

Supporting Information for
Application of a Negative Multinomial Model Gives Insight into Rarity-Area
Relationships

Additional Methods

Detailed derivation of Eq. 5 in the main text

The joint distribution of $N_{i1}, N_{i2}, \dots, N_{iq}$ in Eq. 5 can be derived from the product of two probability functions by

$$\begin{aligned}
 & P(N_{i1} = n_1, \dots, N_{iq} = n_q | \lambda) \\
 &= P(N_{i1} = n_1, \dots, N_{iq} = n_q | N_{i,A} = n, \lambda) \times P(N_{i,A} = n | \lambda) \\
 &= \frac{\Gamma(n+1)}{\prod_{i=1}^q \Gamma(n_i + 1)} \prod_{i=1}^q \left(\frac{a_i}{A} \right)^{n_i} \times \frac{e^{-\lambda} \lambda^n}{n!} \\
 &= \prod_{i=1}^q \left[\left(\frac{a_i}{A} \lambda \right)^{n_i} \times \frac{e^{-\frac{a_i}{A} \lambda}}{n_i!} \right]
 \end{aligned}$$

Detailed derivation of Eq. 7 in the main text

The marginal distribution of $N_{i,a}$ can be accomplished by summing up all possibilities of $N_{i,A-a}$ in the joint probability distribution (Eq. 6 in the main text). That is

$$\begin{aligned}
& P(N_{i,a} = n | k, u) \\
&= \sum_{y=0}^{\infty} P(N_{i,a} = n, N_{i,A-a} = y | k, u) \\
&= \sum_{y=0}^{\infty} \frac{\Gamma(k+n+y)}{\Gamma(n+1)\Gamma(y+1)\Gamma(k)} \left(\frac{u}{u+A}\right)^k \left(\frac{a}{u+A}\right)^n \left(\frac{A-a}{u+A}\right)^y \\
&= \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} \left(\frac{u}{u+a}\right)^k \left(\frac{a}{u+a}\right)^n \times \\
&\quad \sum_{y=0}^{\infty} \frac{\Gamma(k+n+y)}{\Gamma(k+n)\Gamma(y+1)} \left(\frac{u}{u+a}\right)^{-k} \left(\frac{a}{u+a}\right)^{-n} \left(\frac{u}{u+A}\right)^k \left(\frac{a}{u+A}\right)^n \left(\frac{A-a}{u+A}\right)^y \\
&= \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} \left(\frac{u}{u+a}\right)^k \left(\frac{a}{u+a}\right)^n \sum_{y=0}^{\infty} \frac{\Gamma(k+n+y)}{\Gamma(k+n)\Gamma(y+1)} \left(\frac{u+a}{u+A}\right)^k \left(\frac{u+a}{u+A}\right)^n \left(\frac{A-a}{u+A}\right)^y \\
&= \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} \left(\frac{u}{u+a}\right)^k \left(\frac{a}{u+a}\right)^n \sum_{y=0}^{\infty} \frac{\Gamma(k+n+y)}{\Gamma(k+n)\Gamma(y+1)} \left(\frac{u+a}{u+A}\right)^{k+n} \left(\frac{A-a}{u+A}\right)^y .
\end{aligned}$$

Since the summand of the summation in the last equality above is also a standard NBD with success probability $\frac{(u+a)}{(u+A)}$ and aggregation parameter $k+n$. As a consequence, the summation is equal to one, and thus we have the desired result in Eq. 7 of the main text.

Variance estimation of the parameters

The observed information matrix with respect to the likelihood function (8) in the main text is denoted by

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

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).

Alternative likelihood model not accounting for the unseen species

Let $f_m = \sum_{i=1}^S I(N_{i,a} = m)$ denote the number of species with k individuals in the sampled community, thus $(f_1, f_2, \dots, f_\tau)$ shall follow a multinomial distribution with total S and cell probabilities $(\rho_1, \rho_2, \dots, \rho_\tau)$, where $\rho_n = P(N_{i,a} = n)$ is from (Eq. 4) and

$\tau = \max\{N_{1,a}, N_{2,a}, \dots, N_{s,a}\}$, so the likelihood function is expressed by

$$L_1(k, u | f_1, \dots, f_\tau) = \frac{(\sum_{j=1}^{\tau} f_j)!}{\prod_{j=1}^{\tau} f_j!} \prod_{j=1}^{\tau} \rho_j^{f_j} . \quad (\text{S2})$$

R code for calculating the maximum likelihood estimates of k and u , their estimated standard errors, and 95% confidence intervals

```

rm(list = ls())
## Loading the R package "nlme" for the optimization of
## the maximum likelihood function
library(nlme)

## The likelihood function in Eq. 8 #####
## Input data: f is the species frequency counts
## m is the area size denoted by "a" in the main text
cond.logf <- function(x, f, m=1) {
  ## parameter: alpha
  a <- x[1]
  ## parameter: beta
  b <- x[2]

  pp = b/(b + m)

  zz = which(f > 0)
  rhoN = lgamma(zz + a) - lgamma(zz + 1) - lgamma(a) + zz *
    log1p(-pp) + a * log1p(pp - 1) - log(1 - pp^a)
  const = lgamma(sum(f)+1)-sum(lgamma(f+1))
  res <- -sum(f[zz] * rhoN)-const

  res
}

## Calculate the maximum likelihood estimates of alpha and beta,
## their estimated standard errors, and 95% confidence intervals
Result = function(f, m=1){
  cond.logfSol = nlm(b=c(0.2, 0.2), cond.logf, f = f, m = m,
    lower = c(1e-05, 1e-05), upper = c(1000, 1000))

  ### point estimates
  cond.ahat = cond.logfSol$par[1]
  cond.bhat = cond.logfSol$par[2]

  ### estimated variances of alpha and beta
  Fisher.Info = (fdHess(c(cond.ahat, cond.bhat), f = f, m = m, fun = cond.logf)$Hessian)
  var.a = solve(Fisher.Info)[1,1]
  var.b = solve(Fisher.Info)[2,2]

  out=list()

```

```

mle = c(cond.ahat,cond.bhat)
mle = rbind(mle,c(sqrt(var.a),sqrt(var.b)))
rownames(mle)=c("estimate","estimated se")
colnames(mle)=c("alpha","beta")
out[["MLEs and estimated se of alpha and beta"]]= mle

ci=c(lower=round(cond.ahat-1.96*sqrt(var.a),5),
      upper=round(cond.ahat+1.96*sqrt(var.a),5))
ci = cbind(ci, c(lower=round(cond.bhat-1.96*sqrt(var.b),5),
                  upper=round(cond.bhat+1.96*sqrt(var.b),5)))
colnames(ci)=c("alpha","beta")

out[["95% Confidence Intervals"]]= ci

out
}

### Tested by a hypothetical data set #####
### 131 is the number of singletons, 107 is the number of doubletons, ### and so forth.
f=c(131, 107, 72, 61, 49, 38, 26, 23, 15,
    18, 17, 22, 6, 10, 3, 7, 8, 8, 4, 3,
    1, 1, 3, 3, 1, 2, 2, 1, 0, 0, 1, 1, 2,
    1, 1, 0, 0, 1, 2, 1, 0, 0, 1, 0, 0, 0,
    0, 1, 0, 0, 0, 0, 0, 0, 1)

## Start computing
Result(f=f, m=1)

#### Estimating results by our R code
# $`MLEs and estimated se of alpha and beta`
#           alpha      beta
# estimate    0.50179833 0.12075328
# estimated se 0.07975427 0.01297537
#
# $`95% Confidence Intervals`
#           alpha      beta
# lower 0.34548 0.09532
# upper 0.65812 0.14619

```

Additional Tables and Figures

Table S1. Estimated values of k , u , their 95% confidence intervals (CIs) and CV , and the corresponding maximal log-likelihood values ($\ln L_1(k, u)$) for the two forest plots using Eq. S2.

For both plots, the whole plot size is the same as $A = 50$ ha.

Plot	k		u		CV_A	$\ln L_1(k, u)$
	Estimate	95% CI	Estimate	95% CI		
BCI	0.30694	(0.27, 0.35)	0.022095	(0.017, 0.027)	1.81	-748.5
HSD	0.39020	(0.32, 0.46)	0.080393	(0.055, 0.105)	1.60	-372.5

Fig. S1. Multiscale relationships between the aggregation parameter k and another parameter u estimated from the randomly sampled local communities in the two permanent forest plots.

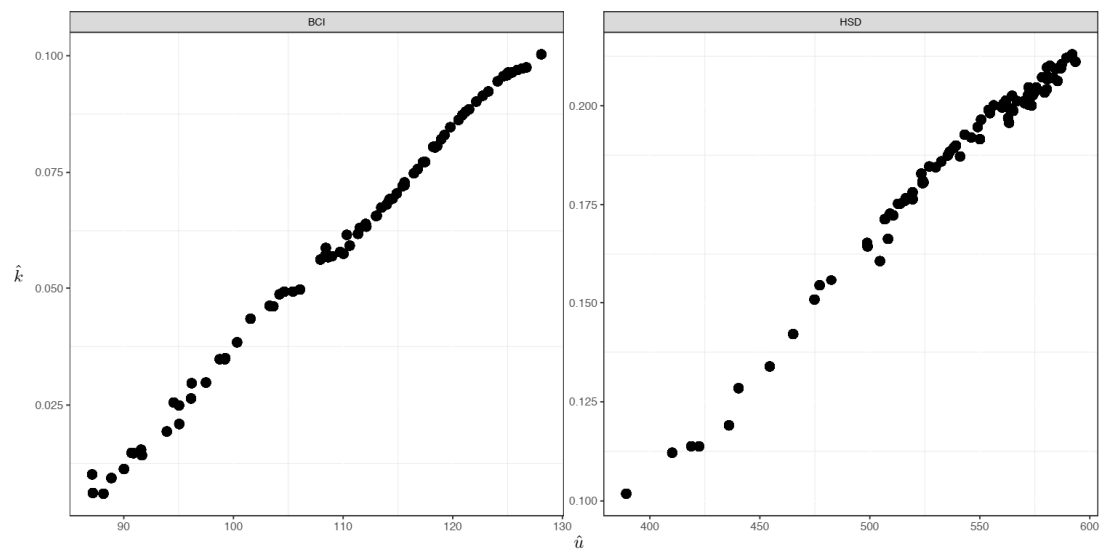


Fig. S2. The theoretical relationships between parameter u and summed probability of having small population size ($\leq 10, 20, 30$) for a single species under the conditions of different fixed k values. In all the subplots, the theoretical area size A is fixed to be 1.

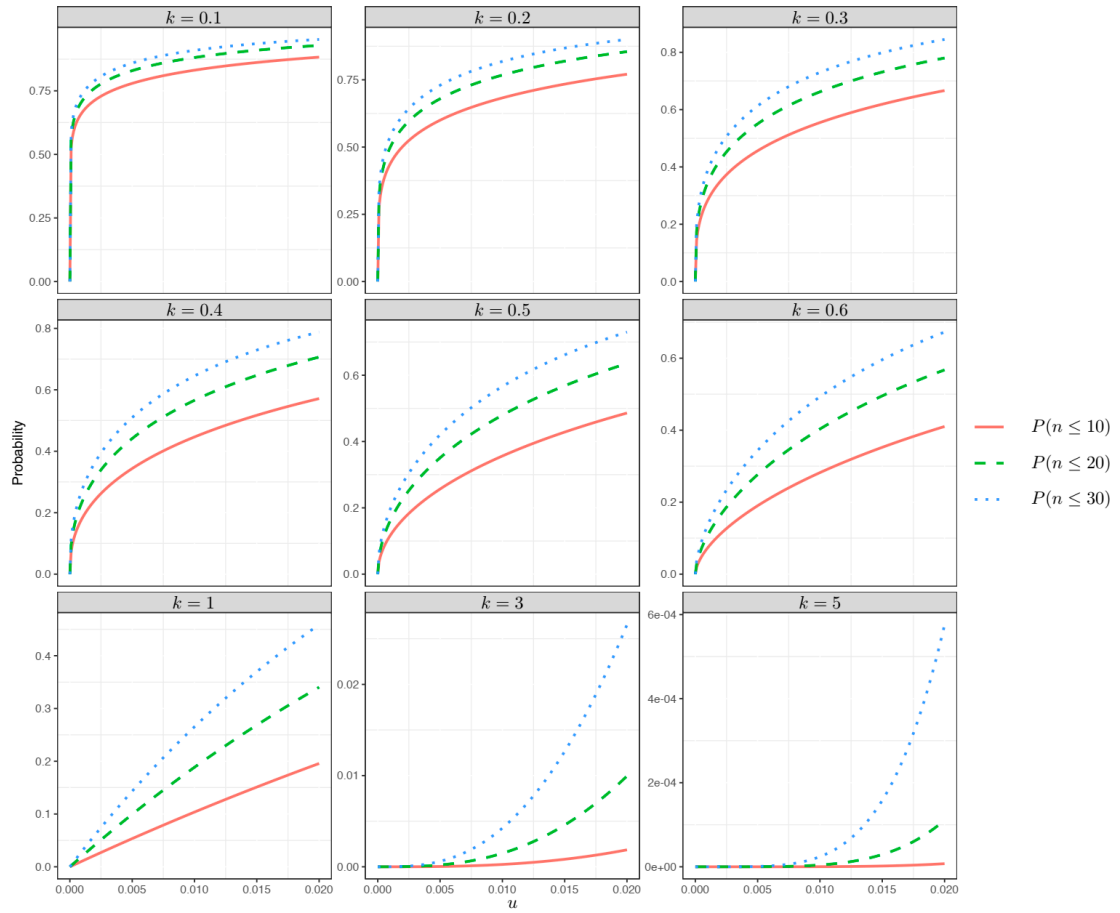


Fig. S3. Multiscale relationships between the ratio a/u and the total abundance of different species (i.e., community size) in the two permanent forest plots. Here a represented the area size (and therefore the spatial scale) for the randomly sampled local communities ($a=A$ when the sampling area is the whole plot), while u was estimated from these sampled communities.

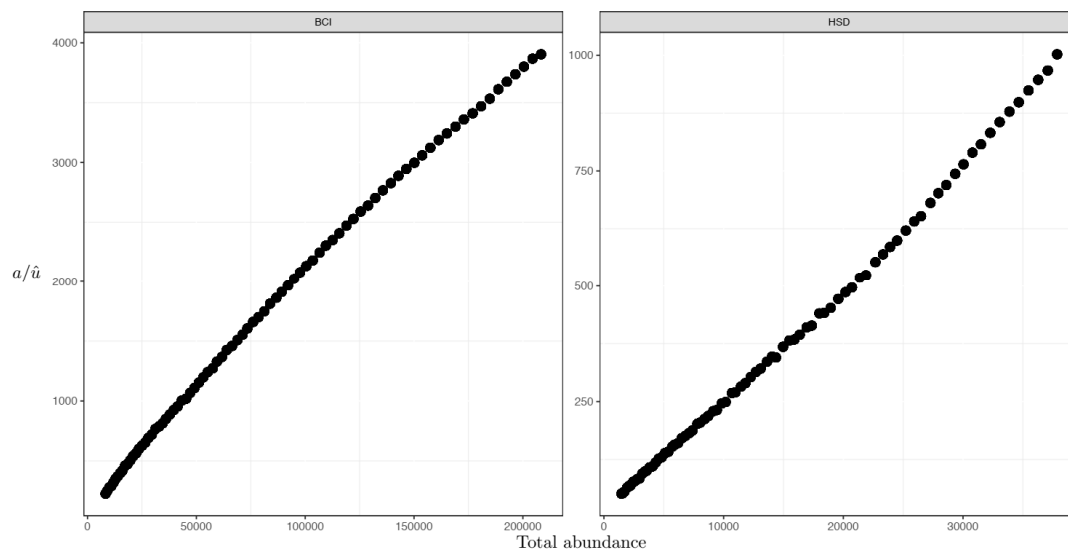


Fig. S4. Multiscale relationships between the ratio u/a and the CV of species abundances in the two permanent forest plots. Here a represented the area size (and therefore the spatial scale) for the randomly sampled local communities ($a=A$ when the sampling area is the whole plot), while both u and CV were estimated from these sampled communities.

