to determine the relationship between the total number of species at the whole scale $p = 1$, S, and the total number of species surveyed at scale p, S_p. For the scale p^*, in the following, we will use the notation $S^* \equiv S_{p^*}$. Note that

$$\mathcal{P}_{sub}(k = 0|p^*) = (S - S^*)/S \tag{S9}$$

$$\mathcal{P}_{sub}(k|p^*) = S^*(k)/S. \tag{S10}$$

Using Eq.(S4), the total number of species in the whole forest, in terms of the data on the surveyed sub-plot is given by

$$\begin{aligned} S &= \frac{S^*}{1 - \mathcal{P}_{sub}(k = 0|p^*)} \\ &= S^* \frac{1 - (1 - \xi)^r}{1 - (1 - \hat{\xi}_{p^*})^r} \end{aligned} \tag{S11}$$

where ξ is given by Eq.(S7).

LS method Let us now suppose that the RSA at the global scale is distributed according to a log-series with parameter x:

$$P(n|1) = P_{LS}(n|x) = \alpha(x)\frac{x^n}{n}, \qquad \alpha(x) = -(\log(1 - x))^{-1}, \tag{S12}$$

where α is the normalization constant.

The probability $\mathcal{P}_{sub}(k|p)$, of finding a species with population $k \geq 0$ in the sub-plot of area $a = pA$ is given by

$$\begin{aligned} \mathcal{P}_{sub}(k|p) &= \sum_{n \geq k} \mathcal{P}_{binom}(k|n, p) P(n|1) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1 - p)^{n-k} \cdot \alpha(x)\frac{x^n}{n} \\ &= \alpha(x)\Big(\frac{px}{1 - x(1 - p)}\Big)^k \frac{1}{k} = \alpha(x)\frac{\hat{x}_p^k}{k}, \qquad k \geq 1 \end{aligned} \tag{S13}$$

$$\begin{aligned}\mathcal{P}_{sub}(k=0|p) &= \sum_{n\geq 1}\mathcal{P}_{binom}(k=0|n,p)P(n|1) = \sum_{k=1}^{\infty}(1-p)^n\cdot\alpha(x)\frac{x^n}{n} \\ &= -\alpha(x)\log(1-x(1-p)) = \frac{\log(1-x(1-p))}{\log(1-x)}, \qquad k=0\end{aligned} \tag{S14}$$

where the parameter x at scale p is found to be:

$$\hat{x}_p = \frac{px}{1-x(1-p)} \tag{S15}$$

which is the same as Eq.(S5). Thus the analog of Eq.(S7),

$$x = \frac{\hat{x}_p}{p+\hat{x}_p(1-p)} \tag{S16}$$

and (S8) also holds in this case. The RSA, $P(k|p)$, is obtained as in Eq.(S6) and it is given by

$$P(k|p) = \frac{\mathcal{P}_{sub}(k|p)}{\sum_{k'\geq 1}\mathcal{P}_{sub}(k'|p)} = \alpha(\hat{x}_p)\frac{\hat{x}_p^k}{k} = P_{LS}(n|\hat{x}_p) \tag{S17}$$

Thus the Fisher log-series, which is a special case of the NB, is of course scale invariant as well.

The number of species with population $k \geq 1$, $S_p(k)$, in the sub-sample of area $a = pA$ is given by

$$S_p(k) \equiv S\mathcal{P}_{sub}(k|p) = S\alpha(x)\frac{\hat{x}_p^k}{k} = \hat{\alpha}\frac{\hat{x}_p^k}{k} \tag{S18}$$

where we gathered both the constants S and $\alpha(x)$ into a unique term $\hat{\alpha}$, which does not depend on the scale p. Again when referring to the scale p^*, we will use the shorthand notation $S^*(k) \equiv S_{p^*}(k)$.

Then the total number of species Sj and the total abundance N^* at the scale p^* are given, respec-

tively, by[?]

$$S^* = \sum_{k=1}^{\infty} S^*(k) = -\hat{\alpha}\log(1-\hat{x}_{p^*}) \tag{S19}$$

$$N^* = \sum_{k=1}^{\infty} kS^*(k) = \hat{\alpha}\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}} \tag{S20}$$

From the sample, because S^* and N^* are known, we can get the $\hat{\alpha}$ parameter by solving the following equation:

$$N^* - \hat{\alpha}\Big(\exp\Big(\frac{S^*}{\hat{\alpha}}\Big) - 1\Big) = 0, \tag{S21}$$

which has been obtained by inserting the expression for $\hat{x}_{p^*}$ from (S19) into (S20).

We now wish to infer information at the global scale $p = 1$ from the information we have at the scale $p = p^*$. We know from previous considerations that the $\hat{\alpha}$ parameter is scale-independent. Therefore, we have the following analogous relations for S and N:

$$S = -\hat{\alpha}\log(1-x) \tag{S22}$$

$$N = \hat{\alpha}\frac{x}{1-x} \tag{S23}$$

from which we obtain

$$S = \hat{\alpha}\log\Big(1+\frac{N}{\hat{\alpha}}\Big), \qquad \hat{\alpha} = S\alpha(x). \tag{S24}$$

In order to deduce the biodiversity S at the global scale, we first require an estimate of the total abundance N. Here we set $N = N^*/p^*$. This is consistent with our theoretical framework that assumes a form-invariant RSA. In fact, it can be easily proved that the mean total abundance scales linearly with the area when one assumes a LS-distributed RSA at the global scale

$$\mathbb{E}(N^*) = \sum_{k=1}^{\infty} kS^*(k) = \sum_{k=1}^{\infty} k\hat{\alpha}\frac{\hat{x}_{p^*}^k}{k} = \alpha\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}} = \hat{\alpha}\frac{px}{1-x} = p^*\mathbb{E}(N), \tag{S25}$$

where we have used Eq.(S15).

The very same result can be obtained if one assumes the RSA distributed as a negative binomial, Eqs.(S3) and (S10):

$$\mathbb{E}(N^*) = \sum_{k=1}^{\infty} kSc(r,\xi\,)\binom{k+r-1}{k}\hat{\xi}_{p^*}^k(1-\hat{\xi}_{p^*})^r = Sc(r,\xi\,)r\frac{\hat{\xi}_{p^*}}{1-\hat{\xi}_{p^*}} = Sc(r,\xi\,)r\frac{p\xi}{1-\xi} = p\mathbb{E}(N). \tag{S26}$$

Another way to infer the total biodiversity at the global scale is by using, as for the NB method, the following relation

$$S = \frac{S^*}{\sum_{k=1}^{\infty}\mathcal{P}(k|\hat{x}_{p^*})} = S^* \cdot \frac{\log(1-x)}{\log(1-\hat{x}_{p^*})}. \tag{S27}$$

In this case, we do not need an estimate of the total number of individuals N within the area A. We applied both the methods to extract biodiversity in our empirical forests and verified that the predictions were essentially the same.

Fisher log-series as a particular limit of the negative binomial We now show that the Fisher log-series is obtainable as the limiting case of the NB Eq.(S1). To do that we observe that

$$\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1)\,\Gamma(r)} \overset{r\approx 0}{\approx} \frac{r}{n+1} \tag{S28}$$

and taking the limit of small r of Eq.(S1)

$$\lim_{r\to 0} c(r,\xi\,)\mathcal{P}(n|r,\xi\,) = lim_{r\to 0}\frac{(1-\xi)^r}{1-(1-\xi)^r}\binom{n+r-1}{n}\xi^n = \frac{\xi^n}{-n\ln(1-\xi)} \tag{S29}$$

which is Eq.(S12) with $x = \xi$.

Assumptions of our method In our analysis, we assume that the probability that an individual tree falls within a given region is proportional to the region's area $a = pA$. This allows us to use

the formalism introduced in Section S.1. We refer to this assumption as the mean field hypothesis. A consequence of the mean field hypothesis is that when we wish to sample the $p\%$ of an area A where every individual has been catalogued into a list according to the species it belongs to, this is equivalent to sampling the $p\%$ of the individuals on this list. This is the only unbiased procedure one can utilize when neither spatial coordinates of the individuals nor spatial correlations are available.

In order for this hypothesis to be satisfied, one must first check if the region under study does not present strong inhomogeneities and anisotropies[14,39,54] – otherwise some species may tend to inhabit specific habitats of the region and therefore the assumption of a homogeneous spatial distribution of the individuals may fail. When extrapolating information to larger scales which present environmental inhomogeneities, we need a large number of randomly located samples in order to cover all the possible habitats, as emphasized by Slik[25].

It may also not be possible to neglect spatial correlations since they could have a strong influence on the spatial distribution of the individuals. For example, we test the influence of spatial correlations between individuals on empirical singleton curves for the French BBS dataset of 2010, which records the occupancy number of 246 species in 1096 cells located all around France. At variance with the case of tropical forests, here the curves obtained by considering or neglecting spatial effects are quite different especially for scales $\lesssim 60\%$. This discrepancy suggests that space cannot be neglected and thus it must be taken into account when analyzing those kinds of datasets.

section S2. Limitation of the LS methods

The LS method suffers from some important limitations. The first, already noted by several groups[14,17,18,28–32], is that in many cases the log-series distribution is not flexible enough[18] to describe the distinct observed RSA patterns: unimodal distributions are the norm, rather than the exception in tropical forests. This fact is reflected in the better performance of the NB method in predicting the biodiversity at larger scales in both artificial forests and in empirical tests (see Section S.4). There are two other important limitations that we describe below in detail.

Lack of flexibility of the LS in describing the singleton curve Using the theoretical framework described above we can determine the number of singletons in a sub-plot whose area is a fraction p of the area of the whole forest. The LS method predicts that the number of singletons is given by (see Eq.(S18)):

$$S_p(1) \equiv S\mathcal{P}_{sub}(k=1|p) \stackrel{LS}{=} S\alpha(x)\hat{x}_p = S\alpha(x)\frac{px}{1-x(1-p)} \qquad \text{for LS} \tag{S30}$$

(note that in Eq.(S18) we used the notation $S^*(k)$ instead of $S_p(k)$ used here). This is a monotonically increasing function of p, since S and $\alpha(x)$ are constants depending only on the composition of the forest at the global scale. In contrast, the number of singletons predicted by our approach, using a single NB, is given by (see Eq.(S10)):

$$S_p(1) \equiv S\mathcal{P}_{sub}(k=1|p) \stackrel{NB}{=} Scr\hat{\xi}_p(1-\hat{\xi}_p)^r \qquad \text{for NB} \tag{S31}$$

This, at variance with Eq.(S30), is not necessarily an increasing function of the sampled area, as we can see in fig. S1, but it depends on the values of the parameters. The negative binomial distribution is therefore more flexible.

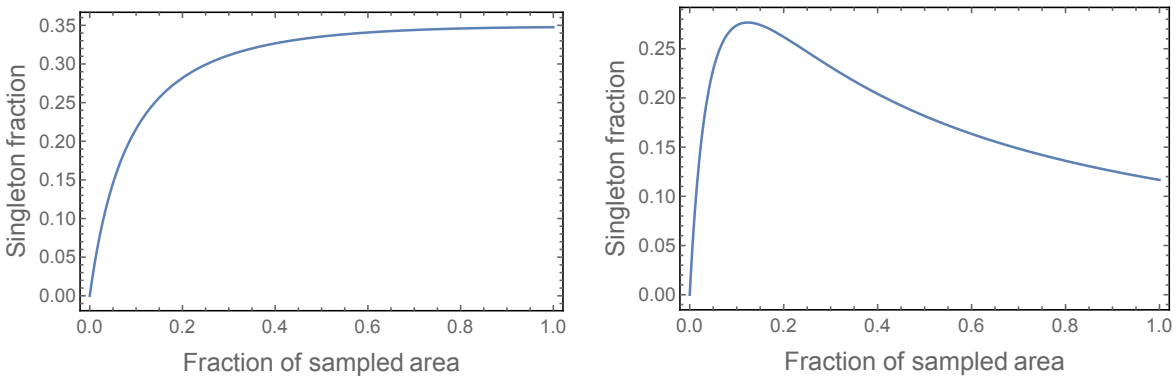


fig. S1. Assuming that the global RSA is distributed according to an NB, we can compute the probability that a species comprises a single individual at the scale p by using eq. S31.

Left panel: singleton fraction as a function of the fraction p of sampled area for a global RSA with parameters $r = 0.1$ and $\xi = 0.9$. Right panel: singleton fraction at different scales p for a global RSA with parameters $r = 0.9$ and $\xi = 0.9$. In contrast with the log-series case the curve does, not necessarily increase monotonically .

Dependence of Fisher's α from the sampling scale Slik et al.[25] showed that Fisher's α, that they deduced from three surveyed macro-regions using Eq.(S21), displays an asymptotic behaviour and they use the corresponding asymptotic value as a reliable estimate for Fisher's α at the global scale. This asymptotic α could be an artifact as its behaviour is affected by having sampled too low a percentage of the area. We have computed Fisher's α for the Amazonian dataset at different scales using the same Eq.(S21) and the empirical values of N^* and S^* (first panel of Figure S2). In particular, because no explicit spatial data were available, but just the RSA of the 4962 recorded species, mean values and error bars at each scale refer to 100 samples and the corresponding fraction of individuals, randomly picked among all the surveyed populations (see Section S.4 for an assessment of the spatial effects). At small scales (up to $\sim 10\%$), we can observe the same increasing behavior as for Slik's curves (see Figure S2d). Nevertheless, when the sampling percentage increases, the α-curve starts to slowly decrease. This means that in some intermediate range, as the sampled area increases, singletons disappear (because other individuals of the same species are found) at a rate faster than that at which new singletons are found. After this regime, the number of singletons reaches an asymptotic value. This phenomenon is even more evident in other cases, such as the Barro Colorado Nature Monument and the Caxiuana forest (second and third column of Figure S2).

The choice of the value of the parameter α strongly affects the predictions of both the number of species and singletons at the global scale, since both estimates are proportional to α itself (see Eqs.(S24) and (S30)). We have computed the number of singletons inferred with NB and LS methods for all the forests in our dataset (see table S1). Except for few cases, where the results are

comparable, usually the number of singletons predicted by the LS method is much larger than the one inferred by the NB approach.

section S3. Flexibility of NB distribution

Our method can be generalized to any linear combination of NBs with the same parameter ξ and different parameters r. For example, this result is particularly useful when dealing with data which present unusual behaviors which cannot be captured by a single NB distribution (see Figure S3). Indeed, one finds that in this case the predicted biodiversity is given by

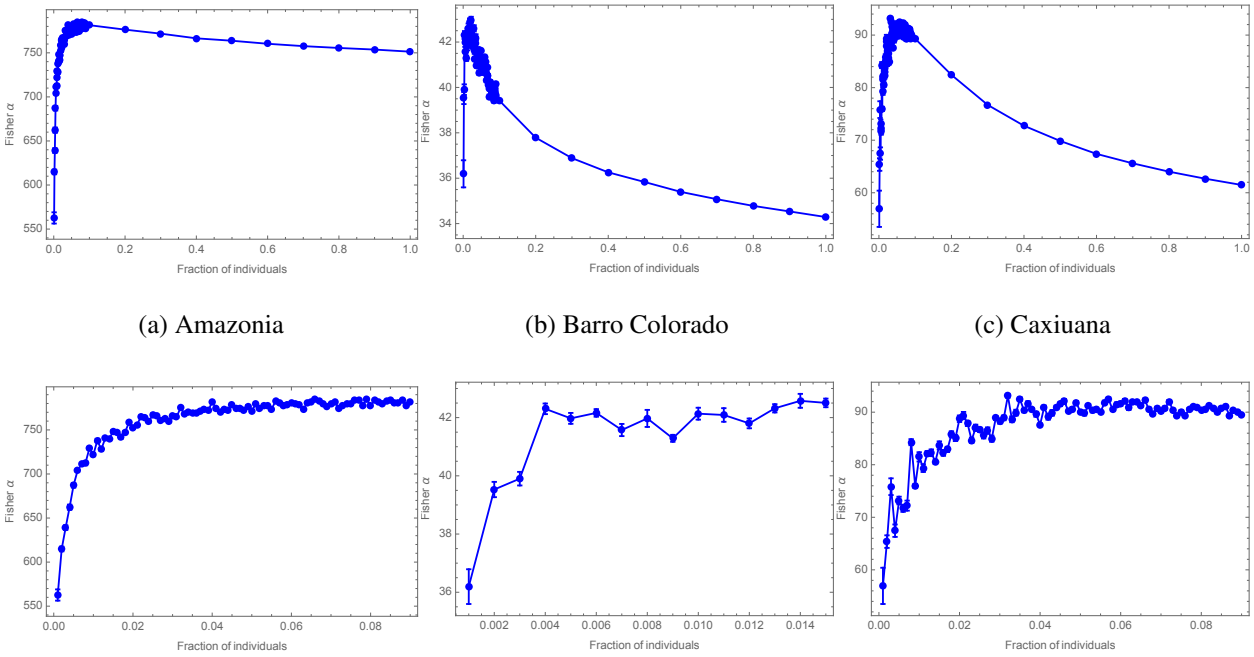
$$S = S^* \frac{\lambda[1-(1-\xi)^{r_1}] + (1-\lambda)[1-(1-\xi)^{r_2}]}{\lambda[1-(1-\hat{\xi}_{p^*})^{r_1}] + (1-\lambda)[1-(1-\hat{\xi}_{p^*})^{r_2}]}, \tag{S32}$$

where $\lambda \in (0,1)$ is the coefficient of the linear combination of the two negative binomials. The parameter ξ is given by Eq.(S7) whereas the parameters $r_1,\ r_2,\ \lambda$ and $\hat{\xi}_{p^*}$ are obtained by the best fit of the RSA of the surveyed area at scale p^* using the linear combination

$$\lambda c(r_1, \hat{\xi}_{p^*}) \cdot \mathcal{P}(k|r_1, \hat{\xi}_{p^*}) + (1-\lambda) c(r_2, \hat{\xi}_{p^*}) \cdot \mathcal{P}(k|r_2, \hat{\xi}_{p^*}). \tag{S33}$$

In principle, one could use a generic combination of an arbitrary number of negative binomials to better fit the distribution. In fact, this can be done to any degree of precision due to the following considerations. The generating functions of the negative binomials, defined by

$$G(z|r,\xi\,) = \sum_{n\geq 0} z^n \mathcal{P}(n|r,\hat{\xi}) = \Big(\frac{1-\xi}{1-z\xi}\Big)^r, \tag{S34}$$



(a) Amazonia (b) Barro Colorado (c) Caxiuana

(d) Amazonia zoom (e) Barro Colorado zoom (f) Caxiuana zoom

fig. S2. Fisher's α for three different rainforests: Amazonia, Barro Colorado Nature Monument, and Caxiuana. For $p < 1$, (top panels) there is no asymptotic limit of Fisher's α. If we zoom the Fisher's α for only the smallest scales (bottom panels), an apparent asymptote is reached. Nevertheless, this is not the real asymptotic Fisher's α. Mean values and error bars at each scale refer to 100 samples and the corresponding fraction of individuals, randomly picked among all the surveyed population.

table S1. Predicted number of singletons in the whole area of each tropical forest obtained by applying our method (NB Method). In the last column, we show the results of the LS method. The NB method yields similar results to the LS method, but without needing an estimate of N, the total number of trees.

Forest	Observed Singletons	NB Method	LS Method
AMAZONIA	645	581	751
BARRO COLORADO NATURE MONUMENT	17	16	34
BUKIT BARISAN	13	2	62
BWINDI IMPENETRABLE FOREST	3	1	19
CAXIUANA	1	1	61
COCHA CASHU MANU NATIONAL PARK	12	3	94
KORUP NATIONAL PARK	0	1	37
MANAUS	11	0	175
NOUABALE NDOKI	0	0	18
PASOH FOREST RESERVE	94	30	118
RANOMAFANA	3	2	40
UDZUNGWA MOUNTAIN NATIONAL PARK	3	1	15
VOLCAN BARVA	5	1	59
YANACHAGA CHIMILLEN NATIONAL PARK	52	58	58
YASUNI NATIONAL PARK	7	4	97

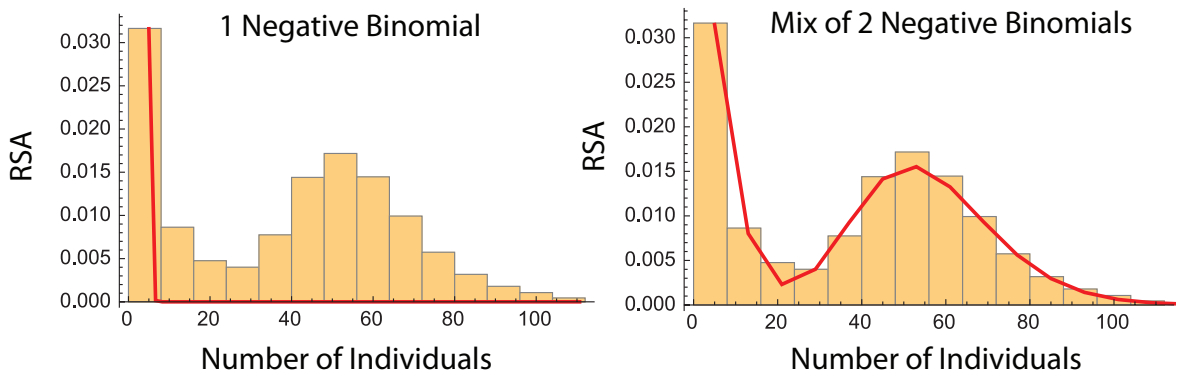


fig. S3. Fit of an RSA consisting of a combination of an LS and a log-normal distribution. On the left the RSA has been fitted through a negative binomial, which cannot capture the unusual behavior of the distribution. On the right is shown an improved fit with a combination of two negative binomials with the same parameter ξ and different clumping parameters like in Eq.(S33).

with $0 \leq z \leq 1$ and $r \geq 0$, generate an algebra of functions when considered as a function of r, with ξ parameter fixed. Such an algebra satisfies the hypothesis of Nachbin theorem[33,34], since its separates the points and the tangent vectors of $\mathbb{C}$. In particular, in our case this second hypothesis reduces to the existence, for each point $x \in \mathbb{C}$, of a function f of the algebra such that $df(x) \neq 0$. Thus, from Nachbin theorem, the algebra is dense in the space of complex functions with the topology induced by the derivatives up the k^{th} order, $C^k(\mathbb{C})$. This implies that any k-times differentiable function of $z \in [0,1]$ can be approximated together with its k derivatives to an arbitrary degree of precision by a linear combination

$$\sum_{i=1}^{\ell} \lambda_i G(z|r_i, \xi) \tag{S35}$$

where ℓ depends on the desired precision and the λ_i's are suitable coefficients. Thus, given an

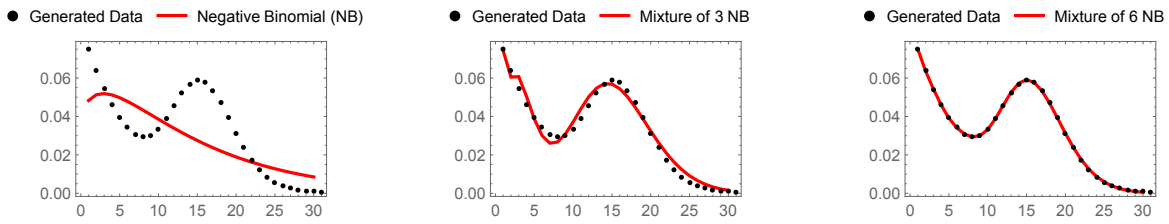


fig. S4. We have generated synthetic data from a combination of discrete distributions (a binomial distribution of parameters $r = 40$ and $\xi = 0.8$, a geometric distribution of parameter $\mu = 0.15$, and a Poisson distribution with parameter $\lambda = 15$) and fit these data with one, three, and six NBs, respectively. As shown, with six negative binomials, we obtain a perfect fit of the data, as suggested by the theorem.

arbitrary (discrete) probability distribution, $P(n)$, we can approximate its generating function, $G(z) = \sum_{n\geq 0} z^n P(n)$, with a linear combination such as Eq.(S35), and its n-th derivative at $z = 0$ gives us the estimate of $P(n)$ at the desired precision. Beware that the approximating linear combination may fail to be a probability generating function. It may not be normalised or some of its derivatives may be negative. Anyway, within our framework the error can be managed as long as we restrict to the case of finite population size.

section S4. Test on computer-simulated forests

In order to test the LS method and our approach based on the NB distribution, we have generated various kinds of artificial forests with and without spatial correlations.

Artificial forests without spatial correlations In this case, the forests are obtained by drawing 5000 species from two of the commonly used RSA for modeling/fitting tropical forest abundances: a negative binomial (NB forest) and a log-normal (LN forest) distribution. We note that an LS forest would be the limit of $r \to 0$ of a NB forest, as shown above. When a zero abundance is generated, the corresponding species is deleted from the dataset.

In the mean field hypothesis (the random sampling described in Section S.1), sampling the fraction p of the whole forest area is equivalent to randomly sample a fraction p of the individuals. We thus tested the two methods predicting the total biodiversity starting from different spatial scales, p. The results are reported in Table S2 (NB forest) and in the Results section of the main text (LN forest). In particular, we see that the NB method, even using a single negative binomial, works well in all cases, while the LS method overestimates the biodiversity when the generated forest has a RSA which is not a log-series. Therefore, the NB method is more flexible and robust even when a negative binomial distribution is not the RSA of the whole forest.

We want to stress that when fitting the sample of the simulated forest with the log-series (LS method), we use as the number of individuals N of the whole area its exact value (that we know as we have generated the forest). We do this to favorably bias the chances of success of the LS method.

Artificial forests with spatial correlations To test the robustness of our method with respect to spatial correlations and sampling methods, we distributed the individuals of the NB and LN forests in a 8900x8900 and in a 4900x4900 grid (where a unit corresponds to 1 meter) respectively, according to two modified Thomas processes[12,38,39]. We recall that this process can be simulated

by first distributing the parents' locations (clusters' centers) according to a Poisson process with intensity ρ. Given then the total number of individuals to be placed within the area of the sample, we randomly assign each of them to one of the previously generated parents. We thus place the offspring at a position drawn from a two-dimensional Gaussian distribution centered at the location of the parent and with variance σ. We have imposed toroidal boundary conditions in order to minimize finite size effects for the whole (artificial) forest. Finally, the parents are removed from the dataset, leaving just the off-springs at their locations.

We set the density of clusters $\rho = 6 \cdot 10^{-5}$ and we chose two clump sizes $\sigma = 15$ and 200 in order to compare the performance of the methods for different degrees of spatial correlations. The area of the global region was chosen with the same ratio N/A of individuals per unit area. We then infer the number of species in the whole area by sampling a percentage $p\% = 1\%$ and 5% of it.

For the NB forest, we consider two different sampling methods: a first one where we survey non-overlapping 1-ha plots at randomly chosen locations within the available area and a second one where we collect data within a unique plot of the same total desired area. Our results are shown in table S2. In fig. S5, we show a schematic presentation of the datasets generated according to different clumping parameters and of the different sampling methods.

We also tested the robustness of the method with respect to different clumping coefficients of the generating Thomas process for the LN forest. Even in this case, no significant difference was recorded (see Results section of the main text). In all cases, the mean values and standard errors refer to 100 trials.

The NB method gives the same results both with different sampling methods and spatial correla-

tions and works well at all spatial scales.

section S5. Comparison with other popular estimators

In this section we compare our method with the most popular ones proposed in literature[20,25], which are summarized in Table S3. We found that our method outperforms that of Chao (denoted with $Chao_{wor}$ in Table 3) for Amazonia, Pasoh and Yasuni. while it is better than LS method and comparable to Chao for the remaining forests. In the latter, the difference between S_{p^*} - the number of observed species at the sampled scale p^* - and S_p - the one at sub-sample scale p - goes to zero very fast as p approaches p^*. At the same time, the number of singletons quickly decreases to (very) small values (even zero). In these cases, Chao's predictions, based on singleton and doubleton species lead to $S_p \approx S_{p^*}$. This effect is very clear when we compare the SAR produced by NB and Chao's methods (see Figure 4 of the main text). In the latter case the SAR remained constant for a great part of the scale range larger than p^*, (see Table S6). Indeed for this range of p^*, the SAR predicted by Chao can be approximated as $S^p_{pred} \approx S_{p^*} + \frac{(\#\text{ singletons})^2}{2\ \#\text{ doubletons}}$. On the other hand, the SAR predicted by the NB method displayed the typical shape of the SAR observed in real ecosystems. We finally found that other methods[19,20] do not converge to S_{p^*} as $p \to p^*$, i.e., they do not have an explicit dependence on the surveyed area, rather they give an upscaled biodiversity estimates only based on the number of singletons or doubletons (see Figure S6). Therefore we excluded these predictors from our analysis.

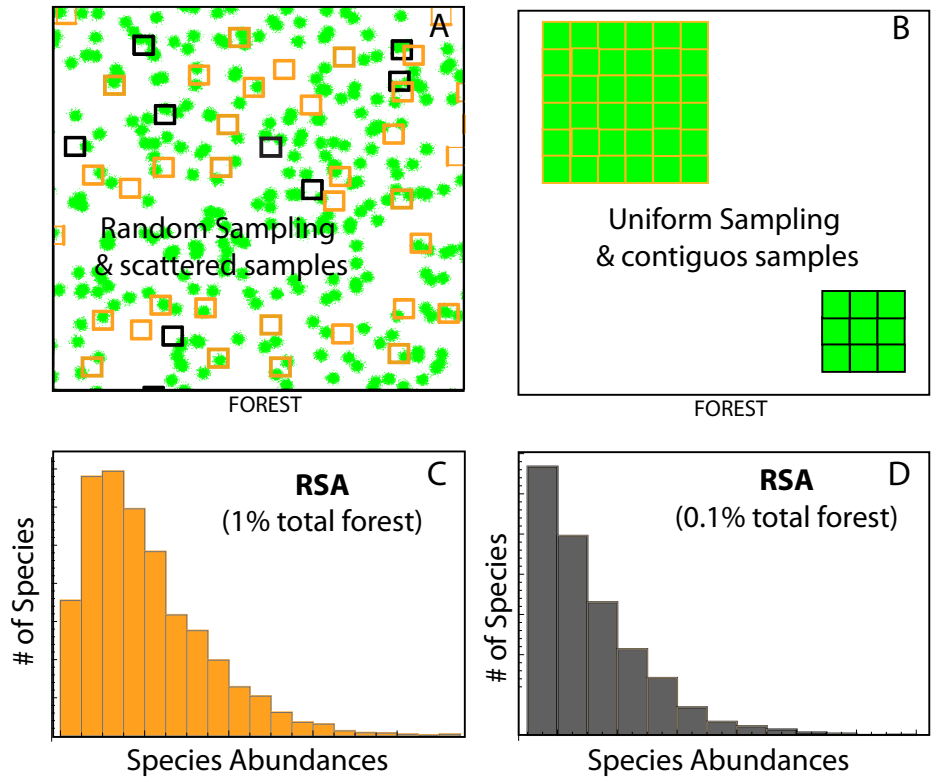


fig. S5. **Robustness of the Method.** A) We test the robustness of the method with respect to different spatial correlations and sampling methods. We distribute the individuals of an "artificial" forest on an area A according to two modified Thomas processes with the same density of clusters $\rho = 6 \cdot 10^{-5}$, two different clump sizes $\sigma = 15,\ 200$ and different RSAs (see Table S2 below and Results section of the main text). In A)-B) green dots are plants' individuals which are either clumped (A) or uniformly distributed (B). We then wish to infer the number of species in the whole area by sampling a fraction p^* of it. We consider two different sampling methods: a first one where we survey non-overlapping plots at randomly chosen locations within the available area (left panel) and a second one where we collect data within a unique plot of the same desired area (right panel). In the figure, orange squares correspond to $p^* = 0.01$ sampling, while black squares represent $p^* = 0.001$ (i.e. 1% and 0.1% respectively). C-D) RSA of the species sampled at the 1% and 0.1% scale. We note that the RSA in (D) does not exhibit a mode due to the effect of the veil-line[27]: the rarest species in the 1% case are not sampled in the 0.1%, leading to a mode of the observed distribution in the 1% case and not in the 0.1% case.

table S2. Prediction of the total number of species obtained by applying both NB and LS methods to the forest generated according to an NB and distributed in 8900 x 8900 units units according to two different modified Thomas processes with the same density ofclusters $\rho = 6 \times 10^{-5}$ and different clump sizes $\sigma = 15$ and 200. Mean values and related standard errors on 100 trials are reported for each percentage of sampling. The NB method works well in all cases and its results are robust with respect both to the sampling method and the presence of spatial correlations. In contrast, the LS method does not give reliable results, because the basic hypothesis of a log-series RSA does not hold.

p=1%		Empirical Data	NB Method	LS Method
High-clustered forest	random samples	4974	4973±5	9823±20
	increasing-area sample	4974	4961±5	9918±35
Low-clustered forest	random sample	4974	4970±4	9834±5
	increasing-area sample	4974	4968±4	9876±21
p=5%		Empirical Data	NB Method	LS Method
High-clustered forest	random sample	4974	4974±1	7448±7
	increasing-area sample	4974	4981±1	7567±26
Low-clustered forest	random sample	4974	4975±1	7440±1
	increasing-area sample	4974	4975±1	7550±26

table S3. Summary table of the most popular biodiversity estimators.

Estimator	**Predicted *S***	**Details**
NB Method	$S^* \dfrac{1-(1-\xi)^r}{1-(1-\hat{\xi}_{p^*})^r}$	$(\xi,\ r)$ NB parameters at scale 1 $(\hat{\xi}_{p^*},\ r)$ NB parameters at scale p^*
Slik	$\hat{\alpha}\log\left(1+\frac{N^*}{\hat{\alpha}}\right)$	$\hat{\alpha}$ s.t. $N^* - \hat{\alpha}\left(\mathrm{e}^{S^*/\hat{\alpha}}-1\right)=0$ $N^* =$ observed individuals at p^*
Chao_1	$S^* + \begin{cases} \dfrac{f_1^2}{2f_2} & f_2>0 \\ \dfrac{f_1(f_1-1)}{2} & f_2=0 \end{cases}$	$f_i =$ # of species with i individuals at the scale p^*
Chao_{bc}	$S^* + \dfrac{N^*}{N^*-1}\dfrac{f_1(f_1-1)}{2(f_2-1)}$	
iChao_1	$S_{Chao_1} + \dfrac{N^*-3}{N^*}\dfrac{f_3}{4f_4}\max\left(f_1 - \dfrac{(N^*-3)f_2f_3}{(N^*-1)2f_4}, 0\right)$	$S_{Chao_1} = S$ predicted by $Chao_1$ method
Chao_{wor}	$S^* + \dfrac{f_1^2}{\frac{N^*}{N^*-1}2f_2 + \frac{p^*}{1-p^*}f_1}$	
Jacknife_1	$S^* + \dfrac{N^*-1}{N^*}f_1$	
Jacknife_2	$S^* + \dfrac{2N^*-3}{N^*}f_1 - \dfrac{(N^*-2)^2}{N^*(N^*-1)}f_2$	
Turing	$S^*_{abun} + \dfrac{S^*_{rare}}{\hat{C}_{rare}}$	$S^*_{abun} = \sum_{n>10} f_n$ $S^*_{rare} = \sum_{n=1}^{10} f_n$ $\hat{C}_{rare} = 1 - f_1/\sum_{n=1}^{10} nf_n$
ACE	$S_{Turing} + \dfrac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare}$	$\hat{\gamma}^2_{rare} = \max\{\gamma - 1, 0\}$ where $\gamma = \dfrac{f_1}{\hat{C}_{rare}}\dfrac{\sum_{n=1}^{10} n(n-1)f_n}{(\sum_{n=1}^{10} nf_n)(\sum_{n=1}^{10} nf_n - 1)}$

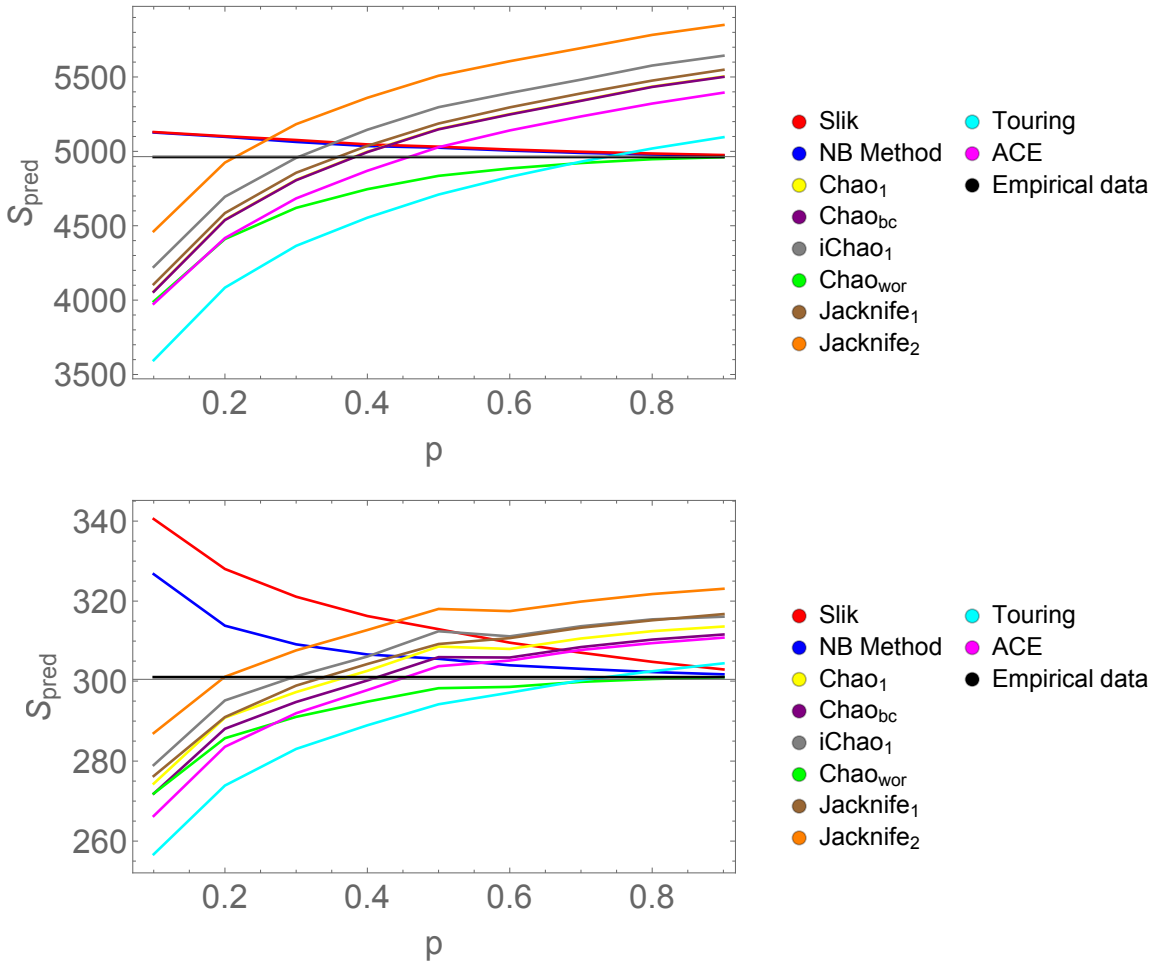


fig. S6**. Comparison between biodiversity estimators for Amazonia and BCI forests.** Predicted biodiversity at the sample scale $p^* = 1$ from sub-samples at scales $p < p^*$ with the most popular estimators summarized in Table S3. While the NB, Slik and Chao$_{wor}$ methods do converge at S_{p^*} as p goes to p^*, all the others[20] have a monotonically increasing behaviour due to the independence, in their predictions, of the scale p.

Finally, we compared our method to another important upscaling technique based on the principle of Maximum Entropy proposed by Harte and collaborators [21]. In Table S4 we display the results on four empirical forests. For each of them, we sub-sampled a fraction $p = 0.1$ of the individuals and predicted the species richness at a larger scale, where the true value of S^* is known (second column of the table). Because Harte's upscaling procedure allows one to scale up by successive factors of two [21], we cannot obtain an estimate at $p = 1$. The last two columns of Table S4 refer to the predictions at $p = 0.8$ and $p = 1.6$, which represent lower and upper bounds for the species richness at the desired scale of $p = 1$. For the first three forests, Harte's method does not perform as well as the others, with typical errors around 20%. For the last forest, the performance is comparable to Chao_{wor}, while being a bit worse than both the NB and LS methods. These two latter methods yield very similar results because the best fitting of the empirical SAD with a negative binomial resulted in an r parameter very close to zero. The SAD, in this case, does in fact resemble a log-series, a result also valid for Hartes method. We also compared the NB, LS, Chao_{wor} and Harte methods on BCI empirical data, when a contiguous area is sampled (see Table 5). More precisely, we sub-sampled a fraction $p = 0.25$ and $p = 0.5$ of the individuals and predicted the species richness at the $p = 1$ scale, where the true value of S^* is known (301 species). In both cases, the NB method outperformed those of Harte, whose estimates are comparable to those of the LS method. This is in accord with theoretical expectations because both the LS and Harte procedures are based on the assumption of a log-series SAD.

table S4. Comparison between NB, LS, Chao$_{\text{wor}}$, and the Harte methods on empirical data. For each tropical forest, we sub-sampled a fraction $p = 0.1$ of the individuals and predicted the species richness at the $p = 1$ scale, where the true value of S^* is known (second column). For Harte's method, two estimates are shown because the iterative method permits upscaling at scales which is a factor of two higher than the previous scale [21]. Here we show predictions at $p = 0.8$ and $p = 1.6$, which are bounds on the species richness at $p = 1$.

FOREST	True S^*	NB		LS		Chao$_{\text{wor}}$		Harte	
		S_{pred}	% err	S_{pred}	% err	S_{pred}	% err	S_{pred}	% err
BCI	301	327	8.6	341	13.3	272	9.6	382/430	26.9/42.9
PASOH	927	910	1.8	1049	13.2	805	13.2	1192/1362	28.6/49.9
AMAZONIA	4962	5127	3.3	5130	3.4	3991	19.6	6060/7107	22.1/43.2
YANACHAGA	209	182	12.9	182	12.9	148	29.2	241/320	15.3/53.1

table S5. Comparison between the NB, LS, Chao$_{\text{wor}}$, and Harte methods on BCI empirical data. We considered two sub-samples consisting of contiguous fractions $p = 0.25$ and $p = 0.5$ of the surveyed area.

p*	True S^*	NB		LS		Chao$_{\text{wor}}$		Harte	
		S_{pred}	% err	S_{pred}	% err	S_{pred}	% err	S_{pred}	% err
0.25	301	310	3.0	325	8.0	287	4.7	333	10.6
0.5	301	306	1.7	313	4.0	298	1.0	315	4.7

section S6. Data set

We collected data of 15 forests around the planet on different tropical field stations of the equatorial zone. The number of observed species and singletons for each forest are reported in Table S6. All datasets are publicly available or upon request.

Datasets of Bukit Barisan, Bwindi Impenetrable Forest, Caxiuana, Cocha Cashu - Manu National Park, Korup National Park, Manaus, Nouabalé Ndoki, Ranomafana, Udzungwa Mountain National Park, Yanachaga Chimillen National Park and Yasuni National Park have been provided by Tropical Ecology, Assessment and Monitoring (TEAM) Network of Conservation International (see http://www.teamnetwork.org/data/use).

The Amazonian dataset came from the paper *Hyperdominance in the Amazonian Tree Flora* by Hans ter Steege et al.[24] (http://science.sciencemag.org/content/342/6156/1243092.figures-only).

The Pasoh and Barro Colorado Island datasets have been provided by the Center of Tropical Research Science of the Smithsonian Tropical Research Institute (http://www.ctfs.si.edu/site).

In particular, for Barro Colorado Island we used the 2005 census and we conducted our analysis by considering all provided species, with no restriction based on dbh (saplings included). Following the analysis in Slik et al.[25], we removed individuals whose taxa were classified as unknown.

table S6. Number of species and singletons in 15 forests in our data set with the percentage $p\%$ (last column of the table) of surveyed area.

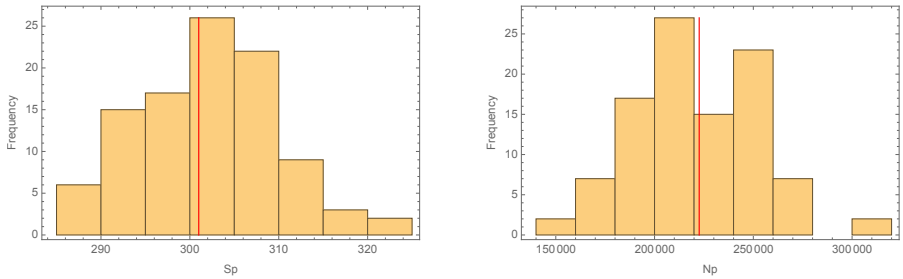
Forest	Species	Singletons	p%=100p
AMAZONIA	4962	645	0.00016
BARRO COLORADO NATURE MONUMENT	301	17	3.20513
BUKIT BARISAN	340	13	0.00169
BWINDI IMPENETRABLE FOREST	128	3	0.01813
CAXIUANA	386	1	0.01818
COCHA CASHU - MANU NATIONAL PARK	489	12	0.00035
KORUP NATIONAL PARK	226	0	0.00473
MANAUS	946	11	0.06000
NOUABALÉ NDOKI	110	0	0.00143
PASOH FOREST RESERVE	927	94	0.35714
RANOMAFANA	269	3	0.01463
UDZUNGWA MOUNTAIN NATIONAL PARK	109	3	0.00302
VOLCAN BARVA	392	5	0.02025
YANACHAGA CHIMILLEN NATIONAL PARK	209	52	0.00372
YASUNI NATIONAL PARK	481	7	0.61100

section S7. Self-consistency and estimation of the critical p^*: How much remains to be sampled?

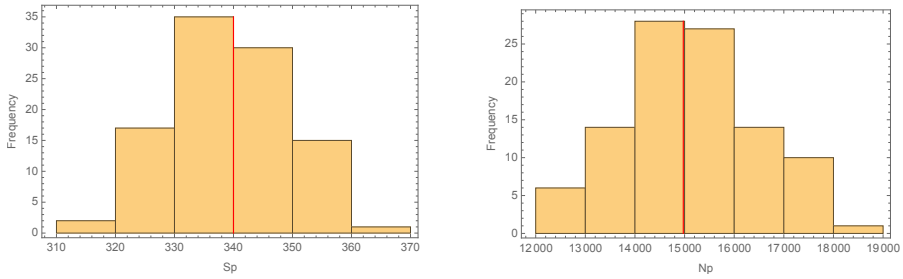
To check the self-consistency of our framework, we run the following test on the empirical forests. We generate the corresponding global forests according to the RSA and the number of species predicted by our method at the global scale. We then sample $N_{p*} = p^* N$ individuals and measure the number of different species (S_{p*}) to which they belong. In summary, from the predicted RSA at the global scale, we can reproduce, by sub-sampling, the empirical values of the number of species, S_{p*}, and the number of individuals, N_{p*}, at the scale $p*$. For each forest, we run the test 100 times and produced the histograms in Figure S7. For all the forests, the red lines representing the empirical values of S_{p*} and N_{p*} in our dataset turn out to be *typical* values.

Using our results on the up-scaled forest biodiversity, it is possible to estimate the percentage of the forest that still needs to be sampled in order to have an estimation error around 5%. We proceed as follows:

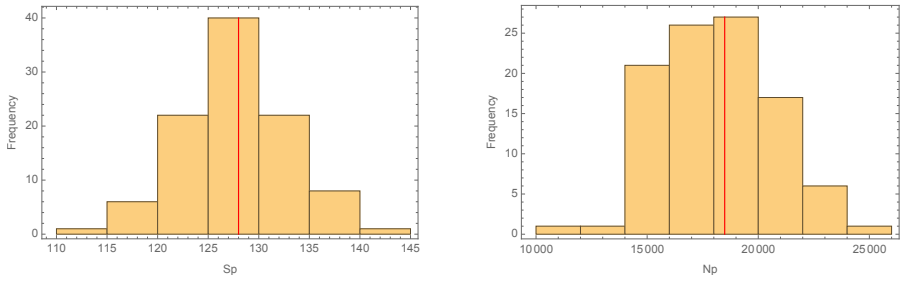
1. employing our estimation of the RSA parameters and of the total number S of the species at the global scale, we generate the predicted forest;
2. we sample the global forest at larger and larger scales p, extracting for each of them 100 samples consisting of $N_p = pN$ randomly chosen individuals;
3. we apply our method to each sample obtaining an estimation S_{pred} of S;
4. we compute for each scale mean values μ and standard deviations σ of the 100 relative errors obtained $(S_{pred} - S)/S$;



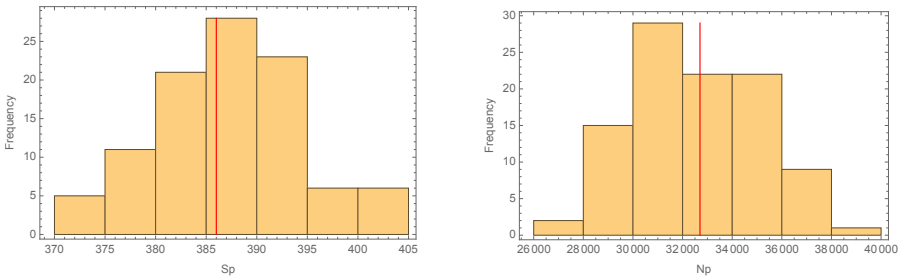
(a) Barro Colorado Nature Monument



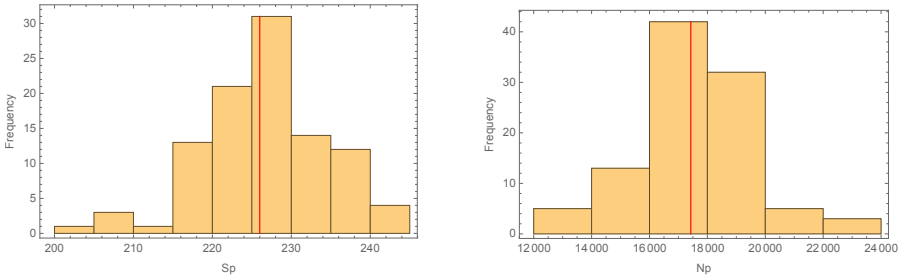
(b) Bukit Barisan



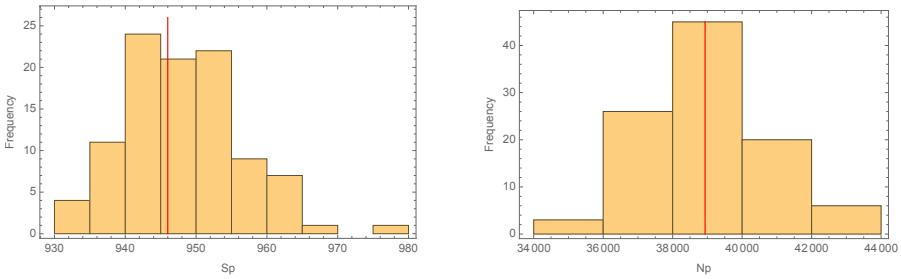
(c) Bwindi Impenetrable Forest



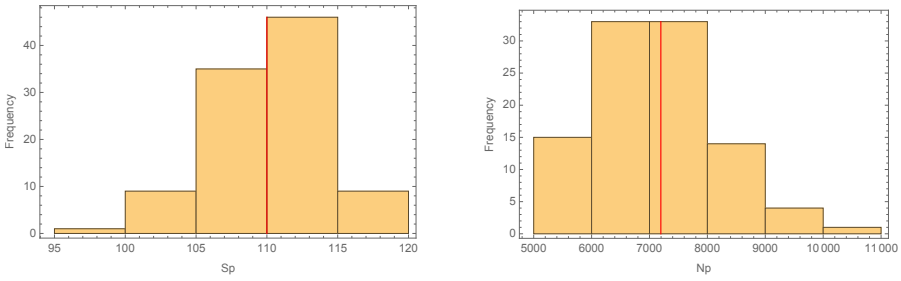
(d) Caxiuana



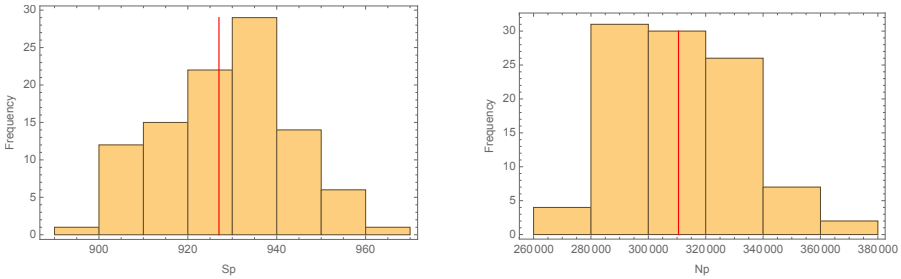
(e) Korup National Park



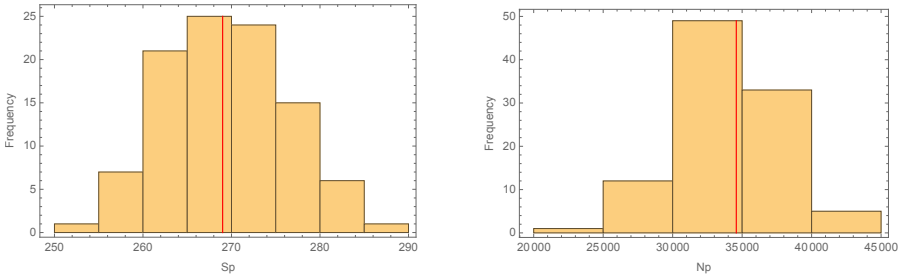
(f) Manaus



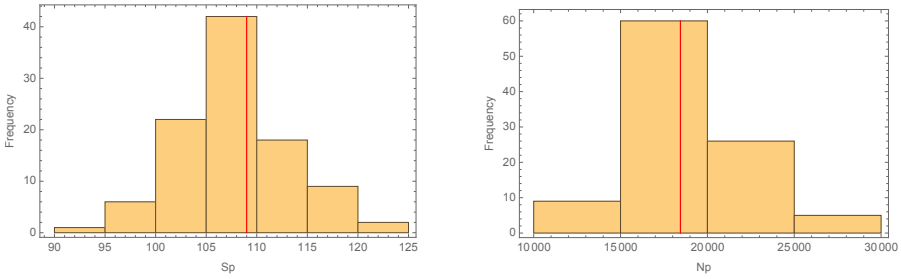
(g) Nouabalé Ndoki



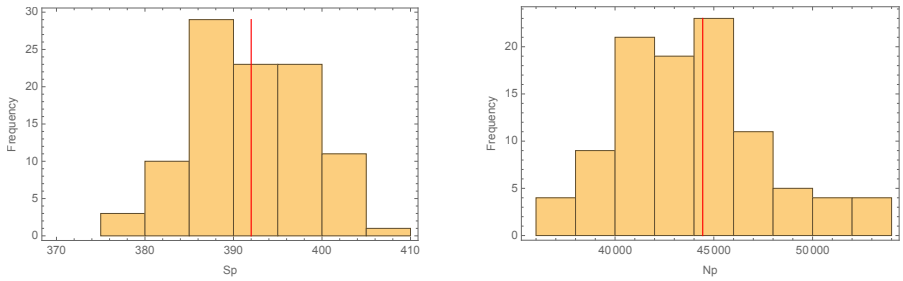
(h) Pasoh Forest Reserve



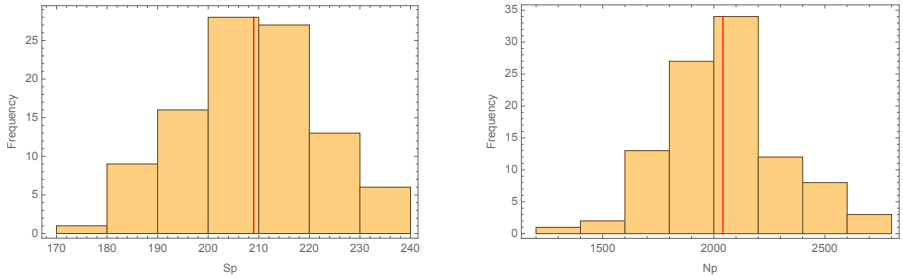
(i) Ranomafana



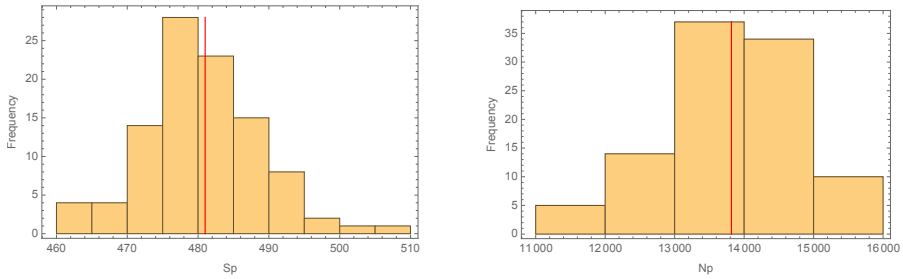
(j) Udzungwa Mountain National Park



(k) Volcan Barva



(l) Yanachaga Chimillen National Park



(m) Yasuni National Park

fig. S7. **Self-consistency test of our framework**. Starting with the RSA and the number of species at the global scale predicted by our method, we generated an artificial forest. We then sampled a fraction p^* of the area and measured the number of different species (S_{p^*}) and the number of individuals (N_{p^*}) at the scale p^*. For each RSA of an empirical forest, we run this test 100 times and produced the histograms depicted above. The red lines represent the empirical value of S_{p^*} and N_{p^*} in our dataset.

5. we select the scale at which 95% of the samples lead to an error less or around 5% with respect to the true value of S (see main text).

In Figure S8, we plot these values against the percentage of hyper-rare species for each forest in log-log scale (see Table 3 of the main text). Intuitively, the higher the number of the rare species of a forest, the bigger the percentage one should sample in order to get an estimate of the total number of species within a given error. Indeed, we can observe a slight increasing trend in the data points. If we exclude the Amazonia dataset, which is clearly an outlier, we get a correlation coefficient of 0.5. If we also exclude the three forests of Bwindi, Udzungwa and Yanachaga, for which we would need a few hundred times the actual sampling to have an estimation precision around 5%, the correlation coefficient rises up to 0.8.

section S8. RSA parameters maximize relative fluctuation in abundances

The parameters of the NB method, which provide the best predictions are very close to $r = 0$ and $\xi = 1$, regardless of the forest. This is somewhat unexpected, because there are neither theoretical nor biological reasons why tropical forests in different geographical locations and with different biodiversity richness should have abundances distributed across species in a very similar manner. As underscored in the main text, a closer look at the NB distribution reveals that, in this region of parameter space, the relative fluctuation of abundances is maximized.

Let (see Eq.(S1)) $P(n|1) = c(r,\xi)\mathcal{P}(n|r,\xi) = \binom{n+r-1}{n}\xi^n(1-\xi)^r/[1-(1-\xi)^r]$ be the RSA in the NB hypothesis and let us compute its first two moments, which we denote by $< n >$ and $< n^2 >$

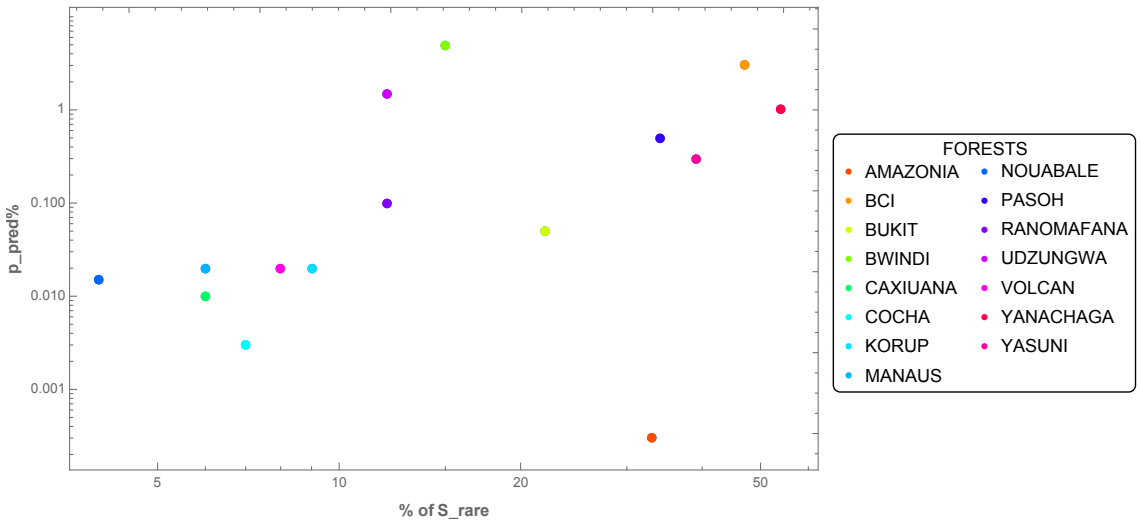


fig. S8. Plot, in logarithmic scale, of the percentage $p_{pred}\%$ that one ought to sample to have a precision estimate of around 5% for the predicted percentage of hyper-rare species, that is, species with fewer than 1000 individuals at the global scale. Data points show a slight increasing trend with few outliers.

respectively. These can be easily calculated to give

$$< n >= \sum_{n=1}^{\infty} nP(n|1) = \frac{\xi r}{(1-\xi)(1-(1-\xi)^r)} \tag{S36}$$

and

$$< n^2 >= \sum_{n=1}^{\infty} n^2 P(n|1) = \frac{\xi r(1+\xi r)}{(1-\xi)^2(1-(1-\xi)^r)} \tag{S37}$$

Then the relative fluctuation in abundances, $F(\xi, r)$, is given by

$$\begin{aligned} F(\xi, r) &= \frac{< (n- < n >)^2 >}{< n >^2} = \frac{< n^2 > - < n >^2}{< n^2 >} \\ &= \frac{(1-(1-\xi)^r)(1+\xi r)}{\xi r} - 1, \end{aligned} \tag{S38}$$

whose contour plot in shown in the main text, Figure 5.