Anne Chao, T. C. Hsieh, Robin L. Chazdon, Robert K. Colwell, and Nicholas J. Gotelli. 2014. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology*.

APPENDIX A: Simulation results based on two theoretical abundnace distributions and four large empirical surveys

Comparisons of the true RAD, the empirical RAD and the proposed RAD for six scenarios

In this appendix, we report the simulation results for six scenarios including the one presented in Fig. 1 of the main text. We simulated data from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and treated four large empirical diversity surveys as the complete assemblages. In the latter case, the species-rank abundance distribution from each survey was assumed to be the “true” complete RAD. The six scenarios are described below. In each case, if ($p_1, p_2, \ldots, p_S$) are fixed parameters, then we give the true value of the CV (coefficient of variation, which is the ratio of standard deviation and mean) of these probabilities. The CV value quantifies the degree of heterogeneity of the probabilities ($p_1, p_2, \ldots, p_S$). When all probabilities are equal, CV = 0. A larger value of CV signifies higher degree of heterogeneity among probabilities. If ($p_1, p_2, \ldots, p_S$) are random variables generated from a distribution (e.g., the log-normal model, as described below), then the average values of CV over 200 simulated data sets are given to approximate the true theoretical value.

<u>Scenario 1</u> (The Zipf-Mandelbrot model with 200 species). The relative abundances of the complete assemblage take the general form $p_i = c / (2 + i)$, i =1, 2, …, 200, where c is a normalized constant such that the sum of the relative abundances is unity; see Magurran (2004) for an introduction of the Zipf-Mandelbrot model. CV = 1.751.

<u>Scenario 2</u>: (The log-normal model with 200 species). We first generated 200 random variables ($a_1, a_2, \ldots, a_{200}$) from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The species relative abundance of species i of the complete assemblage takes the form $p_i = c \exp(a_i)$, where c is a normalized constant. Then all samples were generated from the relative abundances ($p_1, p_2, \ldots, p_{200}$) throughout the simulation. CV = 1.439 for the generated set used in our simulation.

<u>Scenarios 3, 4, 5</u>: We treated three tree data sets collected by Chazdon and colleagues as the complete assemblages; see Norden et al. (2009) for details. The three data sets include species abundance survey data taken, respectively, from the LEP old-growth forest site (*Scenario 3*, 152 species, 943 individuals, CV = 1.545), the LEP older second-growth forest site (*Scenario 4*, 104 species, 1263 individuals, CV = 2.339), and the LS younger second-growth forest site (*Scenario 5,* 76 species, 1020 individuals, CV = 2.305). The species abundance frequency counts for the three surveys are given in Table A1.

<u>Scenario 6</u>: We treated the census data of Miller and Wiegert (1989) for endangered and rare vascular plant species in the central portion of the southern Appalachian Region (USA) as the true assemblage. The species abundance frequency counts for this survey are given in Table A1; a total of 188 species were represented by 1008 individuals. CV = 1.563.

Given the species relative abundances of each complete assemblage, we generated 200 data sets of sample sizes 200, 400 and 800. Then for each generated data set, we obtained the empirical RAD and the proposed RAD. Comparisons of the true RAD of the complete assemblage, the empirical RAD curves, and the proposed RAD curves are shown in Fig. A1. All the patterns are generally consistent with those presented in Fig. 1 of the main text for Scenario 1. Therefore, the findings and interpretations discussed in the main text for Scenario 1 can be applied to all other scenarios. We thus omit the details.

Table A1. Species abundance frequency counts for Scenarios 3–6.

Scenario 3. LEP old-growth forest, 152 species, 943 individuals (Norden et al. 2009)

i	1	2	3	4	5	6	7	8	9	10	11	13	15	16	18	19	20	25	38	39	40	46	52	55
f_i	46	30	16	12	6	5	3	4	5	4	1	3	1	1	1	1	4	3	1	1	1	1	1	1

Scenario 4. LEP older (29 years) second-growth forest, 104 species, 1263 individuals (Norden et al. 2009)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	20	22	39	45	57	72	88	132	133	178
f_i	33	15	13	4	5	3	3	1	2	1	4	2	2	1	2	1	1	1	1	1	1	1	1	2	1	1	1

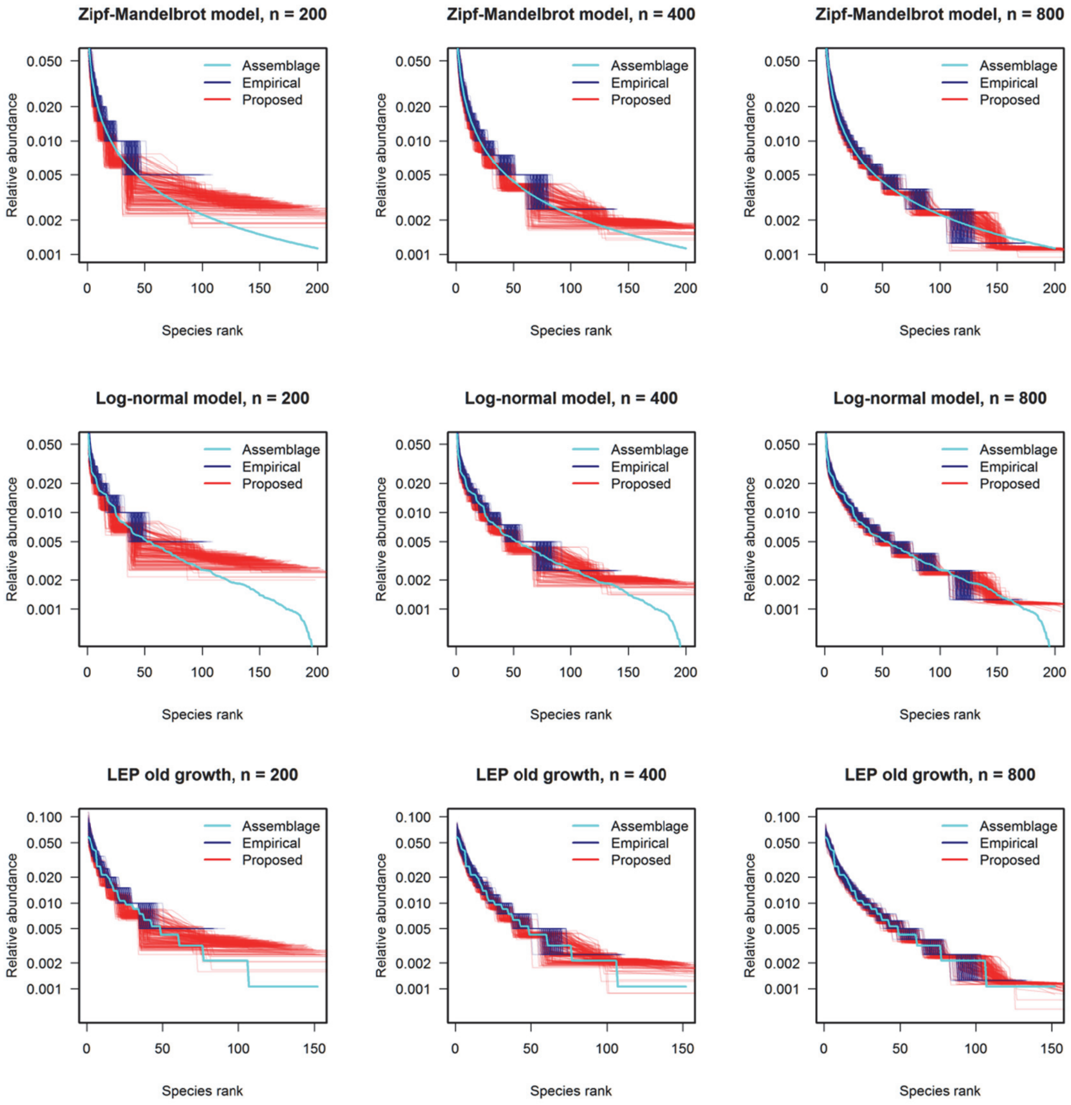
Scenario 5. LS younger (21 years) second-growth forest, 76 species, 1020 individuals (Norden et al. 2009)

i	1	2	3	4	5	7	8	10	11	12	13	15	31	33	34	35	66	72	78	127	131	174
f_i	29	13	5	2	3	4	1	2	2	1	2	2	1	1	1	1	1	1	1	1	1	1

Scenario 6. The extant rare vascular plant species in the southern Appalachians, 188 species, 1008 individuals (Miller and Wiegert 1989)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	61	35	18	12	15	4	8	4	5	5	1	2	1	2	3

i	16	19	20	22	29	32	40	43	48	67
f_i	2	1	2	1	1	1	1	1	1	1



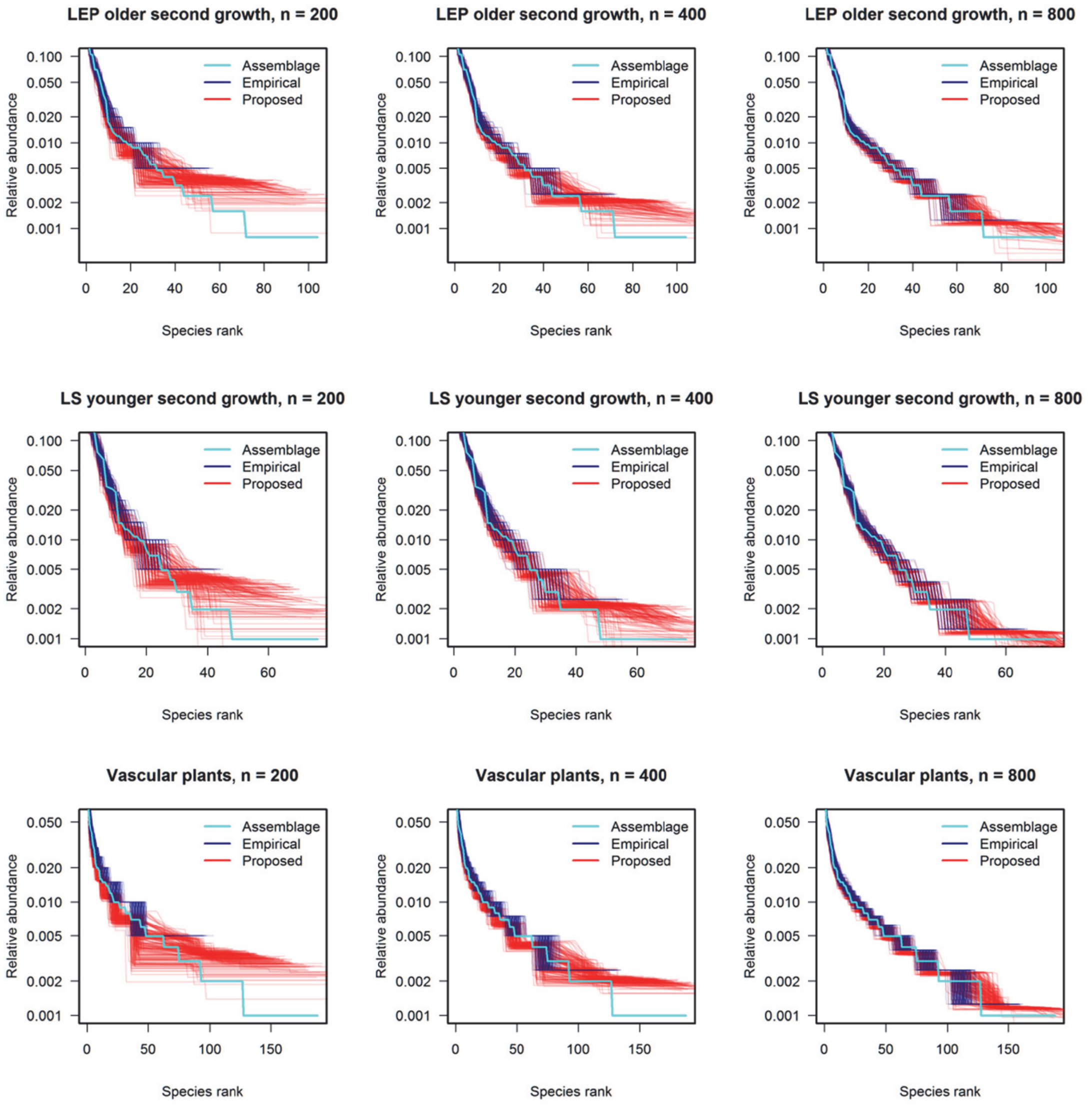


Fig. A1. Comparison of the true RAD of the complete assemblage (light blue line), the empirical RAD curves (superimposed dark blue lines with 200 replications), and the proposed RAD curves (superimposed red lines with 200 replications) for sample sizes 200 (left panels), 400 (middle panels) and 800 (right panels). Data sets were generated from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and four plant assemblages (Scenarios 3–6, see Table A1 for the species abundance frequency counts). For each assemblage and each sample size, 200 data sets were generated, thus there are 200 estimated RADs (200 red lines and 200 dark blue lines). Note that the X-axis is the species list, ranked from most abundant to least abundant, and the Y-axis (relative abundance) is displayed on a $\log_{10}$ scale.

Comparison of the sample relative abundance and the adjusted estimator

In Eq. 4d of the main text, we derive for a detected species *i* (i.e., species sample frequency $X_i > 0$) the following adjusted estimator of the plug-in estimator:

$$\tilde{p}_i = \frac{X_i}{n}(1 - \hat{\lambda} e^{-\hat{\theta} X_i}) . \qquad \text{(A.1)}$$

This adjusted estimator also provides a simple nonparametric relationship between the species sample relative abundance (i.e., X_i / n) and the estimated true relative abundance in the entire assemblage. To examine the relative performance of the two estimators for each species, we conducted simulations by generating data sets from a Zipf-Manelbrot model with 200 species (Scenario 1 described in Appendix A) and from the assemblage of vascular plants with 188 species (Scenario 6 described in Appendix A).

The species in each assemblage are ordered and indexed from the most abundant (with an index 1) to the least abundant (with an index 200 in the Zipf-Mandelbrot model, and an index 188 in the vascular plant assemblage). Given the species relative abundances, we generated 5000 samples for three sample sizes (200, 400 and 800). For each generated sample, the abundance for each detected species was recorded; the plug-in and the adjusted estimators were calculated. For each particular species, we only considered those samples in which that species was detected because the focus was on estimating the relative abundance of a detected species. The number of times that each species was detected in the 5000 generated samples varies with species, and thus the number of simulated samples used to obtain the averages of the relative biases and RMSEs are different. In Figs. A2 and A3, we respectively present the plots of the average relative bias (with respect to the true relative abundance) and the RMSE for each species indexed from the most abundant to the least abundant.

Based on Figs. A2 and A3, we have the following findings:

(1) Both figures reveal that for those relatively abundant species the plug-in estimator works well and is nearly unbiased, and thus no adjustment is actually needed. This is predicted from our theory discussed in the main text. Due to downward bias-correction, the adjusted estimator is nevertheless subject to slight negative bias for relatively abundant species, but the difference in bias between the two estimators for relatively abundant species is generally limited.

(2) For those relatively rare species, the plug-in estimator exhibits positive bias, reiterating our theory (Eq. 1 of the main text). The degree of the positive bias increases with increasing relative rarity of each species, as shown by the increasing trend of the relative bias when species abundance is decreased. Fig. A2 shows that our adjusted estimator for the relatively rare species can reduce the positive bias inherent in the plug-in estimator; the reduction is substantial for smaller sample sizes (*n* = 200 and 400).

(3) Fig. A3 further demonstrates that the adjusted estimator is generally more accurate in terms of smaller RMSE. The improvement over the plug-in estimator is clearly seen for smaller sample sizes (*n* = 200 and 400) and also increases with increasing relative rarity of each species.

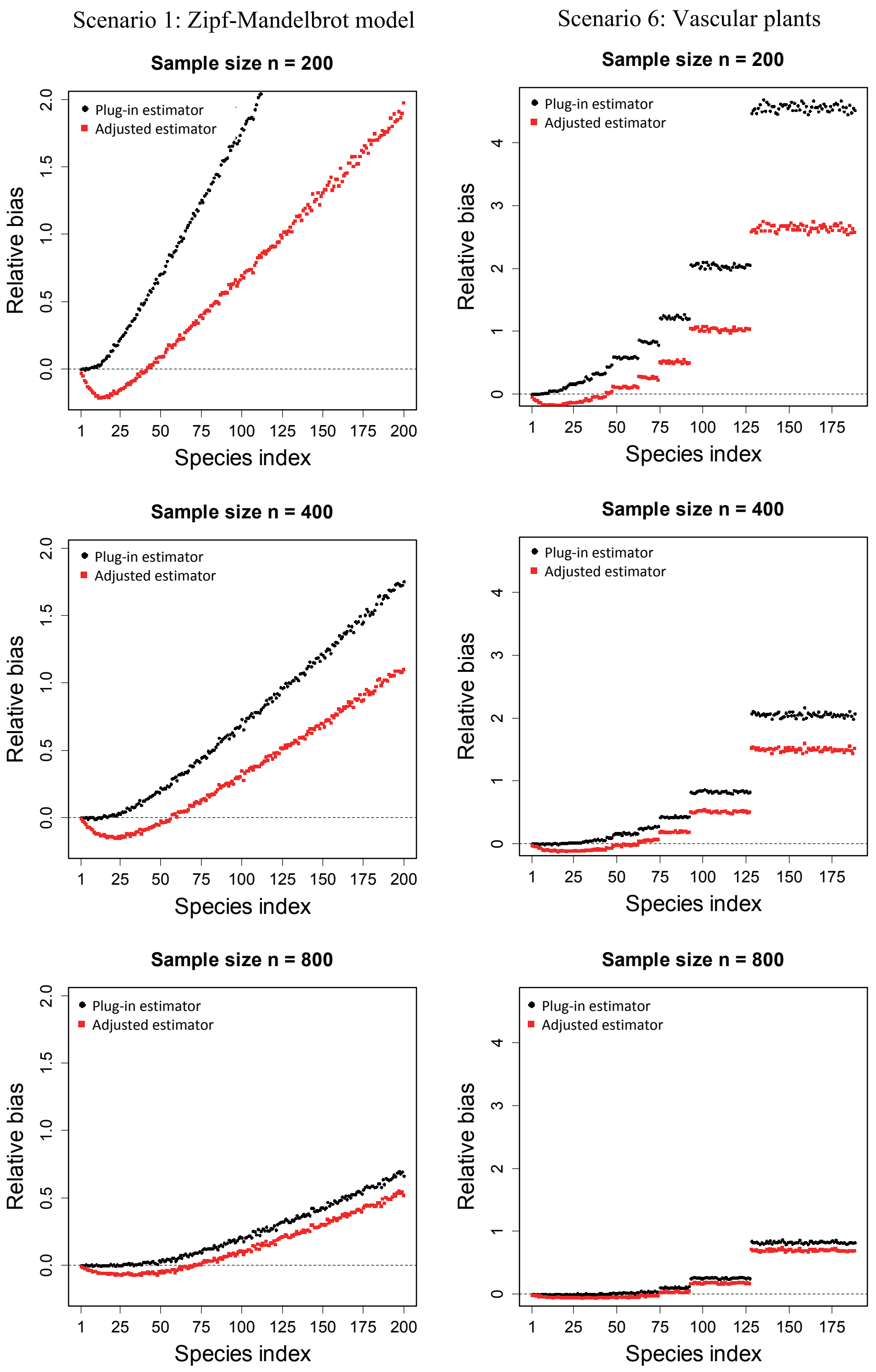


Fig. A2. The average relative bias of the plug-in estimator $\hat{p}_i = X_i / n$ and the proposed adjusted estimator $\tilde{p}_i = (X_i / n)(1 - \hat{\lambda} e^{-\hat{\theta} X_i})$ (Eq. A.1 or Eq. 4d of the main text) for species ordered and indexed from the most abundant to the least abundant under the Zipf-Mandelbrot model (left panels) and the assemblage of vascular plants (right panels) for sample sizes 200 (upper panels), 400 (middle panels) and 800 (lower panels). The horizontal dotted line in each panel represents the relative bias = 0 line.

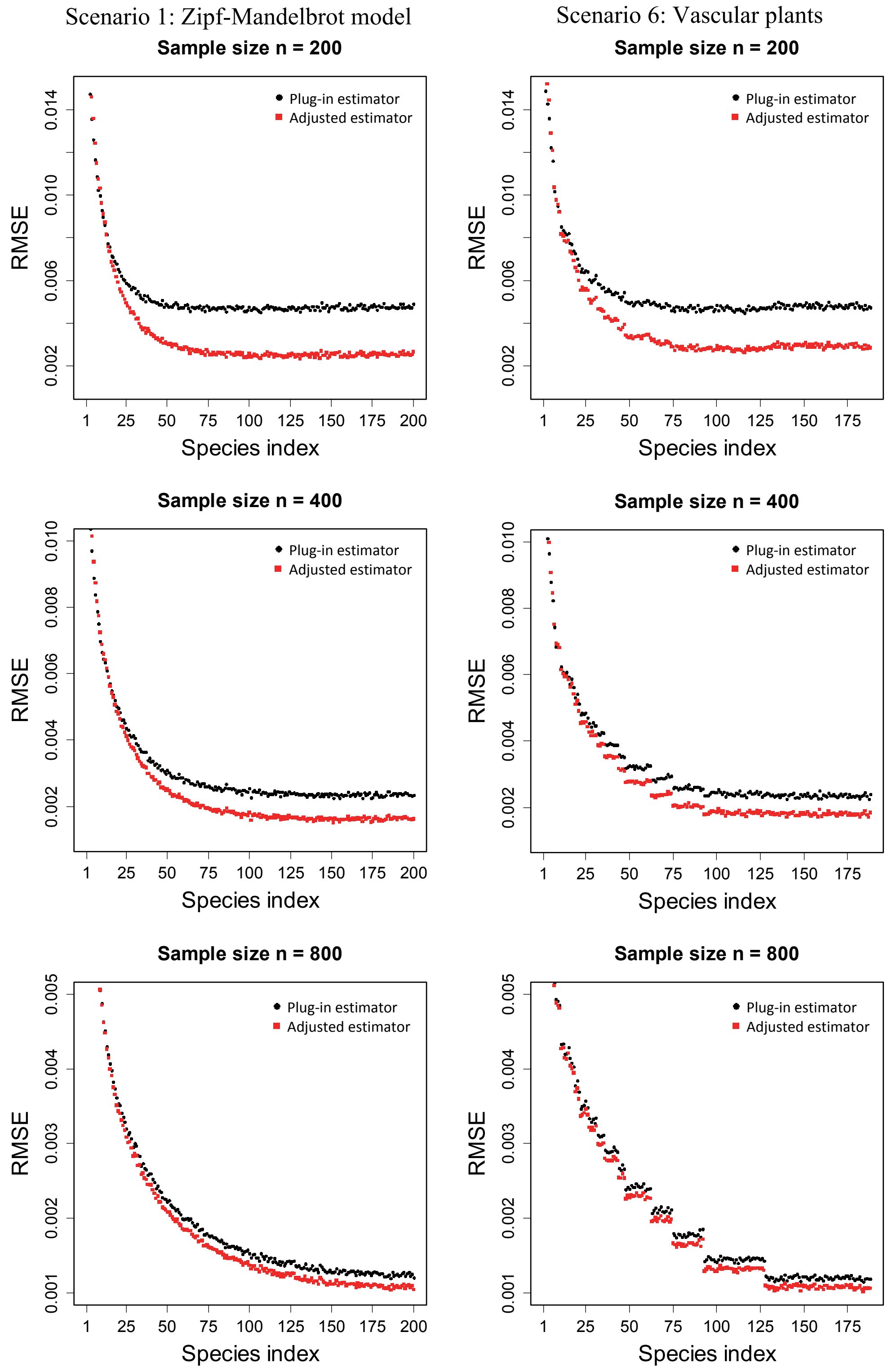


Fig. A3. The root mean squared error (RMSE) of the plug-in estimator $\hat{p}_i = X_i / n$ and the proposed adjusted estimator $\tilde{p}_i = (X_i / n)(1 - \hat{\lambda} e^{-\hat{\theta} X_i})$ (Eq. A.1 or Eq. 4d of the main text) for species ordered and indexed from the most abundant to the least abundant under the Zipf-Mandelbrot model (left panels) and the assemblage of vascular plants (right panels) for sample sizes 200 (upper panels), 400 (middle panels) and 800 (lower panels).

LITERATURE CITED

Miller, R. I., and R. G. Wiegert. 1989. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. Ecology 70:16-22.

Norden, N., R. C. Chazdon, A. Chao, Y.-H. Jiang, and B. Vilchez-Alvarado. 2009. Resilience of tropical rain forests: rapid tree community reassembly in secondary forests. Ecology Letters 12:385-394.