Appendix S1

for

Community-level species' correlated distribution can be scale-independent and related to the evenness of abundance

Youhua Chen[1], Tsung-Jen Shen[2*], Richard Condit[3,4], Stephen P. Hubbell[5,6]

[1] CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, China

[2]Institute of Statistics & Department of Applied Mathematics, National Chung Hsing University, 250 Kuo Kuang Road, Taichung 40227, Taiwan

[3]Field Museum of Natural History, 1400 S. Lake Shore Dr., Chicago, IL 60605, USA

[4]Morton Arboretum, 4100 Illinois Rte. 53, Lisle, IL 60532, USA

[5]Smithsonian Tropical Research Institute, Apartado 0843-03092, Balboa, Panama

[6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA

*Email for correspondence: tjshen@nchu.edu.tw

Ecology. 2018.

Theorem 1: The aggregation parameter *k* will become 2*k* if two neighbouring quadrats with size of *a* is emerged to form a new larger quadrat with size of 2*a* under the independent negative binomial model.

Proof:

At the spatial scale *a* (that is, the area size of each sampling area is *a*), let abundances of a species respectively in grid cell 1 and grid cell 2 be denoted by *X* and *Y*, moreover they follow the same distribution $NBD(k,u)$, under the independence assumption, the joint probability is simply the product of the two probability functions as

$$P(X=x,Y=y)=\frac{\Gamma(k+x)}{\Gamma(k)\Gamma(x+1)}\left(\frac{k}{u+k}\right)^{k}\left(\frac{u}{u+k}\right)^{x}\times\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{u+k}\right)^{k}\left(\frac{u}{u+k}\right)^{y}$$

At the spatial scale 2*a*, when any two neighbouring grid cells are merged together (here for example, grid cells 1 and 2 are merged), we have the probability of species abundance in the new larger cell as

$$\begin{aligned}P(X+Y=z)&=\sum_{y=0}^{z}P(X=z-y,Y=y)\\&=\sum_{y=0}^{z}\binom{k+z-y-1}{z-y}\binom{k+y-1}{y}\left(\frac{2k}{2u+2k}\right)^{2k}\left(\frac{2u}{2u+2k}\right)^{z}\\&=\binom{2k+z-1}{z}\left(\frac{2k}{2u+2k}\right)^{2k}\left(\frac{2u}{2u+2k}\right)^{z}\end{aligned}$$

Therefore, it is evident that *X*+*Y* follows $NBD(2k,2u)$ when the spatial scale *a* is augmented to 2*a*, species distributional aggregation value is expected to increase from *k* to 2*k* (Johnson and Kotz 1969).

Of course, it may be easier to derive the scale dependency pattern of the aggregation value using probability-generating function. To show this, it is well known that the probability-generating function for the negative binomial model $NBD(k,u)$ is

$$G(z)=\left(\frac{\frac{k}{u+k}}{1-(1-\frac{k}{u+k})z}\right)^k=\left(\frac{k}{u+k-uz}\right)^k$$

At the spatial scale *a*, because the random variates *X* and *Y* in the two neighboring grids follow the same $NBD(k,u)$, their probability-generating functions thus are identical as $G_X(z)=G_Y(z)=G(z)$. When the spatial scale is augmented to 2*a* by merging the two neighboring grids, the new random variate becomes *X*+*Y*. By using the product property of probability-generating function, we have

$$\begin{aligned}
&G_{X+Y}(z)=G_X(z)G_Y(z)\\
&=\left(\frac{k}{u+k-uz}\right)^k\left(\frac{k}{u+k-uz}\right)^k\\
&=\left(\frac{k}{u+k-uz}\right)^{2k}\\
&=\left(\frac{2k}{2u+2k-2uz}\right)^{2k}
\end{aligned}$$

From the last equality of the above derivation, it is clear that $G_{X+Y}(z)$ is the probability-generating function of $NBD(2k,2u)$. Thus, we have the desired result $X+Y\sim NBD(2k,2u)$.

Theorem 2: The marginal probability of species abundance in a small area *a* over the region *A* for the negative trinomial model (Eq. 4 in the main text) is identical to the negative binomial model (Eq. 1 in the main text).

Proof:

From equation 4 in the main text,

$$P(N_a=n,N_{A-a}=N-n)=\frac{\Gamma(k+N)}{\Gamma(k)\Gamma(n+1)\Gamma(N-n+1)}\left(\frac{k}{u^*+k}\right)^k\left(\frac{au^*/A}{u^*+k}\right)^n\left(\frac{(1-a/A)u^*}{u^*+k}\right)^{N-n}$$

By taking $n+m=N$, we thus have the marginal probability for species abundance in a small area a by

$$
\begin{aligned}
P(N_a = n) &= \sum_{m=0}^{\infty} P(N_a = n, N_{A-a} = m) \\
&= \sum_{m=0}^{\infty} \frac{\Gamma(k+n+m)}{\Gamma(k)\Gamma(n+1)\Gamma(m+1)} \left(\frac{k}{u^* + k}\right)^k \left(\frac{au^*/A}{u^* + k}\right)^n \left(\frac{(1-a/A)u^*}{u^* + k}\right)^m \\
&= \sum_{m=0}^{\infty} \frac{\Gamma(k+n+m)}{\Gamma(m+1)\Gamma(k+n)} \times \frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{u^* + k}\right)^k \left(\frac{au^*/A}{u^* + k}\right)^n \left(\frac{(1-a/A)u^*}{u^* + k}\right)^m \\
&= \sum_{m=0}^{\infty} \frac{\Gamma(k+n+m)}{\Gamma(m+1)\Gamma(k+n)} \left(\frac{au^*/A + k}{u^* + k}\right)^{k+n} \left(\frac{(1-a/A)u^*}{u^* + k}\right)^m \frac{\frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{u^* + k}\right)^k \left(\frac{au^*/A}{u^* + k}\right)^n}{\left(\frac{au^*/A + k}{u^* + k}\right)^{(k+n)}} \\
&= \frac{\frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{u^* + k}\right)^k \left(\frac{au^*/A}{u^* + k}\right)^n}{\left(\frac{au^*/A + k}{u^* + k}\right)^{(k+n)}} \\
&= \frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{au^*/A + k}\right)^k \left(\frac{au^*/A}{au^*/A + k}\right)^n
\end{aligned}
$$

Because $u^* = E(N)$ and $u = aE(N)/A = au^*/A$,

$$
\begin{aligned}
P(N_a = n) &= \frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{au^*/A + k}\right)^k \left(\frac{au^*/A}{au^*/A + k}\right)^n, \\
&= \frac{\Gamma(k+n)}{\Gamma(k)\Gamma(n+1)} \left(\frac{k}{u + k}\right)^k \left(\frac{u}{u + k}\right)^n
\end{aligned}
$$

which is the same as Eq. 1 of the main text.

Additional Methods

Species abundance models and evenness indices

The probability of observing a species with abundance i for the Fisher's logseries model (Fisher et al. 1943) was calculated as follows:

$$p(i\,|\,\theta) = -\frac{1}{\ln(1-\theta)}\frac{\theta^i}{i},$$

where θ is a uniform random number over [0.95, 1). Selection of this specific parameter range allows the generated dominance-rank curve to vary from even to uneven patterns. This applies to other models as below.

The geometric series model (Motomura 1932, Chen 2014) was used to simulate species abundance as follows:

$$p(i\,|\,\theta) = \theta(1-\theta)^{i-1},$$

where θ is a uniform random number over (0, 0.1].

For the Preston's canonical lognormal distribution (Preston 1962, Kitzes and Harte 2015, Chen and Shen 2017a), species abundance was generated by

$$p(i\,|\,\theta) = C\frac{e^{\theta^2}}{i\ln 2}e^{-\frac{(\ln i-2\theta^2)^2}{4\theta^2}},$$

where C is a normalization constant to allow $\sum_{n=1}^{S} p(n\,|\,\theta) = 1$ and θ is a uniform random number sampled from [0.5, 3].

We also generated species abundance based on a local zero-sum multinomial model without immigration (Hubbell 2001, Alonso and McKane 2004), which was given by,

$$p(i\,|\,\theta) = C\frac{\theta}{i}(1-i/N)^{\theta-1}.$$

Again, C here is a normalization constant to allow $\sum_{n=1}^{S} p(n \mid \theta) = 1$ and θ is a uniform random number sampled over [1,300].

To measure the evenness of the abundance, we computed the coefficient of variation (CV) for each simulated species abundance distribution as in (Chao and Shen 2003):

$$CV = \frac{\sqrt{\sum_{i=1}^{S} \left(p_i - 1/S \right)^2 \Big/ S}}{1/S},$$

where p_i is the proportion of individuals over *N,* belonging to the *i*-th species over *S*.

In addition, we computed the Shannon's and Simpson's evenness indices (Chen 2015), which were calculated as:

$$\begin{cases} E_{Shannon} = \dfrac{-\sum_{i=1}^{S} p_i \ln p_i}{-\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S}} = \dfrac{-\sum_{i=1}^{S} p_i \ln p_i}{\ln S} \\ E_{Simpson} = \dfrac{1 \Big/ \sum_{i=1}^{S} p_i^2}{1 \Big/ \sum_{i=1}^{S} \left(\frac{1}{S} \right)^2} = \dfrac{1 / \sum_{i=1}^{S} p_i^2}{S} \end{cases}.$$

Variance estimation of the parameters

Following a similar application about estimating species richness in Shen and He (2008), we derive the estimated standard errors of $\hat{k}$ and $\hat{u}$ from the observed information matrix with respect to the likelihood function Eq. (8b). Specifically, the observed information matrix is denoted by

$$\Sigma = -\left[\begin{array}{cc} \dfrac{\partial^2 \ln L(k,u \mid f_1,...,f_M)}{\partial k^2} & \dfrac{\partial^2 \ln L(k,u \mid f_1,...,f_M)}{\partial k \partial u} \\ \dfrac{\partial^2 \ln L(k,u \mid f_1,...,f_M)}{\partial u \partial k} & \dfrac{\partial^2 \ln L(k,u \mid f_1,...,f_M)}{\partial u^2} \end{array} \right]_{(k,u)=(\hat{k},\hat{u})} . \quad \text{(S1)}$$

Consequently, the variances of $\hat{k}$ and $\hat{u}$ can be respectively estimated by the (1, 1) entry and (2, 2) entry of Σ^{-1} (the inverse matrix of Σ), and then the estimated standard errors of $\hat{k}$ and $\hat{u}$ are square roots of the two estimated variances.

Confidence intervals of the parameters

Since $(\hat{k},\hat{u})^T$ is asymptotically distributed from a bivariate normal distribution with mean vector $(k,u)^T$ and variance-covariance matrix Σ^{-1}. Consequently, using the estimated variances of $\hat{k}$ and $\hat{u}$, $(1-\alpha)\times 100\%$ confidence intervals of *k* and *u* can be simply established by

$$\left(\hat{k} - z_{\alpha/2}\sqrt{V\hat{a}r(\hat{k})}, \hat{k} + z_{\alpha/2}\sqrt{V\hat{a}r(\hat{k})}\right) \quad \text{(S2a)}$$

and

$$\left(\hat{u} - z_{\alpha/2}\sqrt{V\hat{a}r(\hat{u})}, \hat{u} + z_{\alpha/2}\sqrt{V\hat{a}r(\hat{u})}\right), \quad \text{(S2b)}$$

respectively. Note that $z_{\alpha/2}$ in the confidence intervals is found by $P(Z \geq z_{\alpha/2}) = \alpha/2$ when the confidence level $1-\alpha$ is given, where *Z* is a standard normal random variable.

Fitting of NMD-derived NBD model against alternative models onto regional species distribution

Note that at the regional scale when the entire forest-plot area is studied without being further divided into sub-regions or subsamples, the NMD model is identical to NBD. To further demonstrate the suitability of the NMD-derived NBD model as the regional SAD to the three

forest plots data, we also considered three additional regional SAD models in this analysis. The three models include metacommunity-level neutral zero-sum model (meta-NZSM), Poisson compound-lognormal model (PL) and Poisson compound-continuous logseries model (PCLS).

The expected number of species in the meta-NZSM model (Hubbell 2001, Alonso and McKane 2004) is formulated by

$$E(S_r|\theta) = \frac{\theta}{r}\frac{\Gamma(J+1)}{\Gamma(J+1-n)}\frac{\Gamma(J+\theta-r)}{\Gamma(J+\theta)}, \tag{S3}$$

where θ is a positive parameter and J is the total number of individuals in the entire community. Thus, the relative abundance for species with r individuals, denoted by $p(r)$, can be derived from normalizing the formula in Eq. S3.

PL (Dornelas and Connolly 2008, Connolly et al. 2009) is derived from

$$p(r) = \int_0^\infty P(X = r|n)\phi(n)\,dn, \tag{S4}$$

where $P(X = r|n)$ is the conditional probability mass function (pmf) of a Poisson random variate with intensity n, and $\phi(n)$ is the probability density function (pdf) of a lognormal distribution having the form of

$$\phi(n) = \frac{1}{n\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(\ln n-\mu\right)^2}{2\sigma^2}\right], \tag{S5}$$

where μ and σ are the mean and standard deviation of log-abundance, respectively.

In analogy to the derivation of the model PL, the relative abundance of species with r individuals in the model PCLS can be derived from Eq. S4 but the lognormal distribution in Eq. S5 is replaced by a continuous logseries distribution whose pdf is expressed by

$$\phi(n) = \frac{\theta^n}{n\rho}, \tag{S6}$$

where $0<\theta<1$, $1\le n<\infty$ and ρ is a normalization constant with a form of

$\rho = \int_{-\ln\theta}^{\infty} \frac{e^{-n}}{n} dn$ (Table 1 of Green and Plotkin (2007)). Since numerically computing the relative abundance directly from the integration in Eq. S4 could lead to an inaccurate result, especially for larger r, the tractable formula of $p(r)$ can be further expressed as

$$\begin{aligned} p(r) &= \int_0^{\infty} \frac{n^{r-1}e^{-n}}{\Gamma(r)} \times \frac{\theta^n}{\phi r} dn - \int_0^1 \frac{n^{r-1}e^{-n}}{\Gamma(r)} \times \frac{\theta^n}{\rho r} dn \\ &= \frac{1}{\rho r \left(1-\ln\theta\right)^r} - \int_0^1 \frac{n^{r-1}e^{-n}}{\Gamma(r)} \times \frac{\theta^n}{\rho r} dn \end{aligned}, \tag{S7}$$

for a positive integer r. Note that the relative abundance of unseen species can be numerically computed by

$$p(0) = \int_1^{\infty} e^{-n} \frac{\theta^n}{\rho n} dn . \tag{S8}$$

Finally, as a comparison, we also fit a local neutral model (local-NZSM) in which immigration is allowed. The computation of the local-NZSM model is given by (Volkov et al. 2003, 2007, Chisholm and Pacala 2010),

$$\left\langle \phi_n \right\rangle = \theta \frac{J!}{n!(J-n)!} \frac{\Gamma(\gamma)}{\Gamma(J+\gamma)} \int_0^{\gamma} \frac{\Gamma(n+y)}{\Gamma(1+y)} \frac{\Gamma(J-n+\gamma-y)}{\Gamma(\gamma-y)} \exp(-y\theta/\gamma)\, dy ,$$

where $\left\langle \phi_n \right\rangle$ represents the expected number of species with abundance n in the local community, $\gamma = m(J-1)/(1-m)$, J is the community size of the local community, θ is the fundamental biodiversity number and m represents the immigration rate. Note that here we use this local-NZSM model as a reference only because it is, strictly speaking, a local sampling model and does not contain the aggregation or shape parameter like the other local models presented in Green and Plotkin (2007).

To conduct model comparison and demonstrate the applicability of NMD model in the three forest plots, we utilize Akaike Information Criterion (AIC) (Akaike 1974), which is computed as,

2p-2log($MLKH$), among three models, where p is the number of parameters in the model and $MLKH$ represents the maximum of the likelihood function calculated above for each probabilistic model. Additionally, we also conducted Kolmogorov-Smirnov (KS) and Chi-squared (χ^2) tests (Table S2). In particular, the KS test has to be adjusted because species abundance is a discrete variable. Finally, we also plotted the fitted curves for the four regional SAD models along with the local-NZSM model as a reference and compared each fitted curve to the observed species abundance patterns visually (Fig. 5 of the main text).

Fitting of NMD against alternative models onto local species distribution

For fitting the proposed NMD model with other alternative models and comparing their performance at different local scales, we use two methods, entire-plot-partitioning and intact-subregion-sampling, to collect local multi-species abundances. To be specific, the entire-plot-partitioning sampling scenario is to partition the entire forest plot into small quadrats with a given sampling unit (e.g., 2×2 m). Species abundance at each quadrat is used for constructing likelihood models for the fitting of the parametric models. By contrast, the intact -subregion sampling scenario is to randomly select an intact (or continuous) subregion from the entire forest plot and the limited local-subregion species abundance information is studied. To this end, the difference between these two local sampling scenarios is that the entire-plot-partitioning method will utilize all abundance information from all species found within the study plot.

For the entire-plot-partitioning sampling scenario, each forest plot with a total area size A (e.g., 1000×500 m in the BCI plot) is divided into q non-overlapping and equal-sized quadrats based on a given sampling quadrat size (i.e., $a = 2 \times 2$ m, 5×5 m and 10×10 m). Therefore, each

quadrat has an area size of a (=A/q). Species abundance information distributed among quadrats are counted and used for constructing corresponding likelihood formulas for the NMD, independent Poisson (Chen and Shen 2017b) and independent NBD models, which were respectively given by

$$\begin{cases} Lik_{NMD} = \prod_{s=1}^{S}\left\{\dfrac{\Gamma(k+N_s)}{\Gamma(k)\prod_{i=1}^{q}\Gamma(n_{s,i}+1)}\left(\dfrac{k}{u^*+k}\right)^k\prod_{i=1}^{q}\left(\dfrac{a_i u^*/A}{u^*+k}\right)^{n_{s,i}}\right\} \\ Lik_{Poisson} = \prod_{s=1}^{S}\prod_{i=1}^{q}\left\{\dfrac{e^{-a_i\lambda}\left(a_i\lambda\right)^{n_{s,i}}}{\Gamma(n_{s,i}+1)}\right\} \\ Lik_{NBDs} = \prod_{s=1}^{S}\prod_{i=1}^{q}\left\{\dfrac{\Gamma(k+n_{s,i})}{\Gamma(k)\Gamma(n_{s,i}+1)}\left(\dfrac{k}{a_i u^*/A+k}\right)^k\left(\dfrac{a_i u^*/A}{a_i u^*/A+k}\right)^{n_{s,i}}\right\} \end{cases}, \quad \text{(S9)}$$

where λ denotes the mean intensity across species in the independent Poisson distribution. Note that, as mentioned above, our sampling quadrats have the same area size, i.e., $a_i = a$ ($i = 1,2,...,q$).

In the intact-subregion-sampling scenario, for a given specific fraction of sampling area (e.g., g = 0.2 in the BCI plot indicated that each intact subregion had an area size of 50 ha × g=10 ha), we randomly sampled 200 intact subregions (these subregions may overlap spatially or not) from the entire forest plot. In this study, we considered four area fractions of local subregions as g = 0.2, 0.4, 0.6 and 0.8.

The alternative local aggregation models fitted here for the specie abundance data collected from a subregion are derived from the compound versions of three regional SAD models—the continuous logseries (Eq. S6), gamma (will be shown below) and lognormal (Eq. S5) models. Note that the exponential model is a special case of gamma model when the shape parameter of

the gamma model is set to 1. These local models were discussed in Green and Plotkin (2007) and expressed by the compound of the corresponding parental regional SAD model and a NBD with the parameters related to the sampling fraction (denoted by *g* here) of the targeted subregion (thus they are called "local aggregation models"). Calculation formulas for these locally aggregated models were presented in Table 2 of Green and Plotkin's paper and can be simply expressed as follows:

$$\phi_g\left(y\right)=\int_0^\infty \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{gn}{k+gn}\right)^y\left(\frac{k}{k+gn}\right)^k \phi\left(n\right)dn\,, \tag{S10}$$

where *y* is the number of individuals observed in the local sample with sampling fraction *g* in contrast with the entire study region, *k* is the aggregation parameter, and $\phi(n)$ is any of regional SADs including the continuous logseries model in Eq. S6, the lognormal model in Eq. S5 and a gamma/exponential model with the pdf as

$$\phi(n)=\frac{\lambda^\beta n^{\beta-1}e^{-\lambda n}}{\Gamma(\beta)}\,, \tag{S11}$$

where β and $1/\lambda$ are the shape and the scale parameters, respectively. Therefore, after applying Eq. S10, the corresponding local aggregation models can have specific names given as NBD compound-continuous logseries model, NBD compound-lognormal model and NBD compound-gamma/exponential model.

The parameters of each of local aggregation models enumerated above were estimated by the conditional likelihood function that is similar to Eq. 8b in the main text and is expressed as follows:

$$L(\boldsymbol{\theta}\mid f_1,...,f_M)=\frac{\Gamma\left(\sum_{j=1}^M f_j+1\right)}{\prod_{j=1}^M\Gamma\left(f_j+1\right)}\times\prod_{j=1}^M\left\{\frac{\phi_g(j)}{1-\phi_g(0)}\right\}^{f_j}\,, \tag{S12}$$

where θ is a vector of unknown parameters involved in the local aggregation model from Eq. S10. Note that for the intact-subregion-sampling scenario studied here, the above likelihood model (Eq. S12) implied that the proposed NMD model becomes identical to NBD again (like the situation at the regional scale, and Eq. S12 becomes identical to Eq. 8b in the main text).

Because we randomly took 200 subregions for a given area fraction, we thus had 200 fitted models and 200 AIC values accordingly for each model. To visually demonstrate the fitting results, we used the box-and-whisker plot to present and compare the distribution of the ACI values for the alternative fitted models against our NMD model (Fig. S2). Note that the likelihood function in Eq. S12 was applied and the AIC value was calculated for each fitted model regarding each randomly selected subregion.

References

Akaike, H. 1974. Information theory as an extension of the maximum likelihood principle. Pages 276–281 in B. Petrov and F. Csaki, editors. Second international symposium on information theory. Akademiai Kiado, Budapest.

Alonso, D., and A. McKane. 2004. Sampling Hubbell's neutral theory of biodiversity. Ecology letters 7:901–910.

Chao, A., and T. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10:429–443.

Chen, Y. 2014. Species abundance distribution pattern of microarthropod communities in SW Canada. Pakistan Journal of Zoology 46:1023–1028.

Chen, Y. 2015. Biodiversity and biogeographic patterns in Asia-Pacific Region I: statistical methods and case studies. Bentham Science Publishers.

Chen, Y., and T. Shen. 2017a. A general framework for predicting delayed responses of ecological communities to habitat loss. Scientific Reports:In press.

Chen, Y., and T. Shen. 2017b. Rarefaction and extrapolation of species richness using an area-based Fisher's logseries. Ecology and Evolution 7:10066–10078.

Chisholm, R., and S. Pacala. 2010. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. PNAS 36:15821–15825.

Conlisk, E., J. Conlisk, and J. Harte. 2007. The impossibility of estimating a negaive binomial clustering parameter from presence-absence data: a comment on He and Gaston. American Naturalist 170:651–654.

Connolly, S., M. Dornelas, D. Bellwood, and T. Hughes. 2009. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. Ecology 90:3138–3149.
Dornelas, M., and S. R. Connolly. 2008. Multiple modes in a coral species abundance distribution. Ecology Letters 11:1008–1016.
Fisher, R., A. Corbet, and C. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42–58.
Giam, X., T. Ng, V. Yap, and H. Tan. 2010. The extent of undiscovered species in Southeast Asia. Biodiversity and Conservation 19:943–954.
Green, J., and A. Ostling. 2003. Endemics-area relationships: the influence of species dominance and spatial aggregation. Ecology 84:3090–3097.
Green, J., and J. Plotkin. 2007. A statistical theory for sampling species abundance. Ecology Letters 10:1037–1045.
He, F., and K. Gaston. 2000a. Estimating species abundance form occurrence. American Naturalist 156:553–559.
He, F., and K. Gaston. 2000b. Occupancy-abundance relationships and sampling scale. Ecography 23:503–511.
He, F., and K. Gaston. 2003. Occupancy, spatial variance, and the abundance of species. American Naturalist 162:366–375.
He, F., and S. Hubbell. 2011. Species–area relationships always overestimate extinction rates from habitat loss. Nature 473:368–371.
He, F., and P. Legendre. 2002. Species diversity patterns derived from species-area models. Ecology 83:1185–1198.
Holt, A., K. Gaston, and F. He. 2002. Occupancy-abundance relationships and spatial distribution: a review. Basic and Applied Ecology 3:1–13.
Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32) (Monographs in Population Biology). Princeton University Press.
Johnson, N., and S. Kotz. 1969. Discrete distributions. Houghton Mifflin, Boston.
Kitzes, J., and J. Harte. 2014. Beyond the species-area relationship: improving macroecological extinction estimates. Methods in Ecology and Evolution 5:1–8.
Kitzes, J., and J. Harte. 2015. Predicting extinction debt from community patterns. Ecology 96:2127–2136.
Motomura, I. 1932. On the statistical treatment of communities. Zoological Magazine (Tokyo) 44:379–383.
Peuyo, S., F. He, and T. Zillio. 2007. The maximum entropy formalism and the idiosyncratic theory of biodiversity. Ecology Letters 10:1017–1028.
Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. Lafrankie, and P. S. Ashton. 2000. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of theoretical biology 207:81–99.
Preston, F. 1962. The canonical distribution of commonness and rarity: part I. Ecology 43:185–215.
Ricklefs, R., and F. He. 2016. Region effects influence local tree species diversity. PNAS:doi:10.1073/pnas.1523683113.
Shen, T., and F. He. 2008. An incidence-based richness estimator for quadrats sampled without

replacement. Ecology 87:2052–2060.
Taylor, L. 1961. Aggregation, variance and mean. Nature 189:732–735.
Taylor, L., I. Woiwod, and J. Perry. 1979. The negative binomial as a dynamic ecological model for aggregation, and the density dependence of k. Journal of Animal Ecology 48:289–304.
Volkov, I., J. Banavar, S. Hubbell, and A. Maritan. 2003. A neutral theory and relative species abundance in ecology. Nature 424:1035–1037.
Volkov, I., J. Banavar, S. Hubbell, and A. Maritan. 2007. Patterns of relative species abundance in rainforests and coral reefs. Nature 450:45–49.
Wilber, M., J. Kitzes, and J. Harte. 2015. Scale collapse and the emergence of the power law species-area relationship. Global Ecology and Biogeography:DOI: 10.1111/geb.12309.
Xu, W., G. Chen, C. Liu, and K. Ma. 2015. Latitudinal differences in species abundance distributions, rather than species aggregation, explain beta-diversity along latitudinal gradients. Global Ecology and Biogeography:DOI: 10.1111/geb.12331.
Zillio, T., and F. He. 2010. Modeling spatial aggregation of finite populations. Ecology 91:3698–3706.

Additional Figures and Tables

Fig. S1. Boxplots are displayed to show the distribution of the numbers of unsampled species in the 2000 random replicates selected from each of the three forest plots, where five summary statistics including the minimum (min), the first, second, third quartiles $\left(Q_1, Q_2, Q_3\right)$, and the maximum (max) are given for reference.

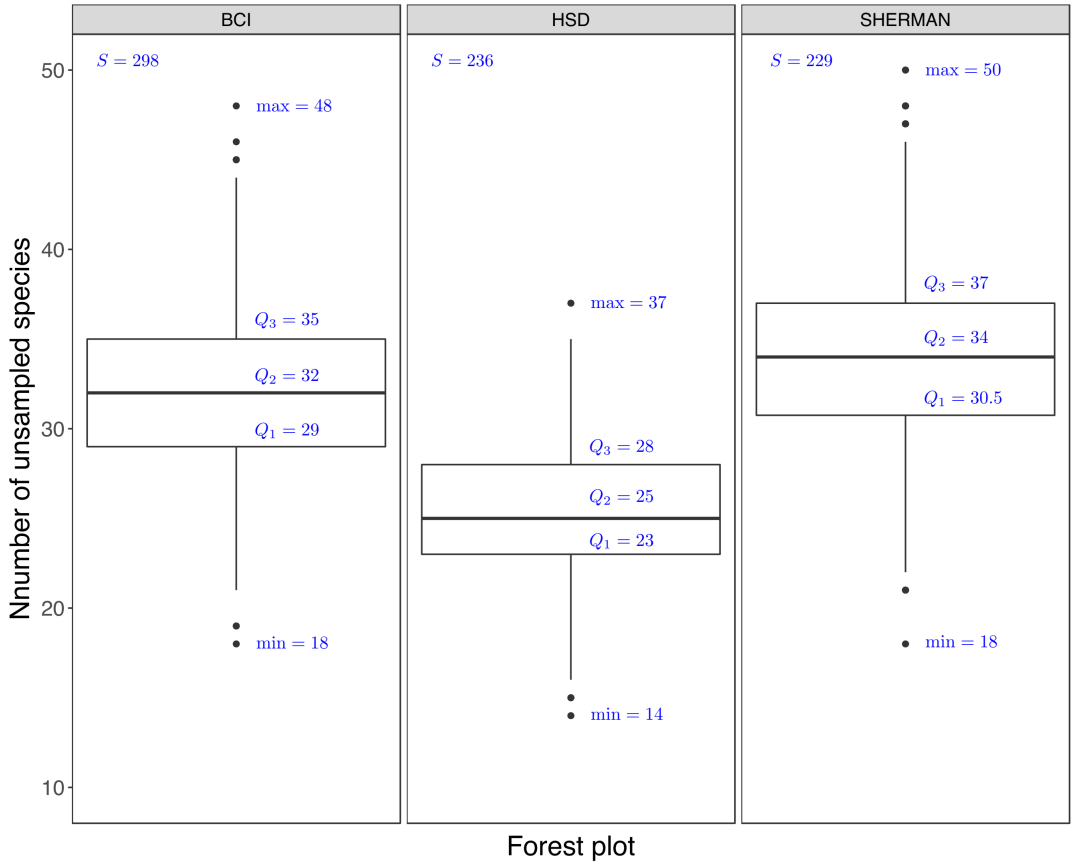
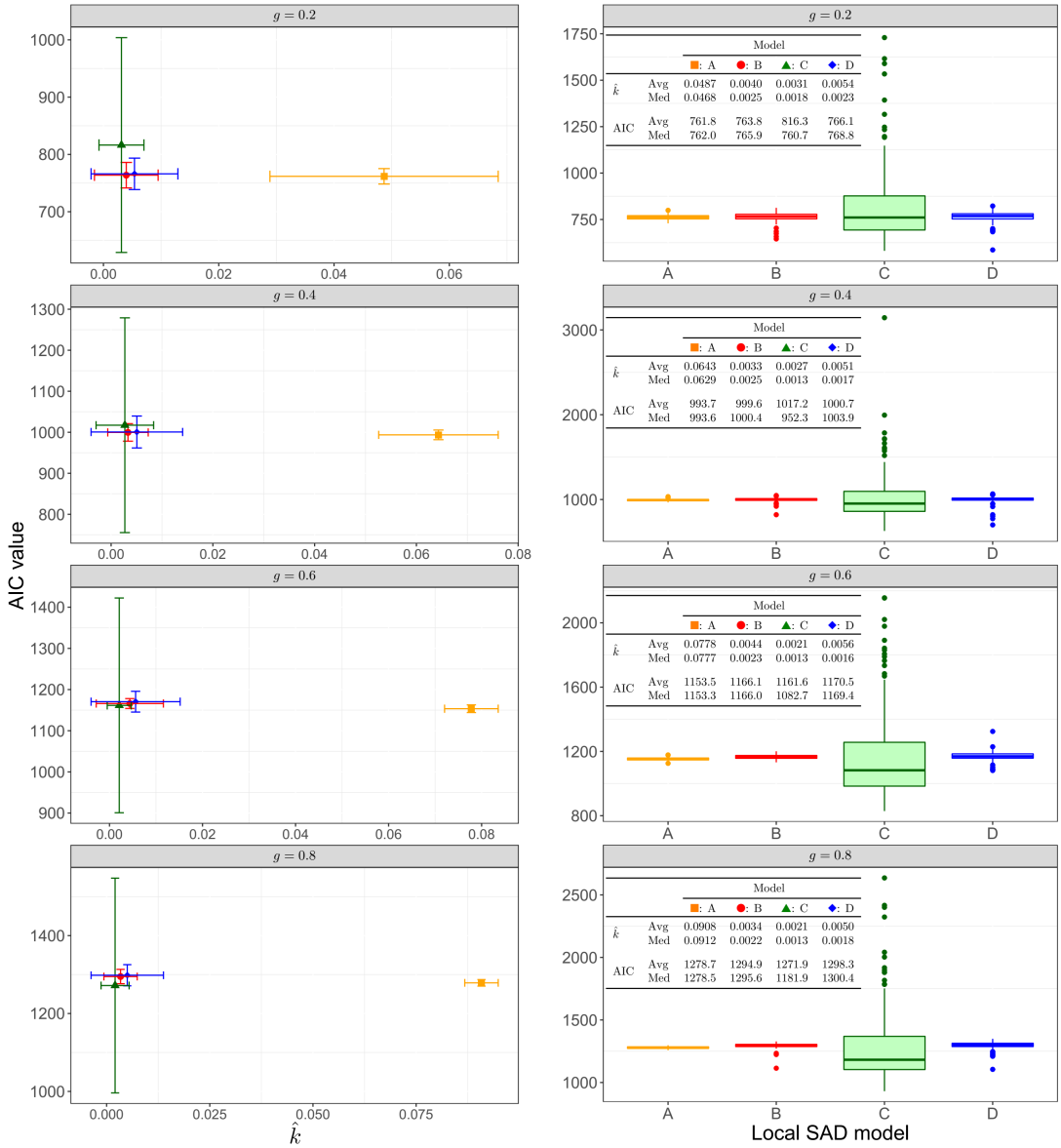
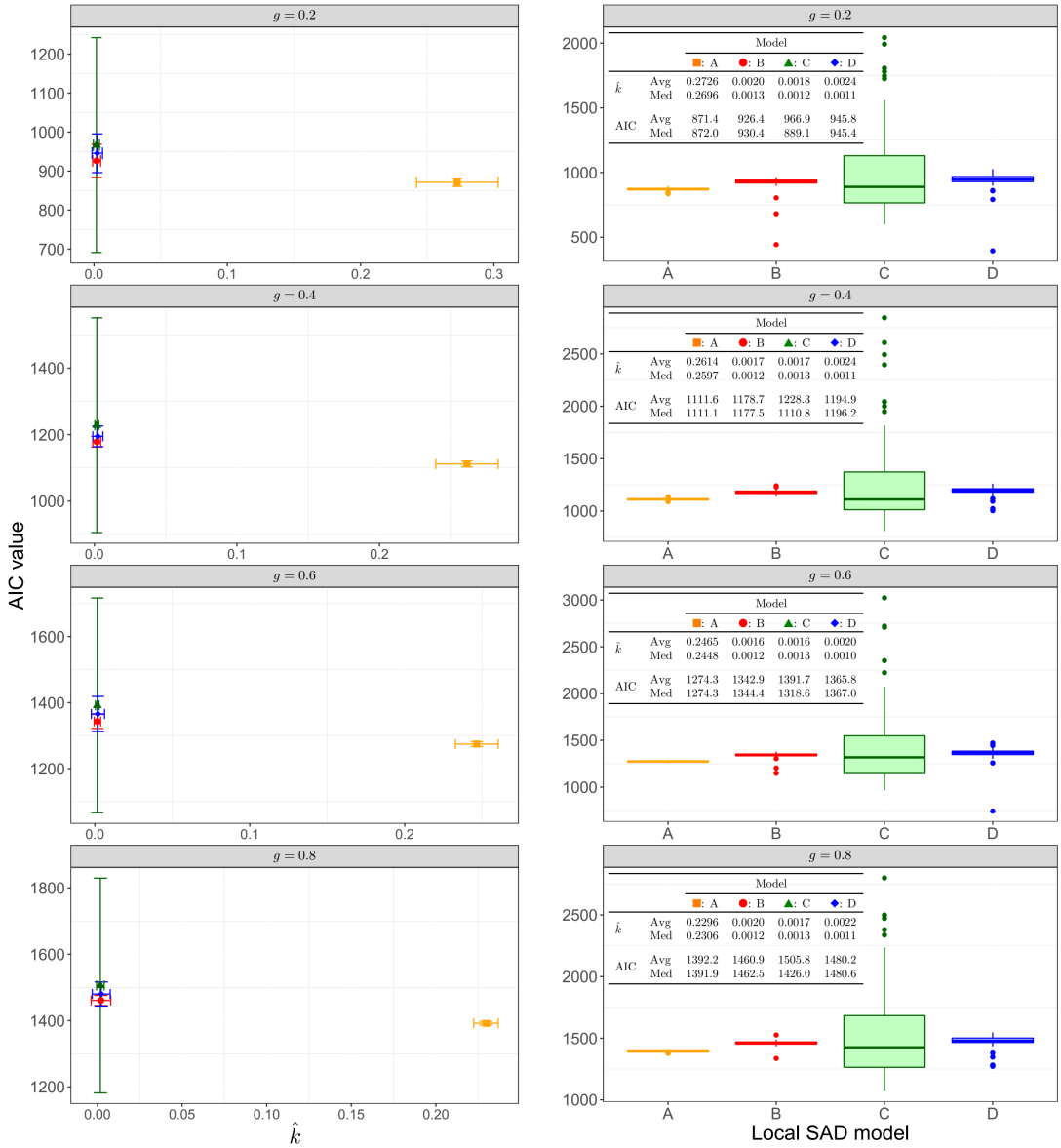


Fig. S2. Fitting of the NMD model (annotated by A: ■) and other alternative locally aggregated models from Green and Plotkin (2007)'s paper (B: ● for the NBD compound-continuous logseries model, C: ▲ for the NBD compound-gamma/exponential model, and D: ◆ for the NBD compound-lognormal model), to the local species abundance data that are collected from each of the 200 randomly selected subregions with a given area size in the three forest plots (a-BCI; b-HSD and c-Sherman). In each of the subplot, the left panel showed the relationship between the estimated k values and the AIC values respectively by horizontal and vertical error bars (mean ± standard error), while the right panel compared boxplots of the AIC distributions for SAD models. Additionally, each table superimposed on the right panel provides detailed information about two location parameters (Avg for the average; Med for the median) calculated from 200 estimated k values and 200 AIC values for each scenario. g represents the fraction of area size of a selected intact subregion with respect to the area size of the entire region. For example, g = 0.2 in the BCI plot indicated that each sampled subregion had an area size of 50 ha × g=10 ha. Note that the Sherman plot has a L-shape geometry and comprises a square plot of 140 × 140 m and a rectangle plot of 100 × 400 m; thus it is unable to consistently use the same sampling scheme as in the BCI and HSD plots which are of the same shape of 1000 × 500 m. Instead, each of 200 subregions in the Sherman plot was composed of small quadrats randomly selected from the plot, where the number of selected quadrats is determined by the sampling fraction g.

a)


g = 0.2
g = 0.4
g = 0.6
g = 0.8
AIC value
$\hat{k}$
Local SAD model
A
B
C
D
Model
■: A ●: B ▲: C ◆: D
$\hat{k}$ Avg 0.0487 0.0040 0.0031 0.0054
Med 0.0468 0.0025 0.0018 0.0023
AIC Avg 761.8 763.8 816.3 766.1
Med 762.0 765.9 760.7 768.8
$\hat{k}$ Avg 0.0643 0.0033 0.0027 0.0051
Med 0.0629 0.0025 0.0013 0.0017
AIC Avg 993.7 999.6 1017.2 1000.7
Med 993.6 1000.4 952.3 1003.9
$\hat{k}$ Avg 0.0778 0.0044 0.0021 0.0056
Med 0.0777 0.0023 0.0013 0.0016
AIC Avg 1153.5 1166.1 1161.6 1170.5
Med 1153.3 1166.0 1082.7 1169.4
$\hat{k}$ Avg 0.0908 0.0034 0.0021 0.0050
Med 0.0912 0.0022 0.0013 0.0018
AIC Avg 1278.7 1294.9 1271.9 1298.3
Med 1278.5 1295.6 1181.9 1300.4

b)


g = 0.2
g = 0.4
g = 0.6
g = 0.8
AIC value
$\hat{k}$
Model
■: A ●: B ▲: C ◆: D
$\hat{k}$ Avg 0.2726 0.0020 0.0018 0.0024
Med 0.2696 0.0013 0.0012 0.0011
AIC Avg 871.4 926.4 966.9 945.8
Med 872.0 930.4 889.1 945.4
$\hat{k}$ Avg 0.2614 0.0017 0.0017 0.0024
Med 0.2597 0.0012 0.0013 0.0011
AIC Avg 1111.6 1178.7 1228.3 1194.9
Med 1111.1 1177.5 1110.8 1196.2
$\hat{k}$ Avg 0.2465 0.0016 0.0016 0.0020
Med 0.2448 0.0012 0.0013 0.0010
AIC Avg 1274.3 1342.9 1391.7 1365.8
Med 1274.3 1344.4 1318.6 1367.0
$\hat{k}$ Avg 0.2296 0.0020 0.0017 0.0022
Med 0.2306 0.0012 0.0013 0.0011
AIC Avg 1392.2 1460.9 1505.8 1480.2
Med 1391.9 1462.5 1426.0 1480.6
A B C D
Local SAD model

c)

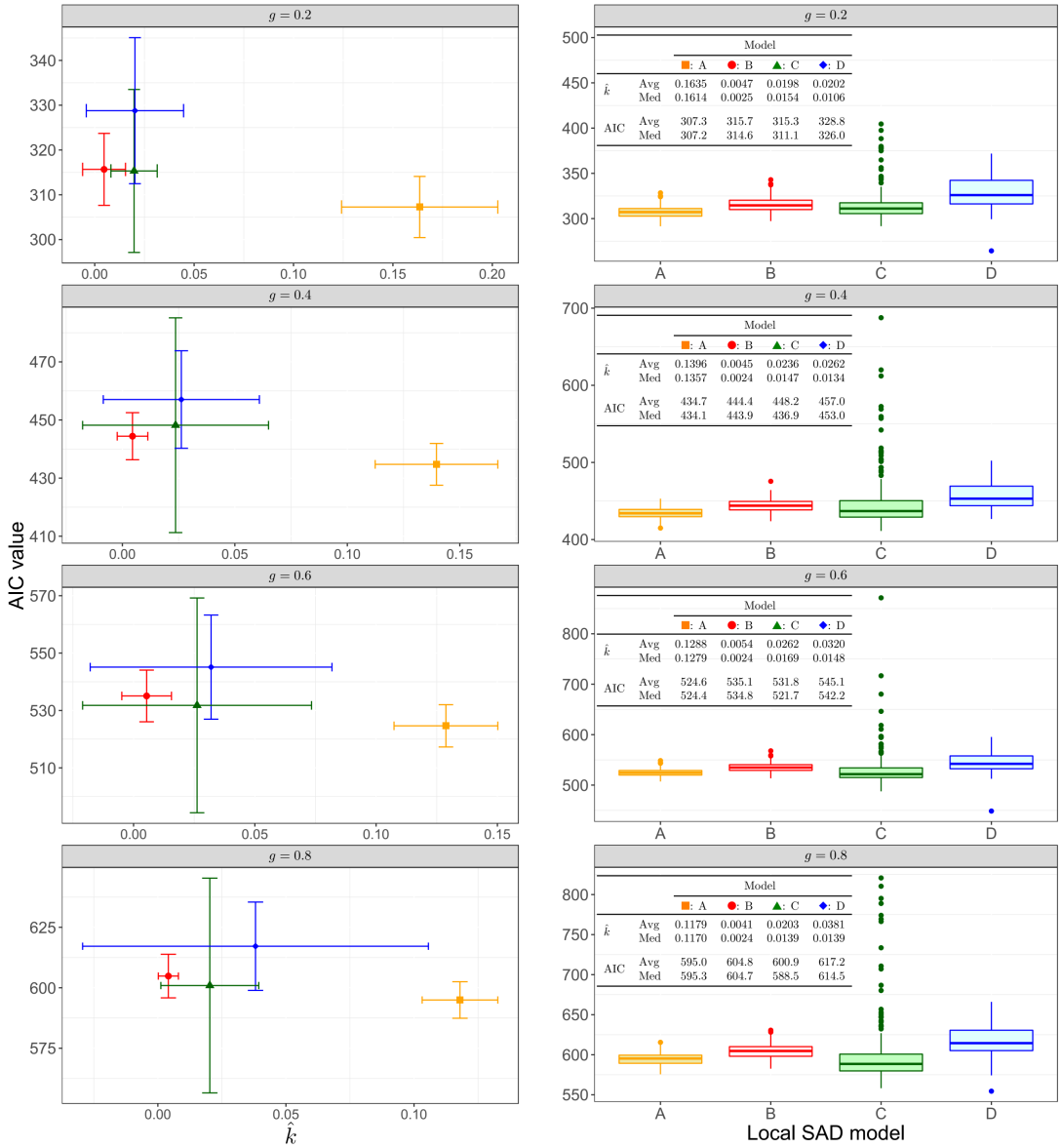

g = 0.2
g = 0.4
g = 0.6
g = 0.8
AIC value
k̂
Local SAD model
Model
■: A ●: B ▲: C ◆: D
k̂ Avg 0.1635 0.0047 0.0198 0.0202
Med 0.1614 0.0025 0.0154 0.0106
AIC Avg 307.3 315.7 315.3 328.8
Med 307.2 314.6 311.1 326.0
k̂ Avg 0.1396 0.0045 0.0236 0.0262
Med 0.1357 0.0024 0.0147 0.0134
AIC Avg 434.7 444.4 448.2 457.0
Med 434.1 443.9 436.9 453.0
k̂ Avg 0.1288 0.0054 0.0262 0.0320
Med 0.1279 0.0024 0.0169 0.0148
AIC Avg 524.6 535.1 531.8 545.1
Med 524.4 534.8 521.7 542.2
k̂ Avg 0.1179 0.0041 0.0203 0.0381
Med 0.1170 0.0024 0.0139 0.0139
AIC Avg 595.0 604.8 600.9 617.2
Med 595.3 604.7 588.5 614.5
A
B
C
D

Table S1. **A review of ecological literature applying the independent negative binomial model to study species aggregation pattern.**

Study topic	References
1. Abundance estimation from occupancy maps; species occupancy-abundance relationship	(He and Gaston 2000a, 2000b, 2003, Holt et al. 2002, Conlisk et al. 2007, Zillio and He 2010)
2. Maximum entropy theory of ecology	(Peuyo et al. 2007, Wilber et al. 2015)
3. Species area relationship or endemic species area relationship	(Plotkin et al. 2000, He and Legendre 2002, Green and Ostling 2003)
4. Species richness estimate and species discovery	(Giam et al. 2010, Ricklefs and He 2016)
5. Species extinction risk or extinction debt estimation	(He and Hubbell 2011, Kitzes and Harte 2014, 2015, Chen and Shen 2017a)
6. Beta diversity or between-site species similarity	(Xu et al. 2015, Chen and Shen 2017a)
7. Population fluctuation: Taylor's power law	(Taylor 1961, Taylor et al. 1979)

Table S2. A comparison of fitting different regional SAD models to tree abundances of different forest plots. Non-significant high *p*-value in KS or χ^2 tests indicates the fitting of the model onto empirical abundance data is better. By contrast, a lower AIC value indicates the fitted model is better.

Forest plot	SAD parameters estimation	KS test		χ^2 test		AIC
		Statistics	*p*-value	Statistics	*p*-value	
BCI						
	NBD $(\hat{k},\hat{u}) = (0.100, 391.6)$	0.06	0.60	53.5	0.31	1,393.6
	meta-NZSM $\hat{\theta} = 33.9$	0.12	0.03	64.8	0.08	1,406.4
	PL $(\hat{u},\hat{\sigma}) = (59.9, 2.5)$	0.13	0.02	55.06	0.26	1,386.4
	PCLS $\hat{\theta} = 0.9999$	0.06	0.57	55.4	0.25	1,403.9
HSD						
	NBD $(\hat{k},\hat{u}) = (0.213, 940.7)$	0.09	0.41	45.4	0.60	1,492.5
	meta-NZSM $\hat{\theta} = 25.6$	0.10	0.29	58.7	0.18	1,537.9
	PL $(\hat{u},\hat{\sigma}) = (149.1, 2.6)$	0.20	0.003	58.1	0.18	1,536.3
	PCLS $\hat{\theta} = 0.9999$	0.09	0.38	54.3	0.29	1,527.7
SHERMAN						
	NBD $(\hat{k},\hat{u}) = (0.114, 50.1)$	0.05	0.70	42.9	0.69	648.5
	meta-NZSM $\hat{\theta} = 35.4$	0.07	0.32	51.8	0.36	653.7
	PL $(\hat{u},\hat{\sigma}) = (13.4, 2.3)$	0.11	0.04	51.4	0.38	674.3
	PCLS $\hat{\theta} = 0.9981$	0.11	0.03	58.0	0.19	661.4