Anne Chao, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2013. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecology Monographs*.

Appendix B: Rarefaction and extrapolation for species richness (abundance data)

This Appendix briefly reviews the abundance data sections of Colwell et al. (2012), and Chao and Jost (2012).

Sample-size-based rarefaction/extrapolation

Let $S_{ind}(m)$ represent the expected number of species in a random sample of *m* individuals from the study assemblage, $S_{ind}(1)=1$. Under the multinomial model (Eq. 1 of the main text), if the true probabilities ($p_1, p_1, ..., p_S$) of each of the *S* species in the assemblage were known, then we have (Good 1953):

$$S_{ind}(m) = S - \sum_{i=1}^{S}(1-p_i)^m \text{ , } m \geq 1. \qquad \text{(B.1)}$$

The plot of $S_{ind}(m)$ with respect to the sample size *m* is the expected species accumulation curve. Individual-based rarefaction estimates the expected species richness for a smaller sample of size $m < n$. Based on the reference sample with observed species abundances X_i, the traditional rarefaction formula is a minimum variance unbiased estimator for $S_{ind}(m)$ (Hurlbert 1971, Smith and Grassle 1977):

$$\tilde{S}_{ind}(m) = S_{obs} - \sum_{X_i>0}\left[\binom{n-X_i}{m} \Big/ \binom{n}{m}\right], \ m < n. \qquad \text{(B.2)}$$

Here, $\binom{a}{b} \equiv 0$ if $a < b$. We use this conventional definition throughout this Appendix.

As discussed in the main text, there are two kinds of variance associated with the estimator $\tilde{S}_{ind}(m)$. A variance that is *conditional* on the reference sample measures only the variation in diversity that would arise from repeatedly resampling (without replacement) the given reference sample. This conditional variance approaches zero as *m* approaches *n* because the diversity of sample size of *n* is fixed (i.e., there is only one combination of all individuals or all sampling units). An *unconditional* variance measures the variation in diversity that would arise if another *new* sample of size *m* were taken from the entire assemblage (rather than from the original reference sample). Therefore, the unconditional variance does not approach 0 when sample size tends to *n*, and all associated confidence intervals are symmetric, which reflects the uncertainty

of the *new* sample. In deriving an unconditional variance, the number of undetected species must be estimated because those undetected species also affect the variation of a new sample. In most applications, unconditional variance is more useful because inferences are not restricted to the reference sample.

Colwell et al. (2012) derived an analytic expression for an asymptotic unconditional standard error (*s.e.*) for $\widetilde{S}_{ind}(m)$ and then applied it to construct a 95% confidence interval by using $\widetilde{S}_{ind}(m) \pm 1.96\ s.e.[\widetilde{S}_{ind}(m)]$. In order to accommodate more complicated rarefaction formulas for Hill numbers (see the main text) within a single unified framework, we suggest an alternative unconditional variance estimator based on a bootstrap method. Details are given in Appendix G.

Abundance-based extrapolation estimates the expected number of species $S_{ind}(n+m^*)$ in an augmented sample of $n + m^*$ individuals from the assemblage ($m^* > 0$). Previous analyses of abundance-based extrapolation include Good and Toulmin (1956), Melo et al. (2003), Shen et al. (2003), Chao and Shen (2004), Mao and Colwell (2005) and Mao (2007). The theoretical formula from Eq. (B.1) for an augmented sample of size $n + m^*$ can be expressed as

$$S_{ind}(n+m^*) = S - \sum_{i=1}^{S}(1-p_i)^{n+m^*}$$

$$= E(S_{obs}) + \sum_{i=1}^{S}[1-(1-p_i)^{m^*}](1-p_i)^n \cdot \qquad \text{(B.3)}$$

Note that as m^* tends to infinity, $S_{ind}(n+m^*)$ tends to species richness. Based only on the reference sample, with observed species frequencies X_i and their frequency counts f_i, we slightly modify the approach of Shen et al. (2003) and consider the following more accurate predictor for the species richness in an augmented sample of size $n + m^*$:

$$\widetilde{S}_{ind}(n+m^*) = S_{obs} + \hat{f}_0\left[1-\left(1-\frac{f_1}{n\hat{f}_0 + f_1}\right)^{m^*}\right]$$

$$\approx S_{obs} + \hat{f}_0\left[1-\exp\left(\frac{-m^* f_1}{n\hat{f}_0 + f_1}\right)\right]. \qquad \text{(B.4)}$$

In this prediction formula, $\hat{f}_0$ can be any proper predicted value of f_0, the number of undetected species present in the assemblage, but not observed in the reference sample. Colwell et al. (2012) suggested using the Chao1 estimator (Chao 1984) or ACE (Chao and Lee 1992) for $\hat{f}_0$. The estimator from the Chao1 estimator is

$$\hat{f}_0 = \begin{cases} [(n-1)/n] f_1^2/(2f_2), & \text{if } f_2 > 0 \\ [(n-1)/n] f_1(f_1-1)/2, & \text{if } f_2 = 0. \end{cases} \quad \text{(B.5)}$$

To obtain an unconditional variance estimator for $\tilde{S}_{ind}(n+m^*)$ and an associated confidence interval, we again used a bootstrap method (Appendix G). As m^* tends to infinity, $\tilde{S}_{ind}(n+m^*)$ approaches the Chao1 estimator, and the analytic variance (Colwell et al. 2012) of $\tilde{S}_{ind}(n+m^*)$ approaches that for the Chao1 estimator. Empirical simulations have indicated that the unconditional variance obtained from the bootstrap method tends to be slightly smaller than the analytic variance.

Colwell et al. (2012) connected the rarefaction part (which plots $\tilde{S}_{ind}(m)$ with respect to m, where $m < n$, see Eq. (B.2)) with the extrapolation part (which plots $\tilde{S}_{ind}(n+m^*)$ with respect to $n + m^*$ for $m^* > 0$; see Eq. (B.4)); the two parts of the curve as well as their corresponding confidence intervals join smoothly at the reference point (n, S_{obs}).

Coverage-based rarefaction/extrapolation

The concept of *sample coverage* (or simply *coverage*) was originally developed by the founder of modern computer science, Alan Turing, and I. J. Good (Good 1953, 2000). Coverage is a measure of sample completeness, and is defined as the total relative abundances of the observed species, or equivalently, the proportion of the total number of individuals in an assemblage that belong to species represented in the sample. For the reference sample of size n from a multinomial model given in Eq. 1 of the main text, sample coverage is defined as:

$$C_{ind}(n) = \sum_{i=1}^{S} p_i I(X_i > 0), \quad \text{(B.6)}$$

where $I(A)$ is an indicator function that equals 1 when A is true and 0 otherwise.

Contrary to most people's intuition, sample coverage can be very accurately and efficiently estimated using only information contained in the sample itself, as long as the sample is reasonably large (Good 1953, Robbins 1968, Esty 1983, 1986). Given a reference sample of size n, the Good-Turing estimator of sample coverage is simply $1 - f_1/n$, where f_1 is the number of singletons. Chao and Jost (2012) used a more accurate estimator:

$$\hat{C}_{ind}(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]. \quad \text{(B.7)}$$

Chao and Jost (2012) suggested that rarefaction and extrapolation be plotted with respect to sample coverage rather than with respect to abundance or number of sampling units. One of their main reasons is that the expected species richness for standardized sample coverage satisfies a replication principle (or doubling property), which the expected species richness for standardized sample size does not obey; see Chao and Jost (2012, their Appendix A) for a proof. They derived sample coverage estimates for rarefied and augmented samples to construct the coverage-based rarefaction/extrapolation curve. For any sample size *m*, let $C_{ind}(m)$ be the expected coverage for a sample of size of *m* individuals and we have (Good, 1953):

$$C_{ind}(m) = \sum_{i=1}^{S} p_i[1-(1-p_i)^m] = 1-\sum_{i=1}^{S} p_i(1-p_i)^m \text{ , } m > 0. \qquad \text{(B.8)}$$

For the rarefaction part of the curve ($m < n$), Alroy (2010) and Jost (2010) suggested algorithmic approaches to estimate this interpolated coverage. Here we adopt the following analytic minimum variance unbiased estimator of the expected coverage $C_{ind}(m)$ derived by Chao and Jost (2012):

$$\hat{C}_{ind}(m) = 1-\sum_{X_i \geq 1} \frac{X_i}{n} \frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}, \quad m < n, \qquad \text{(B.9)}$$

where X_i is the number of individuals of species *i* observed in the reference sample. For the extrapolation part of the curve ($n + m^*$), they derived an estimator for sample coverage for an augmented sample of size $n+m^*$:

$$\hat{C}_{ind}(n+m^*) = 1-\frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1+2f_2}\right]^{m^*+1}. \qquad \text{(B.10)}$$

As m^* tends to infinity, the extrapolated coverage estimator approaches unity, indicating a complete coverage. When $m^* = 0$, Eq. (B.10) reduces to Eq. (B.7), the sample coverage estimator for the reference sample.

As with sample-size-based curves, the coverage-based interpolation (which plots $\tilde{S}_{ind}(m)$ with respect to $\hat{C}_{ind}(m)$, $m < n$, see Eqs. (B.2) and (B.9)) and extrapolation (which plots $\tilde{S}_{ind}(n+m^*)$ with respect to $\hat{C}_{ind}(n+m^*)$ for $m^* > 0$; see Eqs. (B.4) and (B.10)) join smoothly at the reference point ($\hat{C}_{ind}(n)$, S_{obs}), see Eq. (B.7). The confidence intervals of expected species richness based on the bootstrap method also join smoothly.

LITERATURE CITED

Alroy, J. 2010. The shifting balance of diversity among major marine animal groups. Science 329:1191-1194.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao, A., and L. Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533-2547.

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

Chao, A. and T. J. Shen. 2004. Nonparametric prediction in species sampling. Journal of agricultural, biological, and environmental statistics 9:253-269.

Colwell, R. K., A. Chao, N. J. Gotelli, S. Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3-21.

Esty, W. W. 1983. A normal limit law for a nonparametric estimator of the coverage of a random sample. The Annals of Statistics 11:905-912.

Esty, W. W. 1986. The efficiency of Good's nonparametric coverage estimator. The Annals of Statistics 14:1257-1260.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237-264.

Good, I. J. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. Journal of Statistical Computation and Simulation 66:101-111.

Good, I. J., and G. Toulmin. 1956. The number of new species, and the increase in population coverage, when a sample is increased. Biometrika 43:45-63.

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577-586.

Jost, L. 2010. The relation between evenness and diversity. Diversity 2:207-232.

Mao, C. X. 2007. Estimating species accumulation curves and diversity indices. Statistica Sinica 17:761-774.

Mao, C. X., and R. K. Colwell. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology 86:1143-1153.

Melo, A. S., R. A. S. Pereira, A. J. Santos, G. J. Shepherd, G. Machado, H. F. Medeiros, and R. J. Sawaya. 2003. Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? Oikos 101:398-410.

Robbins, H. E. 1968. Estimating the total probability of the unobserved outcomes of an experiment. The Annals of Mathematical Statistics 39:256-257.

Shen, T. J., A. Chao, and C. F. Lin. 2003. Predicting the number of new species in further taxonomic sampling. Ecology 84:798-804.

Smith, W. and J. F. Grassle. 1977. Sampling properties of a family of diversity measures. Biometrics 33:283-292.