Oikos

O17938
Zillio, T. and He, F. 2010. Inferring species abundance distribution across spatial scales. – Oikos 119: 71–80.

Appendix 1

Note: references to equations in the main text are numbered normally, whereas equations in the supplementary material are numbered with the prefix "SM".

1 Mean and Average for the Multiprobability Binomial

Let us have a collection of N heterogeneous binary events, each one associated with a probability p_i, $i = 1, \ldots, N$. We want to calculate on average how many of those events will happen, i.e. if we define I_i such as $I_i = 1$ when the i-th event is successful (with probability p_i) and $I_i = 0$ when the i-th event is not successful (probability $1 - p_i$), then we want the average and the standard deviation of the quantity $I = \sum_i I_i$. Note that if $p_i = p$ for all i, then we should obtain the standard formula for the binomial distribution.

Let $\mathbf{a}$ be a vector of 0's and 1's of length N, let a_i be its elements and let $\mathbf{a}_{(i)}$ be the vector of length $N - 1$ obtained by eliminating the element a_i. Let $G_N \equiv \langle \sum_i I_i \rangle$ for N variables, where the angular brackets denote the average over the p_i's. We will find G_N as a

function of G_{N-1} by summing only on the last variable. To do this, we write G_N explicitly:

$$G_N = \sum_{\mathbf{a}} \left(\sum_{i=1}^{N} a_i \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right)$$

where $\sum_{\mathbf{a}}$ is a sum over all the possible vectors $\mathbf{a}$. Now we rewrite it by isolating the sum over the last variable a_N:

$$\begin{aligned}
G_N &= \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\sum_{i=1}^{N} a_i \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) \\
&= \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\left(a_N + \sum_{i=1}^{N-1} a_i \right) \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) \\
&= \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(a_N \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) + \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\sum_{i=1}^{N-1} a_i \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) \\
&= \sum_{a_N} \left(a_N p_N^{a_N} (1-p_N)^{1-a_N} \right) + \sum_{\mathbf{a}_{(N)}} \left(\sum_{i=1}^{N-1} a_i \prod_{i=1}^{N-1} p_i^{a_i} (1-p_i)^{1-a_i} \right) \\
&= p_N + G_{N-1}
\end{aligned}$$

Applying this equation recursively, we obtain:

$$\left\langle \sum_i I_i \right\rangle = \sum_i p_i$$

When $p_i = p$ we obtain the same expression as the binomial distribution: $\langle \sum_i I_i \rangle = Np$.

We proceed in the same way to calculate the variance of $\sum_i I_i$; let V_N be the variance for N variables, and express it in terms of V_{N-1}. The formula for V_N is:

$$V_N = \sum_{\mathbf{a}} \left(\left(\sum_{i=1}^{N} a_i - G_N \right)^2 \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right)$$

and again isolating the sum over the last variable:

$$V_N = \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\left(\sum_{i=1}^{N} a_i - G_N \right)^2 \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right)$$

$$= \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\left((a_N - p_N) + \left(\sum_{i=1}^{N-1} a_i - G_{N-1} \right) \right)^2 \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right)$$

This last sum can be divided into 3 parts by developing the square inside it. The first part is:

$$\begin{aligned} \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left((a_N - p_N)^2 \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) &= \sum_{a_N} \left((a_N - p_N)^2 p_N^{a_N} (1-p_N)^{1-a_N} \right) \\ &= p_N(1-p_N) \end{aligned}$$

The second part is:

$$\sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left((a_N - p_N) \left(\sum_{i=1}^{N-1} a_i - G_{N-1} \right) \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) = 0$$

To see that this sum is zero, we just need to sum over a_N and notice that:

$$\sum_{a_N} \left((a_N - p_N) p_N^{a_N} (1-p_N)^{1-a_N} \right) = 0$$

The third and final part is:

$$\begin{aligned} & \sum_{a_N} \sum_{\mathbf{a}_{(N)}} \left(\left(\sum_{i=1}^{N-1} a_i - G_{N-1} \right)^2 \prod_{i=1}^{N} p_i^{a_i} (1-p_i)^{1-a_i} \right) = \\ = & \sum_{\mathbf{a}_{(N)}} \left(\left(\sum_{i=1}^{N-1} a_i - G_{N-1} \right)^2 \prod_{i=1}^{N-1} p_i^{a_i} (1-p_i)^{1-a_i} \right) \\ = & V_{N-1} \end{aligned}$$

Putting together the 3 parts we obtain:

$$V_N = p_N(1-p_N) + V_{N-1}$$

and applying this function recursively:

$$V_N = \sum_i p_i(1-p_i)$$

which correctly becomes the binomial variance $V_N = Np(1-p)$ when $p_i = p$.

2 Solution of Eqs. 11 and 12 for $i \to \infty$

Given the recursive system of equations ($i > 0$):

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

$$\phi^{(i)}_1(0) = \sum_N P(0|N)\, \phi^{(i)}_{0,\mathrm{pred}}(N) \tag{SM.2}$$

we want to solve it for $i \to \infty$. The solution is the fixed point of the equation, i.e. the pair $\left(\phi^{(\infty)}_{0,\mathrm{pred}}(N), \phi^{(\infty)}_1(0)\right)$ that fed into Eqs. SM.1 and SM.2 gives as result the same values. In formulas:

$$\phi^{(\infty)}_{0,\mathrm{pred}}(N) = \phi^{(0)}_{0,\mathrm{pred}}(N) + \phi^{(\infty)}_1(0)\, P(N|0) \tag{SM.3}$$

$$\phi^{(\infty)}_1(0) = \sum_N P(0|N)\, \phi^{(\infty)}_{0,\mathrm{pred}}(N) \tag{SM.4}$$

Substituting Eq. SM.3 into Eq. SM.4 we obtain:

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

and solving for $\phi^{(\infty)}_1(0)$:

$$\phi^{(\infty)}_1(0) = \frac{\sum_{N'} P(0|N')\, \phi^{(0)}_{0,\mathrm{pred}}(N')}{1 - \sum_{N'} P(N'|0)\, P(0|N')}$$

which is Eq.12 in the main text. Substituting this into Eq. SM.3 we obtain:

$$\phi^{(\infty)}_{0,\mathrm{pred}}(N) = \phi^{(0)}_{0,\mathrm{pred}}(N) + P(N/0)\frac{\sum_{N'} P(0|N')\,\phi^{(0)}_{0,\mathrm{pred}}(N')}{1-\sum_{N'} P(N'|0)\,P(0|N')}$$

which is Eq.11 in the main text.

3 Geometric Prior

In the main text a lognormal informative prior was used. Here we report the results of the algorithm when an uninformative improper geometric prior ($P(n) \sim 1/n$) is used instead of the lognormal prior. We only report the results for BCI; those for Pasoh are similar.

The procedure to find an uninformative prior in the general case is given by Jaynes (1968), who provided guidelines to be followed in any particular situations. Basically, the solution consists of exploiting the symmetries that this function must be present in order not to give any relevant information. Following Jaynes' guidelines, Pueyo et al. (2007) have shown that the symmetry to be exploited in the case of the species abundance is the invariance under random sampling, i.e. the correct uninformative prior is a function $P(n)$ with the property that a random sample from this abundance distribution has still a distribution of the shape $P(n)$. The only function with this property is $P(N) \sim 1/N$ (see Jaynes (1968) and Pueyo et al. (2007) for more detail). Note that this function is not normalizable; this is a common property for uninformative priors and it is not problematic or paradoxical since such a distribution is not meant to be used alone, but used inside Bayes' rule. In the rare cases where the application of Bayes' rule yields a non-normalizable probability distribution, it means that we do not possess enough data to give a definite prediction (Jaynes 2003). Note also that our prior is only coincidentally the same as the classic "Jeffrey's prior" (Jaynes 1968). Our prior was obtained by imposing invariance under random resampling, while the Jeffrey's prior is obtained by imposing invariance under variable rescaling.

As it may be seen in Fig 1, the use of the geometric prior overestimates the number of rare species. This problem is more evident for small values of a since, as with all Bayesian methods, our algorithm tends to give answers very close to the prior when few data are available (with no data the answer would just be the prior itself). The rare species are the ones suffering more from this effect since they are most likely to be absent in the sub-area from where the extrapolation is done.

When a increases, the overestimation of rare species decreases in magnitude until it becomes unnoticeable for $a = 0.5$.

4 Influence of the Extrapolation of k

To assess the effect caused by an inaccurate extrapolation of the parameter k, we ran a series of extrapolations (Fig 2) where we used the 'true' value of k obtained from the knowledge of all the data in A_0. (But note we only used data in A_1 in all other extrapolations in our study. The computation of the 'true' k is just for comparing the performance of our method based on the A_1 calculated k against that using the 'true' k.) All the differences between Fig 1 and Fig 2 are due to the difference in k. As can be seen in Table 1, the reconstruction with the 'true' k is slightly better both in likelihood and species prediction, but never significantly so, despite the fact that the parameter k appears to be very different between the two cases. This shows that our method is robust with respect to the determination of k, and even a rough extrapolation seems enough to obtain good results.

References

Jaynes, E. T. 1968. Prior Probabilities. - IEEE Transactions on systems science and cybernetics 4: 227-241.

Jaynes, E. T. 2003. Probability Theory – The Logic of Science. - Cambridge University Press.

Pueyo, S., He, F. and Zillio, T. 2007. The maximum entropy formalism and the idiosyncratic theory of biodiversity. - Eco. Lett. 10: 1017–1028.

List of Tables

Table 1:

			Extrapolation k			'True' k		
Plot	a	S_1	k	$S_{0,\mathrm{pred}}$	likelihood	k	$S_{0,\mathrm{pred}}$	likelihood
BCI	0.05	217	0.91	350	-2066.2	1.84	335	-2060.0
	0.1	233	4.89	293	**-2049.6**	2.18	301	**-2051.9**
	0.15	238	3.72	282	**-2049.0**	2.25	286	**-2050.6**
	0.2	250	2.06	297	**-2051.1**	2.44	295	**-2050.2**
	0.3	261	2.17	297	**-2050.6**	3.33	295	**-2049.1 ***
	0.4	272	2.79	299	**-2048.6 ***	4.65	298	**-2047.3 ***
	0.5	277	25.6	294	**-2046.7 ***	7.56	295	**-2046.3 ***

List of Figures

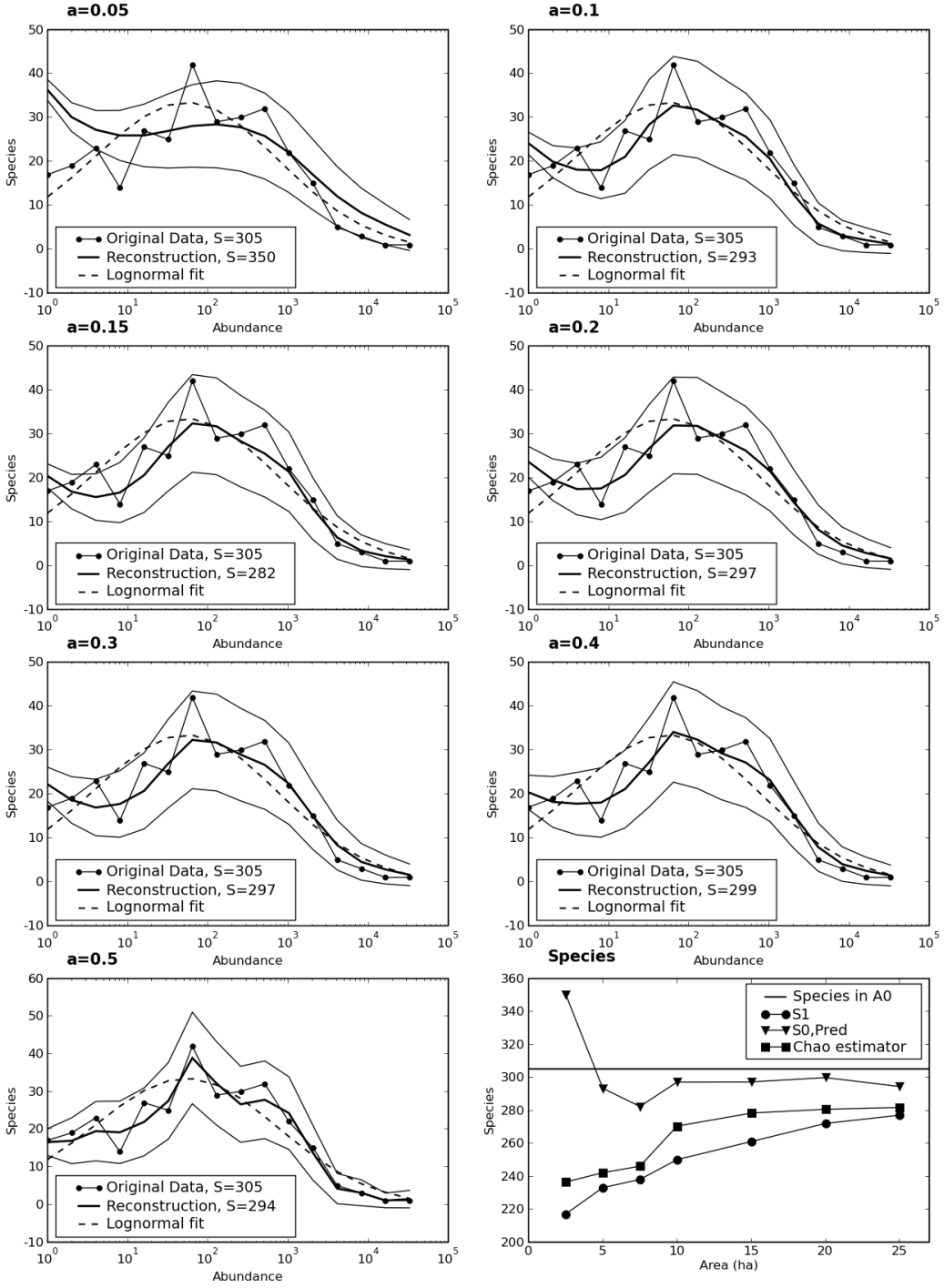
a=0.05
Original Data, S=305
Reconstruction, S=350
Lognormal fit
a=0.1
Original Data, S=305
Reconstruction, S=293
Lognormal fit
a=0.15
Original Data, S=305
Reconstruction, S=282
Lognormal fit
a=0.2
Original Data, S=305
Reconstruction, S=297
Lognormal fit
a=0.3
Original Data, S=305
Reconstruction, S=297
Lognormal fit
a=0.4
Original Data, S=305
Reconstruction, S=299
Lognormal fit
a=0.5
Original Data, S=305
Reconstruction, S=294
Lognormal fit
Species
Species in A0
S1
S0,Pred
Chao estimator
Abundance
Species
Area (ha)

Figure 1.

2 $a = 0.05, \ldots, 0.5$: reconstruction (solid line) with a geometric prior, and with knowledge of the 'true' value of k, of the SAD of the 50 ha BCI forest plot starting from a subarea of extension $A_1 = aA_0$ with various values of a where $A_0 = 50$ ha. The thin lines are the 95% Bayesian standard errors for the reconstruction. The dots represent the real species abundance in the 50 ha plot, and the dashed line is a lognormal fit to the data. Abundance classes are logarithmically binned. The actual and predicted number of species are indicated in the legend. The value of k was determined with the full knowledge of the data, to compare it with the results obtained when k was estimated by extrapolation, 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 of the method, and Chao's estimator is plotted for comparison.

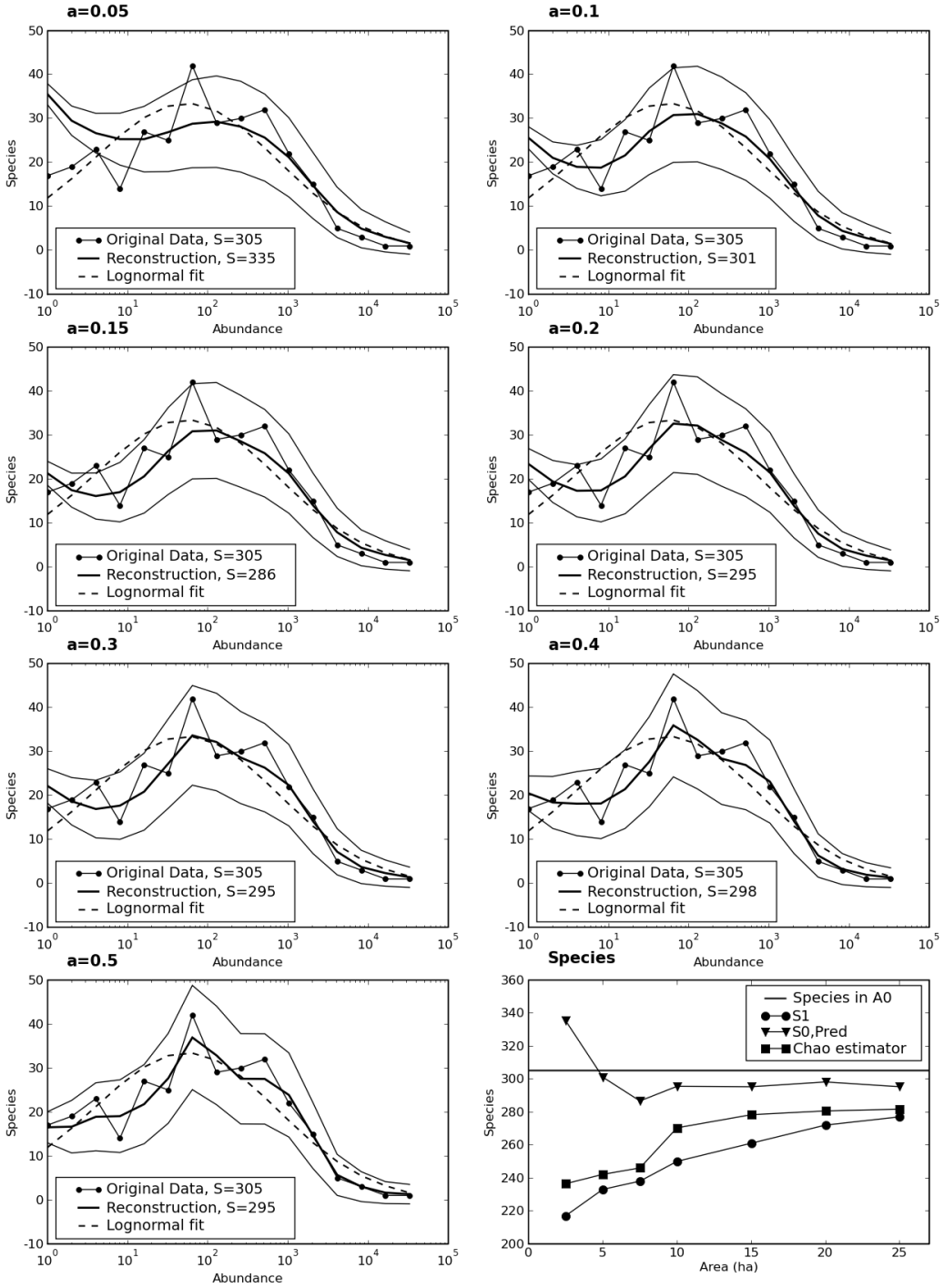
a=0.05
Original Data, S=305
Reconstruction, S=335
Lognormal fit
a=0.1
Original Data, S=305
Reconstruction, S=301
Lognormal fit
a=0.15
Original Data, S=305
Reconstruction, S=286
Lognormal fit
a=0.2
Original Data, S=305
Reconstruction, S=295
Lognormal fit
a=0.3
Original Data, S=305
Reconstruction, S=295
Lognormal fit
a=0.4
Original Data, S=305
Reconstruction, S=298
Lognormal fit
a=0.5
Original Data, S=305
Reconstruction, S=295
Lognormal fit
Species
Species in A0
S1
S0,Pred
Chao estimator
Species
Abundance
Area (ha)

Figure 2.