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 E: Estimating diversity profile based on the proposed RAD estimator (for abundance data).

With the estimated RAD, we can infer many parameters of the focal assemblage. Here we present the estimation of diversity (Hill numbers) profile. Hill (1973) integrated species richness and species abundances into a class of diversity measures later called Hill numbers, or the effective numbers of species, defined as

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}, \quad q \neq 1. \qquad \text{(E.1)} $$

The parameter q determines the sensitivity of the measure to the relative abundances. When $q = 0$, ${}^{0}D$ is simply species richness. For $q = 1$, Eq. E.1 is undefined, but its limit as q tends to 1 is the exponential of the familiar Shannon index, referred to as Shannon diversity (Chao et al. 2014). When $q = 2$, Eq. E.1 yields Simpson diversity, the inverse of the Simpson concentration. A complete characterization of the abundance-based species diversity of an assemblage is conveyed by a diversity profile – a plot of ${}^{q}D$ versus q – from $q = 0$ to $q = 3$ or 4 (beyond 3 or 4 it generally changes little). The diversity profile also completely characterizes the RAD itself.

It is well known that the empirical diversity profile obtained by substituting sample relative abundances (the "plug-in estimator") into Eq. E.1 underestimates the true profile especially for $q \leq 1$. Previous approaches to correcting for this negative bias were proposed only for non-negative integers $q = 0$, 1 and 2 (see Gotelli and Chao 2013 for a review). Now, with our proposed RAD estimator, we can provide a bias-corrected diversity profile for *all* $q \geq 0$, including non-integer values of q, by Eqs. 4d and 6c of the main text.

The variance of the resulting diversity estimator and the associated confidence intervals for the diversity of order q can also be constructed by a bootstrap method based on sampling with replacement from the estimated RAD. The bootstrap procedures are similar to those presented in the main text in the section *Sampling variances of our estimators.* First, a random sample of n individuals is generated from the estimated RAD with replacement to obtain a new set of species sample abundances. Based on this new set, we then calculate new coverage estimates $({}^{1}\hat{C},\ {}^{2}\hat{C})$ and their deficits, a new estimated number of undetected species, and new estimates $(\hat{\lambda}, \hat{\theta})$ and $(\hat{\alpha}, \hat{\beta})$; all these new statistics are then substituted into Eqs. 4d and 6c to obtain a new RAD estimator. That is, all statistics in our RAD estimator are replaced by those computed from the new generated set of species sample abundances. Then a bootstrap diversity estimate ${}^{q}\hat{D}^{*}$, based on the estimated RAD is calculated. The procedure is replicated B times to obtain B bootstrap diversity estimates $\{{}^{q}\hat{D}^{*1}, {}^{q}\hat{D}^{*2}, ..., {}^{q}\hat{D}^{*B}\}$ (B = 1,000 is suggested in confidence interval construction). The bootstrap variance estimator of ${}^{q}\hat{D}$ is the sample variance of these B estimates. Moreover, the 2.5% and 97.5% percentiles of these B bootstrap estimates can be used to construct a 95% confidence interval. Similar procedures can be used to derive variance estimators for any other estimator (e.g., empirical diversity) and their associated confidence intervals (see the example below).

Example (abundance data)

In Fig. E1, we show the diversity profile based on the empirical RAD and the estimated RAD respectively for the spider data discussed in the main text (26 detected species and an estimate of 18 undetected species, Sackett et al. 2011). The associated 95% confidence intervals using 1,000 bootstrap replications are also shown. For any order $q, 0 \le q \le 3$, the estimated diversity of order q is higher than the corresponding empirical diversity. When $q \le 1$, substantial differences exist between the two profiles; the difference arises because in our approach the contribution from undetected species is included whereas the empirical estimates ignore the undetected species. We later use simulation to show the estimated diversity profile based on the proposed RAD estimator eliminates most of the negative bias associated with the empirical diversity profile.

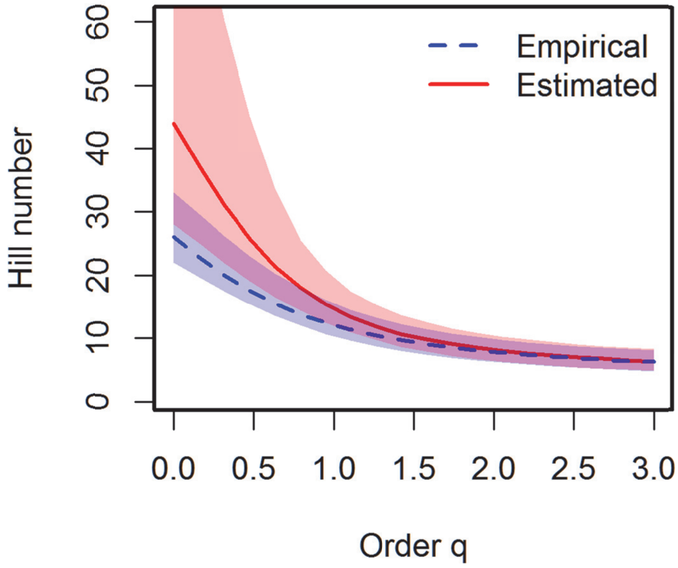
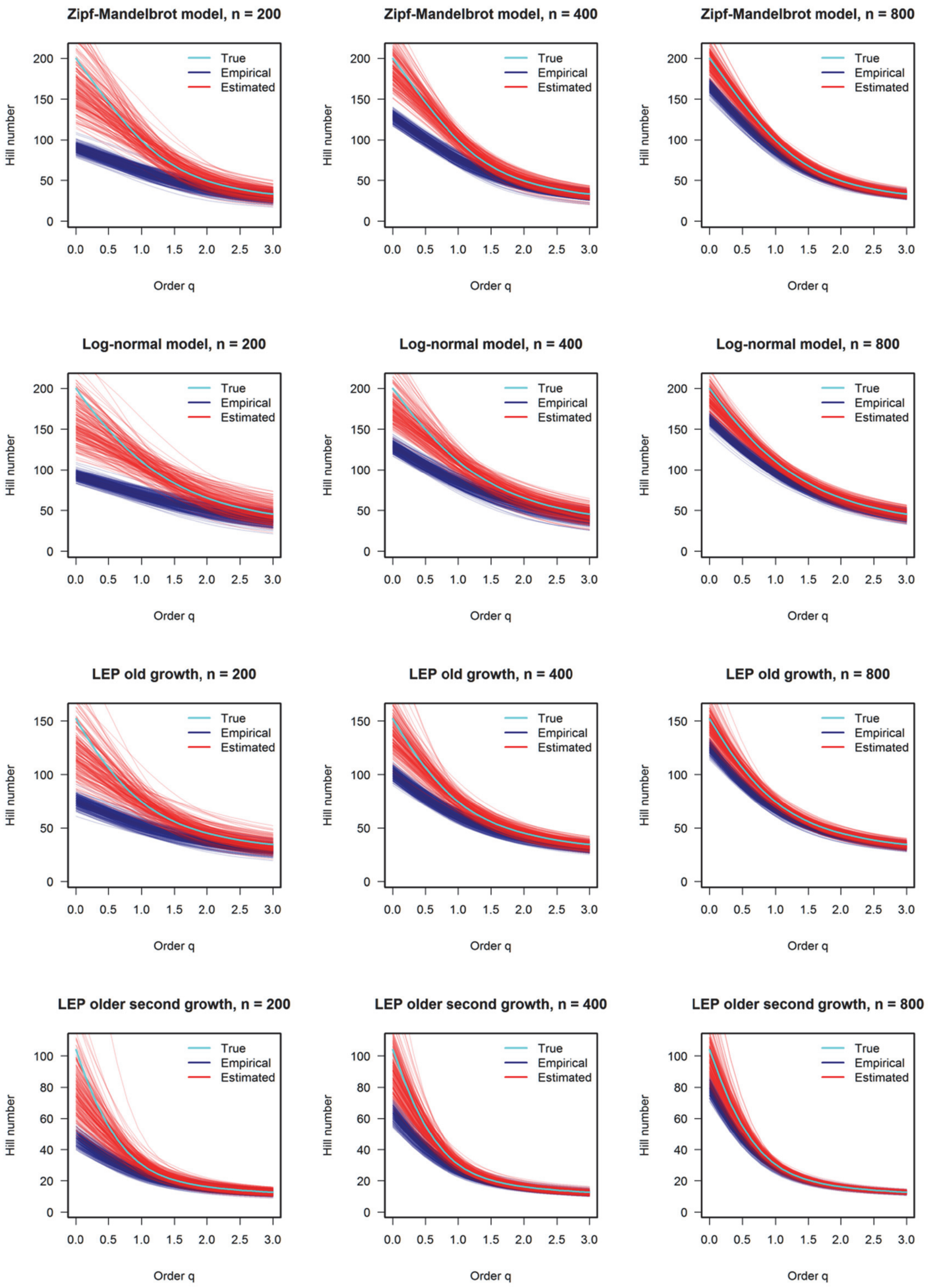


Fig. E1. Comparison of the empirical diversity profile (blue dashed line) and estimated diversity profile (red solid line) as a function of order q, $0 \le q \le 3$, for the abundance data of forest spiders (Sackett et al. 2011). The shaded areas denote the 95% confidence intervals based on 1000 bootstrap replications from the estimated RAD.

Simulation comparisons of empirical diversity profiles and the proposed diversity profiles

Based on the same scenarios used in Appendix A, we examine in Fig. E2 the diversity profiles based on the estimated RAD. For each generated data set under a particular scenario, we compare the true diversity profile of the complete assemblage, the empirical diversity profile, and the estimated diversity profile for $0 \le q \le 3$. All empirical diversity profiles are much below the true curves and thus exhibit negative biases, especially for $q \le 1$. The biases are substantial for $n = 200$. Our estiamted profile eliminates much of the bias and produces nearly unbiased diversity estimates (which fluctuate below and above the true profile) when q is not close to 0. Note that when q is equal to 0, our estimate is the Chao1 estimator, which is an estimated lower bound of the true species richness.

In the literature, the most widely used measures under the framework of Hill numbers are the species richness ($q = 0$), Shannon diversity ($q = 1$) and Simpson diversity ($q = 2$). For $q = 0$, the empirical diversity is the observed species richness; its bias is thus the number of undetected species in the sample. This bias-correction issue is a well known subject in many disciplines; see Chao and Chiu (2012) for a review. For $q = 1$, the empirical diversity is the observed exponential of Shannon entropy, which also exhibits substantial negative bias in hyper-diverse assemblages; see Chao et al. (2013). Our simulation results (Fig. E2) reveal that the diversity estimates based on the proposed RAD perform much better than the empirical diversity measures. However, our estimation procedures require solving nonlinear equations (Eqs. 4b, 4c and Eqs. 6a, 6b) and occasionally the solutions may not lie in the expected range, as discussed after Eq. 4c in the main text. We are currently working on analytic estimators of diversity and entropy profiles.


Zipf-Mandelbrot model, n = 200
Zipf-Mandelbrot model, n = 400
Zipf-Mandelbrot model, n = 800
Log-normal model, n = 200
Log-normal model, n = 400
Log-normal model, n = 800
LEP old growth, n = 200
LEP old growth, n = 400
LEP old growth, n = 800
LEP older second growth, n = 200
LEP older second growth, n = 400
LEP older second growth, n = 800
True
Empirical
Estimated
Hill number
Order q
0
50
100
150
200
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0

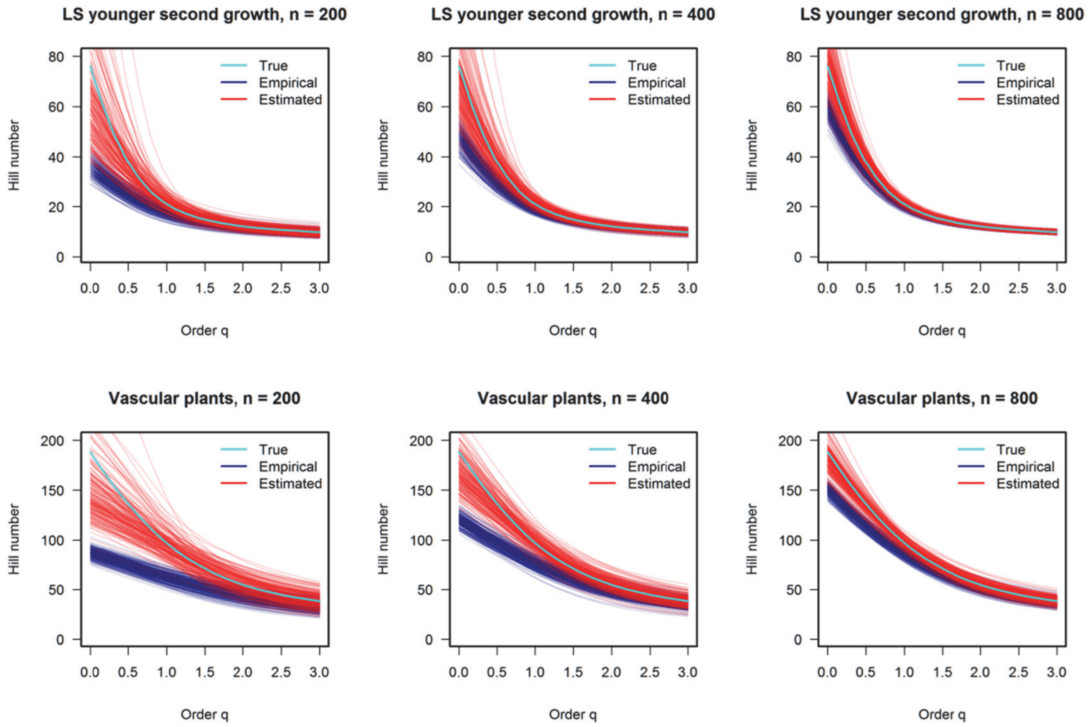


Fig. E2. Comparison of the true diversity profile of the complete assemblage (light blue line in each panel), the empirical diversity profiles (superimposed dark blue lines with 200 replications) and the estimated diversity profiles based on the proposed RAD estimator (superimposed red lines with 200 replications) for sample size 200 (left panels), 400 (middle panels) and 800 (right panels). Data sets were generated from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and four plant assemblages (see Appendix A for data details). For each assemblage and each sample size, 200 data sets were generated, thus there are 200 estimated profiles (200 red lines and 200 dark blue lines). Note that the X-axis is the order of Hill numbers, while the Y-axis is in units of “effective number of species”.

LITERATURE CITED

Chao, A. and Chiu, C. H. 2012. Estimation of species richness and shared species richness. Pages 76-111 in N. Balakrishnan (Ed). Methods and Applications of Statistics in the Atmospheric and Earth Sciences. Wiley, New York, USA.

Chao, A., N. G. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species biodiversity studies. Ecological Monographs 84:45-67.

Chao, A., Y. T. Wang, and L. Jost. 2013. Entropy and the species accumulation curve: a novel estimator of entropy via discovery rates of new species. Methods in Ecology and Evolution 4:1091-1110.

Gotelli, N. J., and A. Chao, A. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. Pages 195-211 in S. A. Levin, ed. Encyclopedia of Biodiversity, 2nd Edition, Vol. 5. Academic Press, Waltham, Massachusetts, USA.

Hill, M. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427-432.

Sackett, T. E., S. Record, S. Bewick, B. Baiser, N. J. Sanders, and A. M. Ellison. 2011. Response of macroarthropod assemblages to the loss of hemlock (*Tsuga canadensis*), a foundation species. Ecosphere 2:art74.