Research

Oikos 119: 71–80, 2010
doi: 10.1111/j.1600-0706.2009.17938.x,

Subject Editor: Bradford Hawkins. Accepted 23 June 2009

Inferring species abundance distribution across spatial scales

Tommaso Zillio and Fangliang He

T. Zillio (tommaso.zillio@gmail.com) and F. He, Dept of Renewable Resources, Univ. of Alberta, Edmonton, AB T6G 2H1, Canada.

A long-standing problem in ecology is to understand how the species–abundance distribution (SAD) varies with sampling scale. The problem was first characterized by Preston as the veil line problem. Although theoretical and empirical studies have now shown the nonexistence of the veil line, this problem has generated much interest in scaling biodiversity patterns. However, research on scaling SAD has so far exclusively focused on the relationship between the SAD in a smaller sampling area and a known SAD assumed for a larger area. An unsolved challenge is how one may predict species–abundance distribution in a large area from that of a smaller area. Although upscaling biodiversity patterns (e.g. species–area curve) is a major focus of macroecological research, upscaling of SAD across scale is, with few exceptions, ignored in the literature. Methods that directly predict SAD in a larger area from that of a smaller area have just started being developed. Here we propose a Bayesian method that directly answers this question. Examples using empirical data from tropical forests of Malaysia and Panama are employed to demonstrate the use of the method and to examine its performance with increasing sampling area. The results indicate that only 10-15% of the total census area is needed to adequately predict species abundance distribution of a region. In addition to species abundance distributions, the method also predicts well the regional species richness.

One of the oldest problems in ecology is to infer general properties of a large ecosystem by sampling only a small part of it. Of this, the most well known example is perhaps Preston's veil line. Preston (1948, 1962) hypothesized that the lognormal species–abundance distribution is a universal curve that gets progressively revealed with increasing sampling effort. Visually, the veil line 'shifts rigidly to the right' of the abundance plot. While the theory of the rigid veil line has been proven wrong (Dewdney 1998, Williamson and Gaston 2005), the problem has generated much interest in search for understanding the dependence of species–abundance distribution (SAD) on spatial scale (Nee et al. 1991, Gregory 1994, Green and Plotkin 2007).

Dewdney (1998) was the first one to establish a formal relationship between local SAD and regional SAD through a hypergeometric sample transformation function. He also showed that the shape of the SAD is retained for randomly distributed species. It is important to remember that a random sample (i.e. a survey on individuals taken at random over the whole area) and a local sample (i.e. a survey on individuals inside a specific sub-area) are not the same since species are generally not distributed randomly but clustered together due to dispersal limitation or habitat heterogeneity. Following this observation, Green and Plotkin (2007) extended the work of Dewdney by correcting for species aggregation. They showed that, as a result of aggregation, the sample SAD is skewed to both rare and abundant species. Random sampling (McKane et al. 2004) and hypergeometric sampling (Etienne and Alonso 2005) have also been used to construct species–abundance distributions of local neutral communities coupled with metacommunity.

The focus of all these studies is to infer species–abundance distribution across scales. However, it is important to distinguish two related but fundamentally different scaling problems: (1) sampling or 'downscaling', and (2) predicting or 'upscaling'. In the former problem the species abundance distribution is assumed to be known for the region of interest. In this context, downscaling is simply the construction of an SAD for a local area nested within the known region through sampling. In contrast, 'upscaling' assumes that the species-abundance is known only at a local scale. The goal is to extrapolate the regional (unknown) SAD from a known local SAD. Therefore, a downscaling problem involves developing a method to accurately sample from a known distribution, whereas an upscaling problem implies predicting a distribution from where the observed data could have been sampled from.

Upscaling biodiversity patterns lies at the heart of macroecology (Willis and Whittaker 2002) and advanced theories and methods are in urgent need for upscaling SAD. The existing literature has so far mostly focused on the downscaling problem: either to understand the effect of sampling intensity on the 'veil line' (Nee et al. 1991, Gregory 1994) or to sample local SAD from a metacommunity (McKane et al. 2004, Etienne and Alonso 2005, Green and Plotkin 2007). An exception is the recent development of a maximum entropy method used to

extrapolate regional species richness and abundance from small sampling areas (Harte et al. 2009). By the method, a logseries SAD is found for a community. Since the logseries distribution is spatially congruent, the SAD will remain logseries when extrapolated to a larger area (although the value of the parameter of the distribution will change).

The importance of the problem has also been well recognized in the literature as reflected by the closing remark of Green and Plotkin (2007): ''Our result on sampling may inform future research aimed at leveraging abundance and aggregation patterns measured at local scales to predict biodiversity patterns at larger, regional scales''. Leveraging over this knowledge to provide a constructive method of SAD prediction is the goal of the present article. We show that the solution to the upscaling problem requires that we have already solved the downscaling problem, since the sampling function is needed to proceed correctly to the extrapolation (see the section 'Sampling probability' below). The downscaling problem aims at finding a sampling function that explains the data at hand and to construct the local distribution from the regional one, while the upscaling problem uses that sampling function to extrapolate the data at a larger scale from the local scale.

When scaling up, new species that are not present in smaller areas will appear. A key step in upscaling SAD is to incorporate those new species to update the SAD for the larger area. As a result, our method not only leads to the construction of a regional SAD but also leads to an estimation of regional richness. Estimating richness itself has been a challenging problem. Many richness estimation methods have been developed so far (Chao 1984, Bunge and Fitzpatrick 1993, Magnussen et al. 2006, Shen and He 2008). Some of them are more sensitive to sample size than others. An important but not yet answered question is how much sample is needed for obtaining a reasonable estimation of regional richness. This study will also contribute to answering this problem.

In this study we propose a Bayesian method to reconstruct the species abundance of an area A_0 from the species abundance of a smaller area A_1 that is nested within A_0. We choose the widely used negative binomial (He and Gaston 2000, Plotkin and Muller-Landau 2002, He and Hubbell 2003, Green and Plotkin 2007) as a sampling function, but the method can be easily generalized to any sampling probability. The performance of the developed Bayesian method is evaluated using two large-scale stem-mapping plots of the tropical forests whose species–abundance distributions are completely known. With the increase of A_1, we expect that the estimated regional SAD for A_0 would approach the actual regional SAD. We will also determine the minimum sampling area needed for predicting the regional SAD.

Material and methods

Statistical framework

Suppose species abundance data are available for a local area A_1 that is nested within a region A_0. With the data, a discrete species–abundance distribution $\phi(n)$, $n>0$, can be constructed for A_1. Here $\phi(n)$ denotes the number of species with abundance n. In some cases, especially in theoretical studies, the distribution is generalized to include $n=0$, which represents those species that are missing from the study plot A_1 but are present in other part of the region A_0. It is important to realize that $\phi(0)$ depends on A_0, because changing the reference area will change the number of the species in A_0 that are not in A_1, and thus the value of $\phi(0)$.

We use P(n) to denote the relative species–abundance that is a probability mass function obtained by normalizing the species abundance $\phi(n)$. If we restrict the abundances to $n>0$, then $\phi(n)=SP(n)$, where S is the number of species present in A_1. If, on the other hand, we want to include the case $n=0$, then S becomes the number of species present in the regional area A_0.

Let P(n|N, a) be the probability that a species is represented by n individuals in A_1 given that the same species is represented by N individuals in A_0 and that A_1 is a fraction a of A_0 (i.e. $a=A_1/A_0$). We call P(n|N, a) the sampling probability. Denoting $\phi_0(n)$ as the number of species with abundance n in A_0 and $\phi_1(n)$ as the number of species with abundance n in A_1, we can obtain the average value of the latter from the former by:

$$\langle \phi_1(n) \rangle = \sum_{N>0} P(n|N, a)\, \phi_0(N) \quad (1)$$

This relation is linked to the species area relationship by noting that:
$S_i = \Sigma_{n>0}\, \phi_i(n),\ i = 0, 1$

Equation 1 is precisely the widespread downscaling formula that links local SAD $\phi_1(N)$ with regional SAD $\phi_0(N)$ (Dewdney 1998, McKane et al. 2004, Etienne and Alonso 2005, Green and Plotkin 2007). However, obtaining the species abundance in A_1 from the one in A_0 is not really helpful to us, since to measure empirically the species abundance in A_0 we are already forced to sample A_1. Our interest is the inverse: to predict the species abundance in A_0 from that of A_1. In the next sections, for the sake of explanation of the reconstruction method, we start by assuming that the sampling probability P(n|N, a) is known a priori. Later on we will discuss what approach should be taken when this sampling function is not known.

Bayes' rule

To achieve upscaling, we need to reverse Eq. 1 to express the regional SAD in terms of the local SAD. To do that, we need P(N|n) — the reverse of the previous sampling probability P(n|N). This reverse involves the conditional probabilities P(N|n) and P(n|N) and can be easily formulated by Bayes' rule:

$$P(A|B) = \frac{P(A)\, P(B|A)}{P(B)} \quad (2)$$

where A, B are arbitrary events. To describe the upscaling problem, the general Bayes' rule can be rewritten as:

$$P(N|n, a) = \frac{P(N|a)\, P(n|N, a)}{P(n|a)} \quad (3)$$

where the left side is the probability that a species has abundance N in A_0 given n individuals of that species are

present in A_1. On the right side P(n|N,a) is the sampling probability, $P(n|a) \equiv \Sigma_N P(N|a)\, P(n|N,a)$ is the normalization factor, and P(N|a) is the prior probability distribution for the species abundance in A_0 (we will further discuss this term below).

Equation 3 gives the probability for the abundance in A_0 for a single species that has abundance n in A_1. Our objective, however, is not the probability that a single species has abundance N, but the average number of species with abundance N in A_0. If we suppose that P (N|n) is the same for all the species, then the number of species with abundance N is just binomially distributed (since every species has the same probability to appear in that abundance class) with an average of SP(N|n), and a standard deviation of SP(N|n)(1 − P(N|n)), where S is the number of species in A_1.

However, in reality P(N|n) is different for different species (either because they have a different abundance n in A_1, or if we consider different sampling probability for different species). We thus need the mean and variance for a collection of random events with different probabilities $\{p_i\}$. In Section 1 of the Supplementary material Appendix 1 we show that the mean and variance are respectively $\Sigma_i pi$ and $\Sigma_i p_i(1-p_i)$. Our prediction for the average number of species per abundance class (i.e. the species abundance distribution) in A_0 can then be written as:

$$\phi_{0,\mathrm{pred}}(N) = \sum_n P(N|n)\, \phi_1(n) \tag{4}$$

with P(N|n) from Eq. 3.

However, the problem of Eq. 4 is immediately apparent: the sum $\Sigma_{N>0}\phi_{0,\mathrm{pred}}(N)$ gives S_1 (the number of species in A_1), while in reality it should be S_0. This means that we should take into account the species present in A_0 but missing in A_1 , otherwise our prediction for the SAD is biased and underestimates the number of rare species (the ones most likely to be missing from A_1). To solve the problem, we must have an independent estimation about the number of missing species ($\phi_1(0)$) in A_1. With this estimation, we can give, as a prediction for $\phi_0(N)$, the quantity $\phi_{0,\mathrm{pred}}(N) + P(N|0)\ \phi_1(0)$. We will describe an iterative system to solve the missing species problem after the next section.

Prior probability

Lets first discuss the prior probability term in Eq. 2 and 3. In the Bayesian framework, one can choose either a prior that expresses complete ignorance of any information (an 'uninformative' prior) or one that reflects the fact that some information is available before we proceed to the inference (an 'informative' prior).

In the case of inferring SAD, overwhelming empirical evidence has shown that widespread of species-abundance data display a lognormal type of distribution (Hubbell 2001, Williamson and Gaston 2005, McGill et al. 2007). It thus appears natural to use an informative lognormal prior:

$$P(N) = \frac{1}{\sqrt{2\pi}N\sigma} \exp - \frac{(\log N - \log\mu)^2}{2\log(\sigma)^2}$$

It is essential to note that the values for the parameters μ and σ of the lognormal are dependent on sampling area and they follow an approximate power relationship with area A as:

$$\mu = \mu_0\, A^{-\theta_\mu}$$

$$\sigma = \sigma_0\, A^{-\theta_\sigma}$$

The values for μ and σ at A_0 can be estimated as follows: 1) divide A_1 into smaller subareas, 2) fit the above relationships to the subareas, and 3) obtain the μ and σ for A_0 by substituting A_0 into the fitted models. This procedure, admittedly not accurate if used alone to predict the species–abundance at a greater area, nevertheless provides a good prior for the Bayesian procedure we described above.

The procedure and results when using an uninformative prior instead of the lognormal one are shown in the Supplementary material Appendix 1.

Missing species prediction

To solve the problem of missing species, we propose an iterative method that predicts the total number of species and the corresponding species abundance distribution. Our 'zeroth order' prediction is Eq. 4 that ignores missing species:

$$\phi^{(0)}_{0,\mathrm{pred}}(N) = \sum_n P(N|n)\, \phi_1(n) \tag{5}$$

where the superscript denotes the first step of the iterative process, and we drop the variable a for simplicity of notation since a is fixed for a given calculation. Starting from the zeroth order $\phi^0_{0,\mathrm{pred}}(N)$ and pretending that this is the true species–abundance at the larger area A_0, we can calculate the number of species that are missing in the smaller area A_1:

$$\phi^{(0)}_1(0) = \sum_N P(0|N)\, \phi^{(0)}_{0,\mathrm{pred}}(N) \tag{6}$$

This number allows us to write a new ('first order') prediction for the species–abundance distribution in A_0:

$$\phi^{(1)}_{0,\mathrm{pred}}(N) = \sum_{n>0} P(N|n)\, \phi_1(n) + P(N|0)\, \phi^{(0)}_1(0) \tag{7}$$

$$= \phi^{(0)}_{0,\mathrm{pred}}(N) + P(N|0)\, \phi^{(0)}_1(0) \tag{8}$$

which in turn allows us to write a new prediction for the number of missing species in A_1:

$$\phi^{(1)}_1(0) = \sum_N P(0|N)\, \phi^{(1)}_{0,\mathrm{pred}}(N)$$

This iterative process can be expressed as a system of two recursive equations for every $i > 0$:

$$\phi^{(i)}_{0,\mathrm{pred}}(N) = \phi^{(0)}_{0,\mathrm{pred}}(N) + P(N|0)\, \phi^{(i-1)}_1(0) \tag{9}$$

$$\phi^{(i)}_1(0) = \sum_N P(0|N)\, \phi^{(i)}_{0,\mathrm{pred}}(N) \tag{10}$$

In Section 2 of the Supplementary material Appendix 1 it is shown that for $i \to \infty$ the system approaches a definite solution:

$$\phi^{(\infty)}_{0,\mathrm{pred}}(N) = \phi^{(0)}_{0,\mathrm{pred}}(N) + P(N|0)\, \phi^{(\infty)}_1(0) \tag{11}$$

$$\phi_1^{(\infty)}(0) = \frac{\sum_{N'} P(0|N')\,\phi_{0,\mathrm{pred}}^{(0)}(N')}{1 - \sum_{N'} P(N'|0)\,P(0|N')} \quad (12)$$

Equation 11 is the prediction for the species abundance distribution that takes into account the missing species. Equation 12 is the prediction for the number of missing species, so that the predicted total number of species at A_0 is:

$$S_{0,\mathrm{pred}} = S_1 + \phi_1^{(\infty)}(0) \quad (13)$$

Sampling probability

The implementation of the above iteration requires a sampling probability P(n|N, a) for a real ecosystem. As explained in the Introduction, this sampling function is a key device for investigating sampling effect on species–abundance distribution. Several functions can be used as P(n|N, a). As a first choice one could simply use a binomial distribution (random sampling):

$$P(n|N,a) = \binom{N}{n} a^n (1-a)^{N-n}$$

However, it is well known that ecosystems do not in general obey a random placement model. Instead, their sampling probability presents some 'clustering'. To capture the nature of the clustering, an appropriate function is the negative binomial distribution (He and Gaston 2000, Plotkin and Muller-Landau 2002, Green and Plotkin 2007):

$$P(n|N,a) = \begin{cases} \dfrac{\Gamma(k+n)}{z\Gamma(k)n!}\left(\dfrac{aN}{aN+k}\right)^n\left(\dfrac{k}{aN+k}\right)^k & n \le N \\ 0 & n > N \end{cases}$$

where the parameter k is a clustering parameter, and normalization is ensured by

$$Z = \sum_{n \le N} \frac{\Gamma(k+n)}{z\Gamma(k)n!}\left(\frac{aN}{aN+k}\right)^n\left(\frac{k}{aN+k}\right)^k$$

for upper tail truncation. The truncation is needed for application because N in the negative binomial distribution is assumed to be infinite.

It is widely known that k is not scale invariant but depends on both A_0 and A_1. Two empirical formula have been derived to extrapolate k across scale based on the tree distributions from the 50 ha Pasoh plot of Malaysia and the 50 ha BCI plot of Panama. The first one given by Plotkin and Muller-Landau (2002) is $k(A_1) = 0.8604 + 0.002923A_1^{0.5450}$ for Pasoh plot, where A_1 is in square meters. The second is derived by He and Hubbell (2003) directly in terms of $a = A_1/A_0$: $k(A_1) = k_0\,(A_1/A_0)^{0.55}$ for the Pasoh and BCI plots. These two parameterizations are both obtained with the assumption that A_0 is constant, which differs from our application in such way that in our case either A_1 or a is held constant but A_0 varies. This is because for this study we do not possess the data of area A_0, but only the data at area A_1 or smaller, thus preventing us from estimating the parameters of these relationships. To overcome the problem, we propose a new parameterization at fixed a:

$$k = k_a A_0^{-\theta_a} \quad (14)$$

This model fits the k data quite well for BCI and Pasoh (Fig. 1). This parameterization is needed when applying the reconstruction method to empirical data.

When extrapolating SAD from A_1 to A_0, information about species distribution, thus k, at A_0 is not available. Equation 14 is used to infer k for A_0,. This is done by fitting the equation to subareas of A_1 following the steps: 1) choose a subarea of A_1, called it as A_0'; 2) choose another subarea A_1' that is nested within A_0', with the size of $A_1' = aA_0'$; 3) calculate k by maximum likelihood between A_0' and A_1'; 4) repeat steps 1–3 by varying A_0'; and 5) finally fit Eq. 14 to the k estimated from the maximum likelihood (Fig. 1). The parameterized Eq. 14 obtained from the last step can then be used to estimate k for any area.

As a summary, the whole procedure of reconstructing SAD is outlined as follows:

1. Enumerate species abundances in A_1 and count empirical species-abundance distribution $\phi_1^{(n)}(0)$. Choose a 'target' area A_0 for extrapolation.
2. Find the value of the parameter k in the negative binomial for the given value of $a = A_1/A_0$ by fitting Eq. 14 to the subareas of A_1 and extrapolating to A_0.
3. Calculate the zeroth-order prediction for the SAD using Eq. 5, by ignoring the missing species.
4. Calculate the number of missing species $\phi_1^{(\infty)}(0)$ using Eq. 12.
5. Calculate the prediction for the SAD using Eq. 11 by including the missing species $\phi_{0,\mathrm{pred}}^{(\infty)}(N)$

Data sets

We applied the Bayesian method to infer tree species-abundance distribution, respectively, for the 50 ha BCI plot (1990 census) of Panama and the 50 ha Pasoh plot (1987

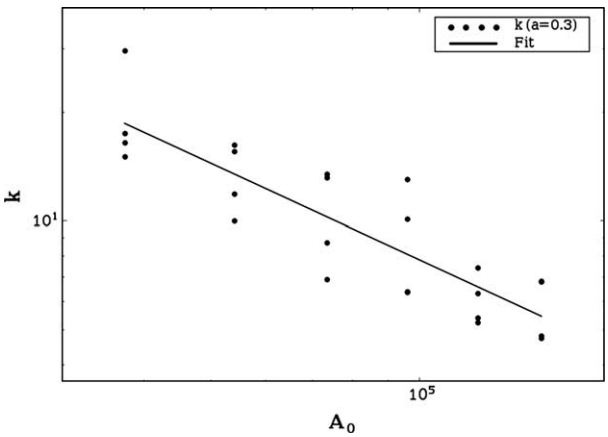

Figure 1. Determination of parameter k of the negative binomial distribution for the BCI plot. The figure shows the values of k for different samples when a is kept fixed ($a = 0.3$) and A_0 varies. The various points at each value of A_0 are the value of k for several samples of size $A_1 = aA_0$, calculated by maximum likelihood. The line is a least squares fit of Eq. 14, with $k_a = 2.1E5$ and $\theta_a = 0.89$, and with $r^2 = 0.867$. k's in Table 1 are obtained in this way for different a for BCI and Pasoh plots. A_0 is measured in square meters.

census) of Malaysia (<www.ctfs.si.edu>). We used the abundances for all the trees and saplings with diameter at breast height (dbh) ≥ 1 cm. To check that our reconstruction was working properly, we selected a rectangular region of area $A_1 = aA_0$ with a = 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5 located at the center of the whole 50 ha plot and we pretended that A_1 was the only area sampled, with the goal to use the reconstruction method to predict the species abundance distribution for A_0. The lognormal prior was used for all reconstruction (results for the uninformative geometric prior are reported in Section 3 of the Supplementary material Appendix 1 for comparison). The sampling probability was assumed to be a negative binomial, and the parameter k was calculated by Eq. 14; the parameters k_a and θ_a used in Eq. 14 were obtained by fitting the model to the k estimated by maximum likelihood method for subareas of A_1. The reconstructions were also performed, for comparison, using the 'true' value of k, and are reported in Section 4 of the Supplementary material Appendix 1, the result being that the extrapolation of k do not introduce any significant error.

Due to a considerable interest in surveying entire 1564 ha Barro Colorado Island (Stephen Hubbell pers. comm.), here we performed an exercise to predict SAD for $A_0 = 100$, 150, 200, 300 and 500 ha, based on the 50 ha BCI plot. Note, as it is shown below, at the present stage our method cannot give reliable estimates for larger A_0, e.g. 1000 ha, of which the 50 ha BCI plot is only 5%.

Measures of goodness of prediction

To asses the goodness of our prediction we calculated the log likelihood of the prediction against the data, i.e. if we denote the empirical abundances in A_0 with $\{N_i\}$, i = 1, . . ., S, and our prediction is $\phi^{(\infty)}_{0,\mathrm{pred}}(N)$ from Eq. 11, then the log likelihood of our prediction is:

$$L_{\mathrm{pred}} = \sum_{i=1}^{S} \ln\left(\phi^{(\infty)}_{0,\mathrm{pred}}(N_i)\right) \tag{15}$$

These results are reported in Table 1. Note that there is no adjustable parameter in the prediction $\phi^{(\infty)}_{0,\mathrm{pred}}(N)$ (the parameter k has already been fixed by fitting Eq. 14 as explained above).

As a benchmark, we also calculated the log likelihood for the maximum likelihood lognormal fit to the data:

$$L_{\mathrm{logn}} = \max_{\mu,\sigma} \sum_{i=1}^{S} \ln\left(\frac{1}{N_i \sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln N_i - \mu)^2}{2\sigma^2}\right)\right) \tag{16}$$

These results are reported in the caption of Table 1. To compare the two goodness of fit we used the F statistics and the χ^2 asymptotic distribution of the likelihood.

Results

Results for the reconstruction of the species-abundance distribution for the BCI and Pasoh plots (Eq. 11) are shown in Fig. 2 and 3 and in Table 1. As a benchmark of the goodness of the predictions provided by our method, we fitted a lognormal curve to the species abundance of the entire 50 ha BCI and Pasoh plots, and calculated the log likelihood for both the prediction and the lognormal fit. In both cases the likelihood of the prediction tends to increase with a, as expected. For the Pasoh plot, the reconstruction at large A_1 ($a \geq 0.2$) even outperforms the direct fit of lognormal to the data.

Bayesian methods take account both prior knowledge and data. If there is little data, the method's output is dictated by the prior. The influence of the prior decreases with the increase of data. As a result, our method tends to give a prediction close to the prior when the input data are scarce (i.e. when a is small). This, along with the unreliability of the extrapolation of the parameters of the priors for small a's, explains the poor performance of the method when a = 0.05.

Our prediction for the number of species is compared in Table 1 and in Fig. 2 and 3 with the Chao estimator (Chao

Table 1. Reconstruction of the SAD of the 50 ha BCI and Pasoh forest plots starting from a subarea $A_1 = aA_0$ where $A_0 = 50$ ha (Fig. 2). k: predicted value for parameter k of the negative binomial distribution, extrapolated from data at scales smaller than A_1. S_1: number of observed species in A_1. $S_{0,\mathrm{pred}}$: predicted number of species in A_0 (Eq. 13). Chao: Chao's estimator for the number of species. Likelihood: log likelihood of the reconstruction curve against the data (Eq. 15), to be confronted with the log likelihood for a lognormal fit to the entire BCI and Pasoh data (Eq. 16), ln L = −2054.5 in BCI and ln L = −5404.9 in Pasoh. The boldface likelihood indicates nonsignificant difference from the lognormal fit; an asterisk indicates that the reconstruction is significantly better than the lognormal fit. Significance is calculated using F statistic and the χ^2 asymptotic distribution of the likelihood. The actual number of species in BCI is 305 and 817 in Pasoh.

Plot	a	k_a	θ_a	r^2_k	k	'true' k	μ	σ	S_1	$S_{0,\mathrm{pred}}$	Chao	likelihood
BCI	0.05	5.4E2 ± 1.8E2	0.49 ± 0.78	0.17	0.91 ± 13.6	1.84	8185	2.97	217	219	236	−4500.6
	0.1	1.1E1 ± 0.3E1	0.062 ± 0.103	0.13	4.89 ± 1.33	2.18	10.0	10.5	233	288	242	**−2048.8**
	0.15	1.0E2 ± 0.1E2	0.25 ± 0.13	0.34	3.72 ± 1.35	2.25	19.6	9.90	238	272	246	**−2049.8**
	0.2	9.3E3 ± 0.01E3	0.64 ± 0.22	0.76	2.06 ± 1.56	2.44	24.9	9.86	250	284	270	**−2049.8**
	0.3	9.7E4 ± 0.01E4	0.82 ± 0.42	0.80	2.17 ± 1.86	3.33	33.0	10.7	261	286	278	**−2049.4**
	0.4	9.8E4 ± 0.02E4	0.80 ± 0.44	0.70	2.79 ± 1.72	4.65	47.0	9.43	272	290	281	**−2050.4**
	0.5	2.0E-1 ± 0.1E-1	−0.36 ± 0.59	0.67	25.6 ± 1.4	7.56	62.2	9.33	277	288	282	**−2049.8**
Pasoh	0.05	1.4E1 ± 0.5E1	0.16 ± 0.15	0.21	1.61 ± 1.75	1.44	3222	1.62	598	598	639	−29848
	0.1	1.0E1 ± 0.3E1	0.09 ± 0.09	013	3.10 ± 1.31	1.65	8.76	6.15	663	859	688	−5426.3
	0.15	2.9E0 ± 2.8E0	0.05 ± 0.09	0.05	6.00 ± 1.26	2.08	11.2	8.20	695	808	711	**−5400.2**
	0.2	3.0E1 ± 0.4E1	0.14 ± 0.13	0.22	5.06 ± 1.31	2.69	13.6	9.06	720	812	743	**−5397.6***
	0.3	1.0E3 ± 0.03E3	0.41 ± 0.30	0.48	4.63 ± 1.57	4.45	26.0	8.62	751	808	766	**−5391.1***
	0.4	1.4E4 ± 0.01E4	0.61 ± 0.40	0.59	4.87 ± 1.66	7.03	42.7	7.30	768	803	782	**−5390.1***
	0.5	3.3E6 ± 0.2E6	1.05 ± 1.54	0.88	3.06 ± 2.33	10.2	46.4	8.39	781	813	800	**−5390.4***

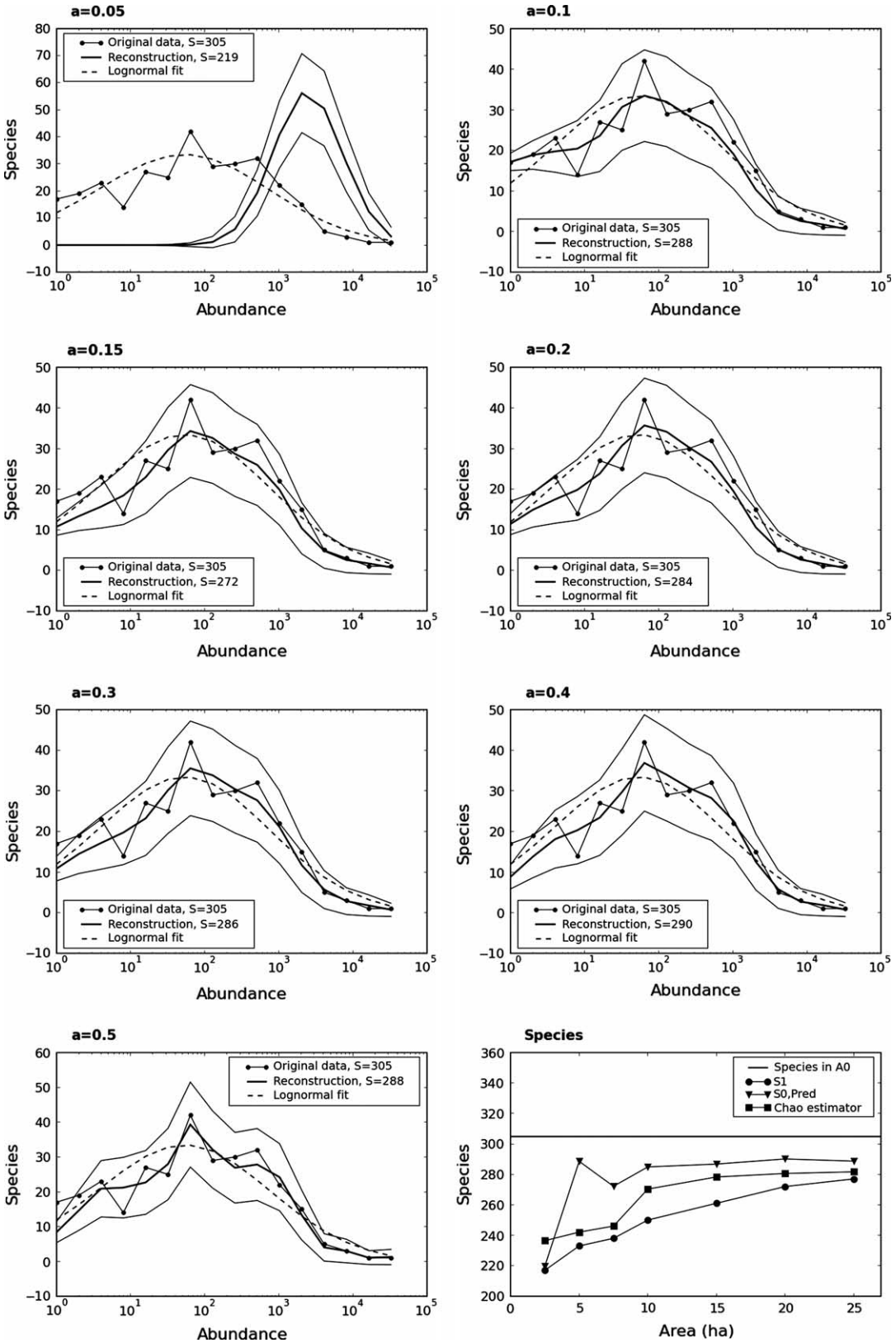

Figure 2. a = 0.05, . . . , 0.5: reconstruction (solid line) with a lognormal prior of the SAD of the 50 ha BCI forest plot starting from a subarea of extension A1 = aA_0 with various values of a where $A_0 = 50$ ha. Thin lines are the 95% Bayesian standard errors for the reconstruction; dots represent the real species abundance in the 50 ha plot, and the dashed line is a lognormal fit to the SAD at 50 ha. Abundance classes are logarithmically binned. The actual and predicted number of species are indicated in the legend. For the value of k and goodness of fit see Table 1. Species: species prediction performance. The horizontal line shows the true number of species at $A_0 = 50$ ha. S_1 is the true number of species present at $A_1 = aA_0$, $S_{0,\,\mathrm{pred}}$ is the prediction given by Eq. 13, and Chao's estimator is plotted for comparison.

1984) $S_0 = S_1 + (f_1)^2/2f_2$, where f_1 and f_2 are the number of singletons and doubletons in A_1, respectively.

The application of our method to predict SADs of Barro Colorado Island at scales larger than 50 ha results in Fig. 4. How well our predictions do? The question cannot be answered before we have species abundance data for the entire island. However, we may partially assess the goodness of prediction by examining the species–area curve. From species checklist, Condit et al. (2005) estimate there are 436 tree/shrub species (dbh $\geq$1 cm) in the entire island. Judged by the species–area curve constructed from the 50 plot and the 436 species, our predictions seems not to be unreasonable. The triphasic species–area curve shown in Fig. 4 is consistent to the observation of Hubbell (2001, p. 199).

Discussion

With few exceptions, information on biodiversity of a region has to be inferred from samples. The essence of the problem is to predict diversity across scales. Numerous methods have been developed to estimate population sizes (Seber 1982), species richness of animals and plants (Bunge and Fitzpatrick 1993, Colwell and Codington 1994), and species composition (Condit et al. 2002, Chao et al. 2005) at varying scales. Surprisingly, methods to explicitly extrapolate the species abundance distribution across scales are absent from the literature, although SAD is indisputably the most well studied diversity pattern and it has long been recognized to vary with sampling scale, both theoretically and empirically (Preston 1948, Nee et al. 1991, Gregory 1994, Dewdney 1998, Williamson and Gaston 2005, Green and Plotkin 2007). To the best of our knowledge, the only exception to this last statement is Harte et al. (2009) who proposes a method to upscale the species richness based on the maximum entropy framework that also predicts the SAD. Because Harte et al. method is parametric, they can only predict logseries-shaped SADs. Our method, on the other hand, is non-parametric and can predict SADs of arbitrary shape.

For the first time, we have proposed a non-parametric method for upscaling regional SAD from smaller samples. The Bayesian method takes a reasonable prior (lognormal) and assumes a negative binomial spatial distribution of species. Both these choices are guided by the information already available on the system at hand. The method works very well as shown by the predictions of BCI and Pasoh SAD's (Fig. 2, 3). With as small as 10–15% of sampling area, we can adequately predict the SAD's of both plots. With a larger (20% or higher) sampling area, our method can capture the detailed shape of the SAD and reconstructs the full SAD data. In Pasoh plot, the prediction works even more accurately than a direct lognormal fit. This is a conservative statement since the lognormal is fitted by maximum likelihood, so we in fact give some advantage to the fit in the comparison. This is remarkable if we mention that many richness estimators require 60% or even more samples to obtain a reasonable estimate of the true species richness for BCI and Pasoh plots (Shen and He 2008); keep in mind richness is a single number, not a spectrum of distribution like SAD which is much more difficult to predict. As shown in Fig. 4, our method also does a very good job in estimating richness. With the increase of sampling area, the estimated richness converges to true richness very quickly. Only 15–20% of sampling area is needed for a reasonable estimation of richness. These results are not surprising if one notices that, for instance, the 15% sample plot contains 78% of the total species in BCI and 85% species in Pasoh, thus providing enough information to reconstruct the whole species abundance. Since it is not guaranteed that the situation is the same in other cases, it is possible that when our method is applied to a different ecosystem the threshold of 15% sample won't hold.

Unlike other richness estimators (jackknifing, bootstrapping, Chao 1984, Chao et al. 2005 and Shen and He 2008), our current method can only handle data from single quadrat due to the computational requirement of the method. This means we likely missed some very rare species with sample size smaller than 50%. That is why the estimated richness is somewhat smaller than the actual richness. This problem can only be solved by increasing sample size.

The major shift in our reconstructions occurs between 5 and 10% sample due to the property of the Bayesian method: the method is dictated by prior probability when only a little data is available but the importance of prior reduces with the increase of data. Here the prior probability is a lognormal whose parameters have been extrapolated as was explained above. Therefore, the Bayesian method is bound to give imprecise results when the sample is small.

Although the method remains to be tested for other ecosystems than tropical forests, there are good reasons to be confident because species are widely discovered to be aggregated for the negative binomial sampling function to hold (He and Gaston 2003). What needs to be readjusted is perhaps the lognormal prior probability because not every ecosystem is known to follow lognormal distribution (Williamson and Gaston 2005, McGill et al. 2007). An uninformative prior can also be used, at the expense of accuracy in the reconstruction of the rare species part of the SAD (Supplementary material Appendix 1).

If in case the negative binomial sampling is not an appropriate choice, how one may find an appropriate sampling function? A common feature of Bayesian methods is that their performance improves with the available input information (in our case, with the value of a) only if their underlying assumptions are correct (or at least approximately correct). If the underlying assumptions are not correct, increasing a will actually sway the predictions away from the desired result. In our case the monotonic increase of the likelihood with a is a testimony that the negative binomial sampling is a good choice. If the same reconstructions were attempted with a different sampling (e.g. assuming random sampling) the likelihood will decrease with increasing a (not shown here). Thus in systems other than tropical forests one should set up a reconstruction scheme like the one we performed in Fig. 2 and 3, i.e. with the results known a priori, and test different sampling assumptions to gauge their validity by their response to varying a.

In this study we proposed an approximated parameterization, Eq. 14, for extrapolating k using data at areas smaller than A_1. Theoretically, k should be proportionally (not as a power law) increased with sampling area (Johnson and Kotz 1969). However, this relationship has rarely

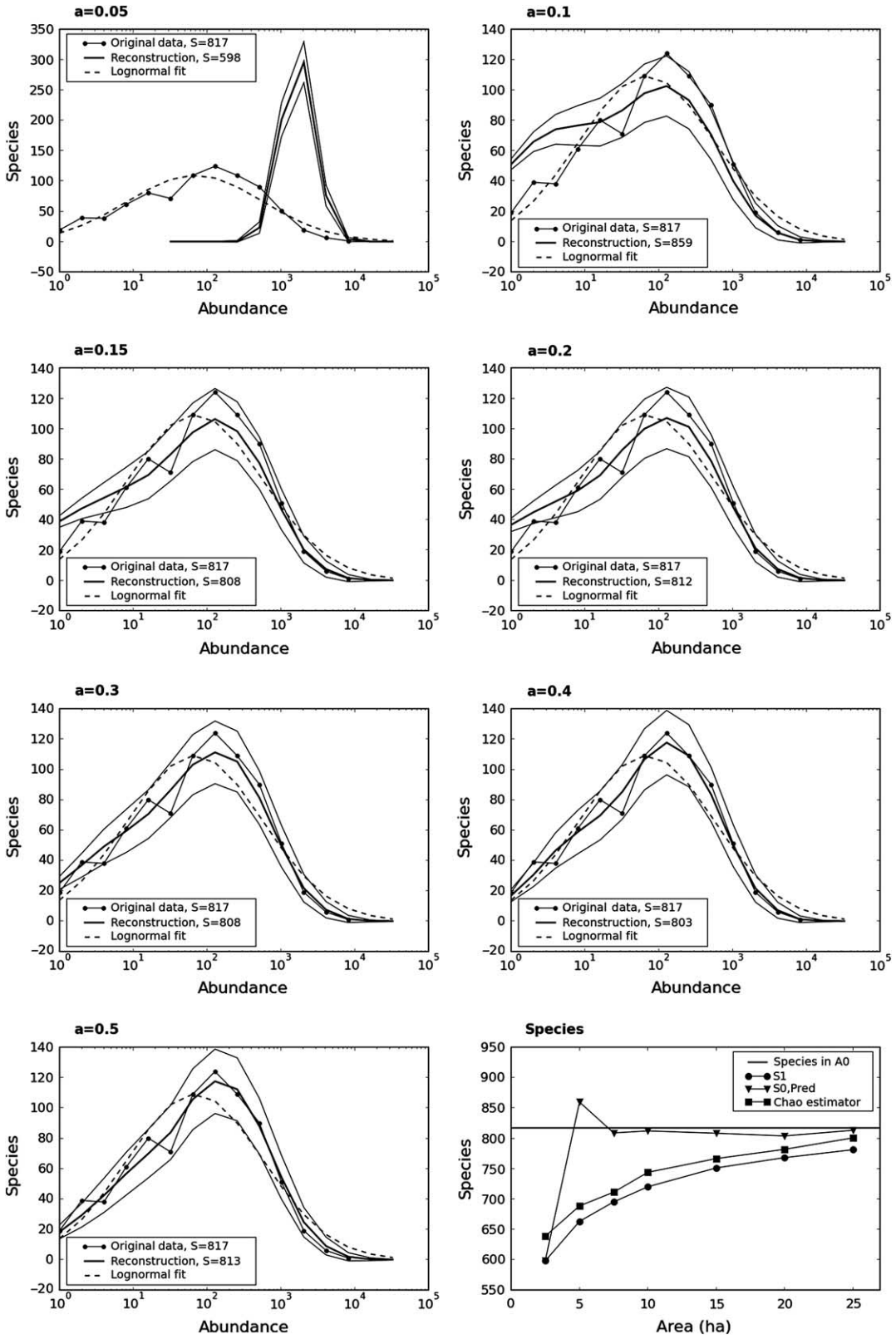

Figure 3. a = 0.05, . . . , 0.5: reconstruction (solid line) with a lognormal prior of the SAD of the 50 ha Pasoh forest plot starting from a subarea of extension $A_1 = aA_0$ with various value of a where $A_0 = 50$ ha. Thin lines are the 95% Bayesian standard errors for the reconstruction, dots represent the real species abundance in the 50 ha plot, and the dashed line is a lognormal fit to the SAD at 50 ha. Abundance classes are logarithmically binned. The actual and predicted number of species are indicated in the legend. For the value of k and goodness of fit see Table 1. Species: species prediction performance. The horizontal line shows the true number of species at $A_0 =$ 50 ha. S_1 is the number of species present at $A_1 = aA_0$, $S_{0,\mathrm{pred}}$ is the prediction given by Eq. 13, and Chao's estimator is plotted for comparison.

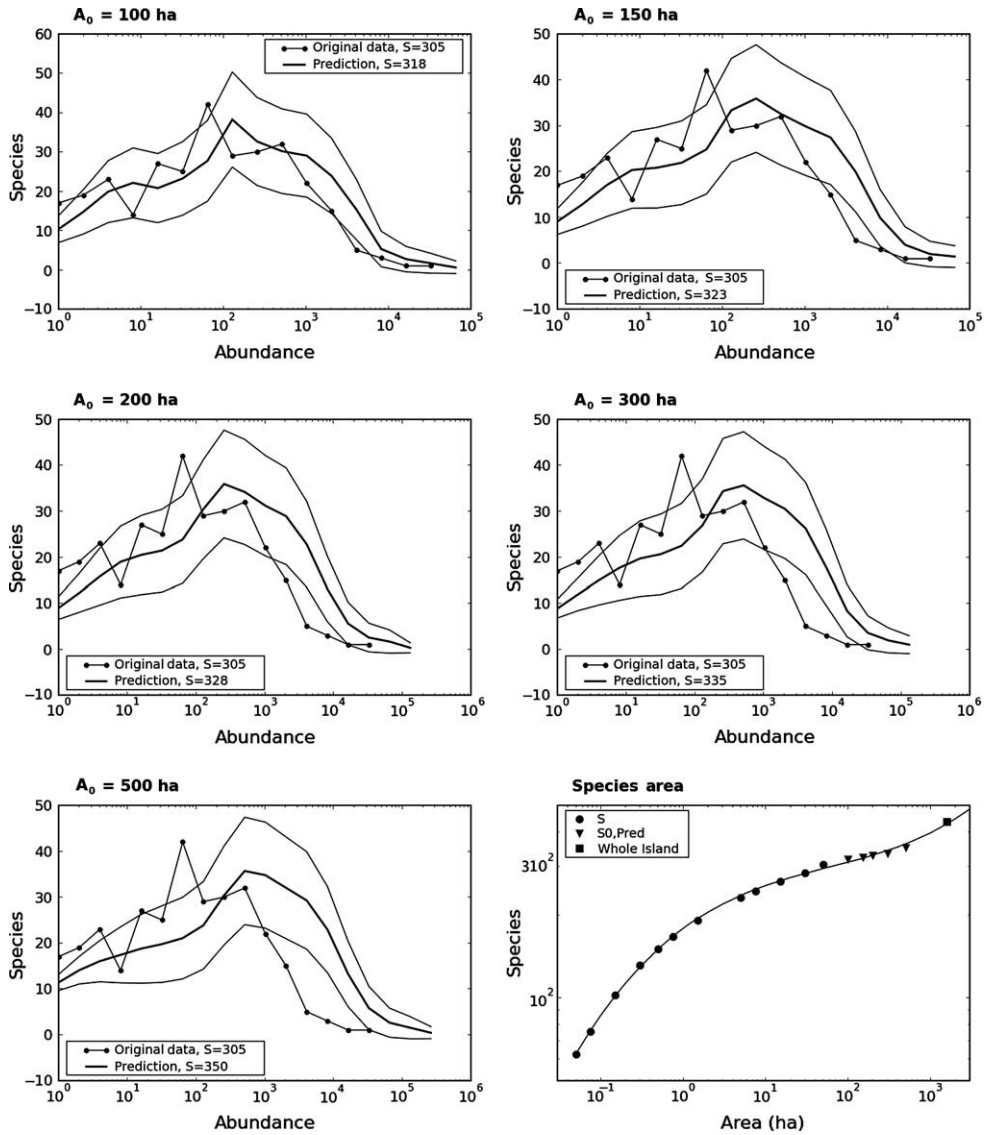

Figure 4. $A_0 = 100, \ldots, 500$ ha: Species–abundance prediction (solid line) with a lognormal prior for BCI for areas greater than the the area actually sampled (50 ha). Dots represent the data of the 50 ha plot. Species–area: the dots are the observed species–area within the 50 ha plot (S), the triangles are the predictions from our algorithm ($S_{0,\mathrm{pred}}$) and the square is the total number of species present on the whole island. The smooth curve is the fit of a third-order polynomial.

found to describe empirical distributions well. Empirical studies instead have shown that the power law is a reasonable model for parameterizing the relationship between k and sampling area (Plotkin and Muller-Landau 2002, He and Hubbell 2003). These power law models assume A_0 fixed with varying A_1, while our Eq. 14 was proposed to deal with the situation that keeps either a or A_1 fixed with varying A_0. The results show that the method works well (Fig. 2, 3), and the extrapolation of the correct value of k is not a major source of error due to the relative insensitivity of our method to k, as seen by performing the reconstructions using the 'true' value of k. By all means, when the value of k becomes larger than about 5, the resulting negative binomial is almost indistinguishable from the random Poisson distribution (theoretically, when $k \to \infty$ the negative binomial converges to Poisson). This means that the higher the value of k, the less sensitive the model is to the inaccuracy in the determination of k.

As a first approximation, we assumed that k was the same for all the species in a given community. In reality, however, k has different values across species even at the same scale, depending on the level of aggregation or 'clustering' of each species. This is likely a source of inaccuracy in our prediction. Because an enormous additional computational load is required to perform extrapolation with a different sampling probability for every species, we currently are unable to show how much more improvement we may gain on the prediction (Fig. 2, 3) we have already achieved by considering different k for different species. Although in principle this can be done, the computation involved will be very complicated (in particular, Eq. 4 will not be valid any more because different $P(N|n,a)$ is needed for different species, and the results the section 'Missing species prediction' need to be generalized to different sampling probabilities). In this study, the computational time spent on the calculation of the

sampling probability for various values of n and N was already heavy; when this probability is the same for every species, the results can be stored in a vector and recycled for every species. This approach is not possible with a different value of k for every species. At present this is the major limitation of our method and the one that is most likely to give an improvement in prediction if solved. As a caveat, it is worthwhile to note that although k is assumed to be the same across species in a plot, k is not assumed scale-invariant in our study. Instead, Eq. 14 and its estimation procedure reveal that k varies with both A_0 and A_1 (not just a function of a). What we assume is that at a given scale k is the same for every species.

In addition to considering different k for different species, another future improvement can be to consider data from multiple samples rather than a single sample as formulated in this study, taking into account the species similarity decay with distance between samples. While we recognize the limitations of the current method, this first ever effort for predicting regional SAD from local data works reasonably well. More importantly, our formulation offers a useful approach for sampling species-abundance distribution.

Acknowledgements – The authors thank Guillaume Blanchet for constructive comments that improved this study. The Center for Tropical Forest Science generously provided the BCI and Pasoh data. The large-scale forest plot at Pasoh Forest Reserve is an ongoing project of the Malaysian Government, initiated by the Forest Research Institute Malaysia through its Director General, and under the leadership of N. Manokaran, P S. Ashton and S. P. Hubbell. The work was supported by the National Excellent Centre for the Sustainable Forest Management Network of Canada and the Natural Sciences and Engineering Research Council of Canada.

References

Bunge, J. and Fitzpatrick, M. 1993. Estimating the number of species: a review. – J. Am. Stat. Ass. 88: 364–373.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.

Chao, A. et al. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. – Ecol. Lett. 8: 148–159.

Colwell, R. and Codington, J. 1994. Estimating terrestrial biodiversity through extrapolation. – Philos. Trans. R. Soc. Lond. 345: 101–118.

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

Condit, R. et al. 2005. Geographic ranges and β-diversity: discovering how many tree species there are where. – Biol. Skr. 55: 57–71.

Dewdney, A. K. 1998. A general theory of the sampling process with applications to the ''veil line''. – Theor. Popul. Biol. 54: 294–302.

Etienne, R. S. and Alonso, D. 2005. A dispersal-limited sampling theory for species and alleles. – Ecol. Lett. 8: 1147–1156.

Green, J. L. and Plotkin, J. B. 2007. A statistical theory for sampling species abundance. – Ecol. Lett. 10: 1037–1045.

Gregory, R. D. 1994. Species abundance patterns of british birds. – Proc. Bio. Sci. 257: 299–301.

Harte, J. et al. 2009. Biodiversity scales from plot to biomeswith a universal species–area curve. – Ecol. Lett. 12: 789–797.

He, F. and Gaston, K. J. 2000. Estimating species abundance from occurrence. – Am. Nat. 156: 553–559.

He, F. and Gaston, K. J. 2003. Occupancy, spatial variance, and the abundance of species. – Am. Nat. 162: 366–375.

He, F. and Hubbell, S. P. 2003. Percolation theory for the distribution and abundance of species. – Phys. Rev. Lett. 91: 198103.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.

Johnson, N. L. and Kotz, S. 1969. Discrete distributions. – Houghton Mifflin.

Magnussen, S. et al. 2006. An assessment of sample-based estimators of tree species richness in two wet tropical forest compartments in Panama and India. – Int. For. Rev. 8: 417–431.

McGill, B. et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. – Ecol. Lett. 10: 995–1015.

McKane, A. et al. 2004. Analytic solution of Hubbell's model of local community dynamics. – Theor. Popul. Biol. 65: 67–73.

Nee, S. et al. 1991. Lifting the veil on abundance patterns. – Proc. Bio. Sci. 243: 161–163.

Plotkin, J. B. and Muller-Landau, H. C. 2002. Sampling the species composition of a landscape. – Ecology 83: 3344–3356.

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

Preston, F. W. 1962. The canonical distribution of commonness and rarity. – Ecology 43: 185–215.

Seber, G. A. 1982. The estimation of animal abundance and related parameters. – Griffin.

Shen, T.-J. and He, F. 2008. An incidence-based richness estimator for quadrats sampled without replacement. – Ecology 89: 2052–2060.

Williamson, M. and Gaston, K. J. 2005. The lognormal distribution is not an appropriate null hypothesis for the species–abundance distribution. – J. Anim. Ecol. 74: 409–422.

Willis, K. J. and Whittaker, R. J. 2002. Species diversity – scale matters. – Science 295: 1245–1248.

Supplementary material (available online as Appendix O17938 at www.oikos.ekol.lu.se/appendix). Appendix 1.