**Anne Chao, T. C. Hsieh, Robin L. Chazdon, Robert K. Colwell, and Nicholas J. Gotelli. 2015. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology* 96:1189-1201.**

APPENDIX C: Other models for the relative abundances of the undetected species.

In our estimation procedures, we adopt a nonparametric method for adjusting sample relative abundances for species detected in a sample (Eq. 4d of the main text). The adjustment method for detected species is valid for all species abundance distributions. For estimating the species relative abundances for undetected species, a functional form of parametric model is needed. In the main text, we assume a geometric series model based on an estimated number of undetected species $\hat{f}_0$ (see Eq. 5a of the main text). There are other models or distributions that could be applied to the set of undetected species indexed by 1, 2, …, $\hat{f}_0$.

*The broken-stick model*

Under a broken-stick model with $\hat{f}_0$ species, Pielou (1975) and Baczkowski (2000) showed that the *i*-th ordered species relative abundance of the set ( $p_1 > p_2 > ... > p_{\hat{f}_0}$ ) takes the form:

$$p_i = \frac{K}{\hat{f}_0}\left(\frac{1}{i}+\frac{1}{i+1}+...+\frac{1}{\hat{f}_0}\right)=\frac{K}{\hat{f}_0}\sum_{j=0}^{\hat{f}_0-i}\frac{1}{\hat{f}_0-j}, \quad i = 1, 2, \ldots, \ \hat{f}_0,$$

where $K$ is a normalized constant such that the sum of the relative abundances is equal to the coverage deficit estimator ${}^1\hat{C}_{def}$ (Eq. 2c in the main text), i.e.,

$$\sum_{i=1}^{\hat{f}_0} p_i = {}^1\hat{C}_{def}.$$

Because the expected relative abundances have an analytic form as indicated above, the RAD is fully determined by the number of undetected species. The model is thus not flexible.

*The Poisson log-normal model*

As indicated in Magurran (2004, p. 39), the continuous log-normal model should only be applied to continuous abundance data, such as biomass or cover measures, rather than to discrete data. For individual-based data, Bulmer (1974) presented a Poisson log-normal model. When Bulmer's model is used to the set of undetected species, the species relative abundances are:

$$p_i(M,V) = \frac{1}{\sqrt{2\pi V}}\int_0^{\infty}\frac{\lambda^{i-2}e^{-\lambda}}{(i-1)!}e^{-\frac{(\log\lambda-M)^2}{2V}}d\lambda, \quad i = 1, 2, \ldots, \hat{f}_0,$$

where $M$ and $V$ denote respectively the mean and variance of the log-normal distribution. Based on the Eqs. 6a and 6b in the main text, we have the following two equations for the

undetected species in terms of two parameters $M$ and $V$:

$$\sum_{i \in \text{undetected}} p_i \approx \sum_{i=1}^{\hat{f}_0} p_i(M,V) = {}^1\hat{C}_{def};$$

$$\sum_{i \in \text{undetected}} p_i^2 \approx \sum_{i=1}^{\hat{f}_0} [p_i(M,V)]^2 = {}^2\hat{C}_{def} \times \frac{\sum_{X_i \geq 2} X_i(X_i - 1)}{n(n-1)}.$$

*Comparison of models using spider data*

For the spider data (with 26 detected species and an estimate of 18 undetected species) discussed in the main text, we compare in Fig. C1 the empirical RAD and three fitted RADs including the proposed RAD curve based on a geometric series model, the broken-stick model, and the Poisson log-normal model. These three models were only fitted to the undetected species; for the 26 detected species, Eq. 4d of the main text is used for adjusting the plug-in estimator for all three models. So there are no differences in the adjusted relative abundances for the 26 detected species, as shown in Table C1 where the estimated complete RAD under each of the three models are listed.

For the proposed geometric series model, we also obtained the bootstrap s.e. based on 5000 replications. See the main text for the description of the bootstrap method. From Fig. C1 and Table C1, the Poisson log-normal and our proposed geometric series model yield almost identical RADs for this example. Here the parameters for the Poisson log-normal model are estimated to be $\hat{M} = 6.089$ and $\hat{V} = 4.779$ by the above system of nonlinear equations. (For some other data sets, solutions may not exist.) Except for a few species in the broken-stick model, the empirical RAD which ignores the set of the undetected species gives higher relative abundances than those in the other three models. The broken-stick model yields substantial low abundances for the tail of the fitted RAD.

Our simulation studies have suggested that the iterative steps for the Poisson log-normal models in many cases failed to converge to proper solutions. This is the main drawback to fit a Poisson log-normal model. For the simple geometric series model we used, all the iterative steps converged quickly. As discussed in the main text, we use these models only for modeling the undetected tail distribution; unless the assemblage is poorly sampled, the relative abundances of those undetected species (i.e., in the "tail" of the estimated RAD) are typically very small. Thus, the choice of the model for estimating the relative abundances of undetected species is a minor issue in our approach.

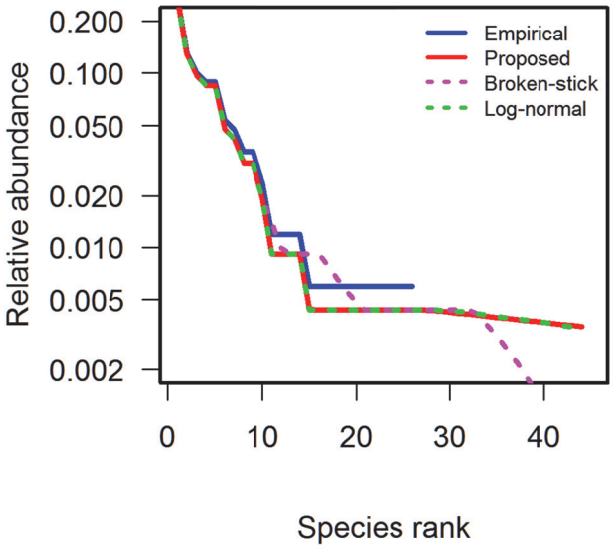


**Fig. C1.** Comparison of the empirical RAD curve (blue solid line) and three fitted RAD curves for the spider data discussed in the main text: the proposed RAD curve based on a geometric series model (red solid line), the broken-stick model (fuchsia dotted line), and the Poisson log-normal model (green dotted line). All these three models were fitted only to the set of undetected species. The adjustment method (Eq. 4d of the main text) for the 26 detected species is independent of species abundance distributions or models.

**Table C1**. Comparison of the estimated species-rank abundances of the empirical RAD and the estimated RADs based on three models (geometric series, Poisson log-normal, and broken-stick) fitted to the set of undetected species for the spider data discussed in the main text. The estimated abundances for the detected species are in black print; the estimated abundances for the undetected species are in red print. The empirical RAD applies only to 26 observed species whereas the other three models additionally consider 18 undetected species. The three models differ only in the estimated abundances for undetected species as the estimated abundances of the 26 detected species for all three models are based on the same adjusted formula (Eq. 4d of the main text). The bootstrap s.e. of the adjusted estimator for each detected species under the geometric series model is based on 5000 bootstrap replications.

| Species rank | Empirical RAD | Geometric series model (s.e.) | Poisson log-normal | Broken-stick model |
|---|---|---|---|---|
| 1 | 0.2738 | 0.2735 (0.0349) | 0.2735 | 0.2735 |
| 2 | 0.1310 | 0.1282 (0.261) | 0.1282 | 0.1282 |
| 3 | 0.1012 | 0.0973 (0.0228) | 0.0973 | 0.0973 |
| 4 | 0.0893 | 0.0849 (0.0213) | 0.0849 | 0.0849 |
| 5 | 0.0893 | 0.0849 (0.0215) | 0.0849 | 0.0849 |
| 6 | 0.0536 | 0.0482 (0.0162) | 0.0482 | 0.0482 |
| 7 | 0.0476 | 0.0422 (0.0151) | 0.0422 | 0.0422 |
| 8 | 0.0357 | 0.0306 (0.0130) | 0.0306 | 0.0306 |
| 9 | 0.0357 | 0.0306 (0.0128) | 0.0306 | 0.0306 |
| 10 | 0.0238 | 0.0194 (0.0100) | 0.0194 | 0.0194 |
| 11 | 0.0119 | 0.0091 (0.0060) | 0.0091 | 0.0138 |
| 12 | 0.0119 | 0.0091 (0.0060) | 0.0091 | 0.0099 |
| 13 | 0.0119 | 0.0091 (0.0061) | 0.0091 | 0.0091 |
| 14 | 0.0119 | 0.0091 (0.0060) | 0.0091 | 0.0091 |
| 15 | 0.0060 | 0.0044 (*) | 0.0044 | 0.0091 |
| 16 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0091 |
| 17 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0079 |
| 18 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0066 |
| 19 | 0.0060 | 0.0044 (0.0038) | 0.0044 | 0.0056 |
| 20 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0048 |
| 21 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0044 |
| 22 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0044 |
| 23 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0044 |
| 24 | 0.0060 | 0.0044 (0.0037) | 0.0044 | 0.0044 |
| 25 | 0.0060 | 0.0044 (0.0039) | 0.0044 | 0.0044 |
| 26 | 0.0060 | 0.0044 (0.0038) | 0.0044 | 0.0044 |

| | | | |
|---|---|---|---|
| 27 | 0.0044 (0.0039) | 0.0044 | 0.0044 |
| 28 | 0.0044 | 0.0044 | 0.0044 |
| 29 | 0.0043 | 0.0044 | 0.0044 |
| 30 | 0.0042 | 0.0043 | 0.0044 |
| 31 | 0.0042 | 0.0043 | 0.0044 |
| 32 | 0.0041 | 0.0042 | 0.0044 |
| 33 | 0.0041 | 0.0042 | 0.0041 |
| 34 | 0.0040 | 0.0041 | 0.0036 |
| 35 | 0.0040 | 0.0040 | 0.0031 |
| 36 | 0.0039 | 0.0040 | 0.0026 |
| 37 | 0.0039 | 0.0039 | 0.0022 |
| 38 | 0.0038 | 0.0038 | 0.0019 |
| 39 | 0.0038 | 0.0038 | 0.0015 |
| 40 | 0.0037 | 0.0037 | 0.0012 |
| 41 | 0.0037 | 0.0036 | 0.0010 |
| 42 | 0.0036 | 0.0036 | 0.0007 |
| 43 | 0.0036 | 0.0035 | 0.0005 |
| 44 | 0.0035 | 0.0031 | 0.0002 |

*An undetected species, so its bootstrap s.e. is not obtainable; see the subsection *Sampling variances of our estimators* of the main text for explanation.

LITERATURE CITED

Baczkowski, A. J. 2000. The broken-stick model for species abundances: an initial investigation. http://www1.maths.leeds.ac.uk/~sta6ajb/drep0010.pdf. (Accessed on October 16, 2014).

Bulmer, M. G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics 30:101-110.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.

Pielou, E. C. 1975. Ecological diversity. Wiley, New York, USA.