Oikos

o20751

Rosindell, J. and Cornell, S. J. 2013. Universal scaling of species-abundance distributions across multiple scales. – Oikos 122: 1101–1111.

Appendix 1 and 2

Universal scaling of species-abundance distributions across multiple scales supplementary material

James Rosindell [1,2,*] and Stephen J. Cornell [1]

17th October 2012

1 Faculty of Biological Sciences, The University of Leeds, Leeds, LS2 9JT, UK

2 Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire, SL5 7PY, UK

* Corresponding author, both authors contributed equally to this work

E-mails James@Rosindell.org and S.J.Cornell@Leeds.ac.uk

Appendix 1: Fits to species abundance distributions at small areas

We show in the main body of the paper that the species abundance distributions (SADs) for intermediate to large survey areas reduce to a family of curves described by a single parameter $A\nu L^{-2}$, which represents the ratio of the survey area A to the average species range. This description breaks down at smaller areas, where the dispersal distance L introduces an additional length scale to the geographical length scale $L\nu^{-\frac{1}{2}}$.

In this Appendix we investigate empirically the SADs at smaller scales by fitting a variety of analytical forms to them. The candidate SADs we used were:

One parameter distributions:

- Exponential (i.e. geometric) $\phi_n \propto e^{-\lambda n}$
- Log-series $\phi_n \propto \frac{\theta^n}{n}$. This is the SAD for very large sample areas (see the main text).
- Poisson $\phi_n = \frac{\lambda^n}{n!}$. This represents individuals of each species being uniformly and independently distributed over all space.

Two-parameter distributions

- Negative binomial $\phi_n \propto \frac{\Gamma(k+n)}{n!}\left(1+\frac{k}{m}\right)^{-n}$. This represents a Poisson mixture model, where individuals of each species are independently distributed but different species have local densities that follow a Gamma distribution **?**.
- Gamma $\phi_n \propto n^{k-1}e^{-\frac{n}{\theta}}$
- Weibull $\phi_n \propto n^{k-1}e^{-(n/\lambda)^k}$
- Stretched exponential $\phi_n \propto e^{-(n/\lambda)^k}$

Three of these models (Log-series, Poisson, Negative binomial) have particular mechanistic interpretations as described above. The Gamma, Weibull, and Stretched Exponential are usually used for data that take continuous rather than discrete values, but are included here to provide a heuristic comparison with the Negative Binomial.

Statistical procedure

A sample area contains a fixed number of individuals, consequently the abundances of different species are not independent (the abundance of one species can only increase at the expense of another). While it is possible to take account of this when computing sampling formulae for the spatially implicit system **??**, we are not able to extend those methods to the spatially explicit case. However, we can treat the species abundances as approximately independent provided (i) it is very rare for any species to have abundance comparable to the total survey area and (ii) we have many independently sampled survey areas (for these data, from 98 to 521 independent samples were obtained for each parameter set). Under these assumptions, if on average we expect ϕ_n species of abundance n then the likelihood that, over N samples, we see s_1 species of abundance 1, s_2 species of abundance 2 etc is

$$P(\{s_i\}) = \prod_{i=1}^{n} e^{-N\phi_i} \frac{(N\phi_i)^{s_i}}{s_i!}. \tag{1}$$

A maximum likelihood fit for each of the candidate distribution was found by maximizing $P(\{s_i\})$ over the parameters of the distribution (using the `optim()` and `optimize()` functions in R **?**). A goodness of fit for the model was obtained by simulating a sample from the maximum likelihood fit to the actual data, fitting the new maximum likelihood parameters to the simulated data, and counting the fraction p of realizations for which the

simulated data had a higher maximum likelihood than the original data. The statistic p is a p-value for the test of the null hypothesis that the original data represent independent species with a distribution generated by a model of the candidate class, because under that null hypothesis we would expect p to be uniformly distributed between 0 and 1.

Results

Table 1 shows the p-values, maximized log-likelihoods, and fitted maximum likelihood parameters for fits to nine different SAD data sets (sample area 16×16, Gaussian kernel with $L \in \{8, 16, 32\}$, speciation rates $\nu \in \{10^{-6}, 10^{-4}, 10^{-2}\}$). Figure A1 shows a comparison of the fitted distributions with the original data. A spreadsheet containing the fitted values and goodness of fit indicators for the other parameter combinations we simulated is included as further online material . Several clear patterns emerge:

- The Negative Binomial, Gamma, and Weibull distributions all give good fits to the data, with comparable p-values (> 0.05 in all cases) when the same data are fitted to each of the three distributions. In figure A1 the fitted negative binomial (black) and gamma (blue) distributions are practically indistinguishable, while the Weibull (magenta) distribution is also very similar

- The stretched exponential only gives good fits to the four data sets for which there are relatively few abundance values ($L = 32$ with $\nu = 10^{-2}$, 10^{-4},10^{-6}, and $L = 16$ with $\nu = 10^{-2}$, giving estimated $p < 0.0001$ for the others.

- The exponential distribution shows a similar pattern, though generally with worse fits, than the stretched exponential

- The log-series and Poisson distributions give very poor fits to the data.

Conclusions

We can draw the following conclusions:

1. The data are clearly good enough to distinguish between distributions. In particular, we can reject the log-series (which describes SADs for large areas) and the Poisson distribution.

2. Three of the two-parameter distributions fit the data equally well, but a fourth one fits much worse. The ones that fit the distribution well have an algebraic dependence (or approximately so, for the negative binomial) at small abundances, whereas the ill fitting one does not, so we can conclude that this is an essential feature of the data. On the other hand, the data do not have the resolving power to distinguish between a stretched exponential behavior at large abundances (Weibull) and a simple exponential one (Gamma).

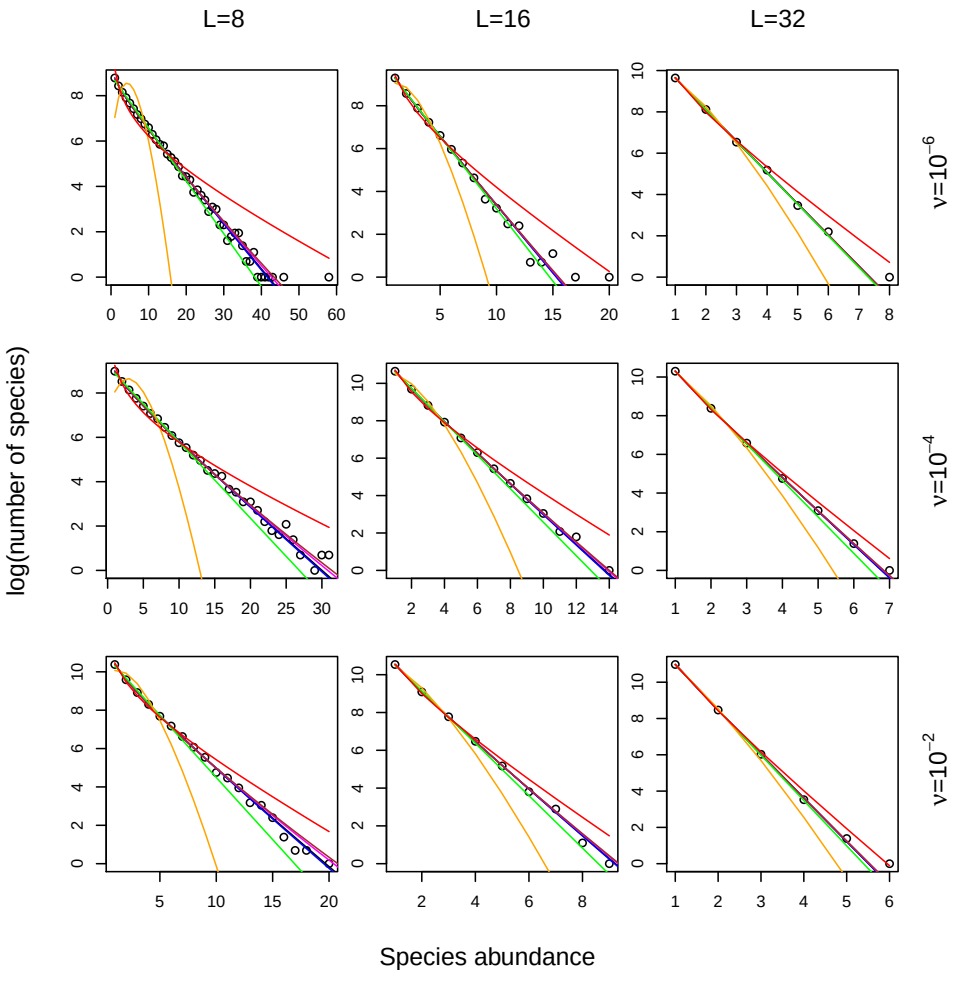
Table 1: Goodness of fit (p-value), maximized log-likelihood (LL), and maximum likelihood parameter values for the seven candidate models fitted to nine species abundance distributions. Gaussian kernel, survey area 16×16. The p-values are estimated from 10000 simulated data sets from the candidate distribution.

ν	L	Neg. Bin.	Gamma	Weibull	Str. Exp.	Exponential	Log-series	Poisson
10^{-6}	8	p =0.9025	p =0.9051	p =0.8531	p =0.0001	p =0.0000	p =0.0000	p =0.0000
		LL=-148.17	LL=-147.13	LL=-151.60	LL=-157.65	LL=-235.23	LL=-1859.99	LL=-22736.59
		m =3.508	θ =5.051	λ =3.959	λ =3.252	λ =0.233	θ =0.927	μ =4.772
		k =0.752	k =0.799	k =0.406	k =0.853			
10^{-6}	16	p =0.1833	p =0.1725	p =0.1771	p =0.0000	p =0.0017	p =0.0000	p =0.0000
		LL=-61.29	LL=-61.08	LL=-61.64	LL=-62.01	LL=-69.35	LL=-406.47	LL=-1812.43
		m =0.945	θ =1.606	λ =1.393	λ =1.298	λ =0.676	θ =0.723	μ =1.640
		k =0.803	k =0.857	k =0.526	k =0.931			
10^{-6}	32	p =0.2855	p =0.2797	p =0.2865	p =0.2824	p =0.3363	p =0.0000	p =0.0000
		LL=-28.61	LL=-28.62	LL=-28.60	LL=-28.60	LL=-28.68	LL=-76.84	LL=-140.91
		m =0.268	θ =0.663	λ =0.643	λ =0.636	λ =1.530	θ =0.375	μ =0.510
		k =0.938	k =0.962	k =0.657	k =0.987			
10^{-4}	8	p =0.8565	p =0.7789	p =0.9351	p =0.0001	p =0.0000	p =0.0000	p =0.0000
		LL=-107.15	LL=-107.70	LL=-107.82	LL=-110.22	LL=-206.53	LL=-945.83	LL=-10712.66
		m =2.097	θ =3.574	λ =2.566	λ =2.007	λ =0.344	θ =0.879	μ =3.311
		k =0.658	k =0.726	k =0.423	k =0.820			
10^{-4}	16	p =0.8779	p =0.8626	p =0.8852	p =0.0000	p =0.0000	p =0.0000	p =0.0000
		LL=-52.94	LL=-52.87	LL=-53.35	LL=-53.89	LL=-94.54	LL=-559.93	LL=-3052.30
		m =0.577	θ =1.263	λ =1.004	λ =0.905	λ =0.895	θ =0.622	μ =1.161
		k =0.679	k =0.766	k =0.541	k =0.899			
10^{-4}	32	p =0.9277	p =0.9230	p =0.9337	p =0.7775	p =0.0438	p =0.0001	p =0.0000
		LL=-24.22	LL=-24.21	LL=-24.26	LL=-24.28	LL=-29.72	LL=-42.59	LL=-156.60
		m =0.123	θ =0.601	λ =0.452	λ =0.418	λ =1.883	θ =0.275	μ =0.340
		k =0.511	k =0.633	k =0.633	k =0.903			
10^{-2}	8	p =0.7694	p =0.8084	p =0.3101	p =0.0000	p =0.0000	p =0.0000	p =0.0000
		LL=-75.84	LL=-74.87	LL=-82.50	LL=-88.38	LL=-342.91	LL=-586.65	LL=-8130.68
		m =0.760	θ =2.041	λ =1.232	λ =0.925	λ =0.647	θ =0.737	μ =1.724
		k =0.463	k =0.561	k =0.452	k =0.783			
10^{-2}	16	p =0.9015	p =0.9259	p =0.8277	p =0.0529	p =0.0000	p =0.0000	p =0.0000
		LL=-34.23	LL=-33.96	LL=-35.00	LL=-35.38	LL=-58.59	LL=-110.62	LL=-663.64
		m =0.231	θ =0.844	λ =0.606	λ =0.541	λ =1.387	θ =0.423	μ =0.605
		k =0.508	k =0.624	k =0.579	k =0.879			
10^{-2}	32	p =0.5800	p =0.5656	p =0.5885	p =0.5917	p =0.3084	p =0.0004	p =0.0000
		LL=-21.50	LL=-21.51	LL=-21.48	LL=-21.48	LL=-22.66	LL=-34.35	LL=-73.80
		m =0.071	θ =0.426	λ =0.364	λ =0.351	λ =2.498	θ =0.156	μ =0.174
		k =0.655	k =0.763	k =0.701	k =0.949			

Table 1:

Figure A1

Comparison of maximum likelihood fits of the seven candidate models (lines) with the original nine SAD data sets (points). Black: negative binomial; blue: Gamma; magenta: Weibull; brown: stretched exponential; green: exponential; red: log-series; orange: Poisson. Parameters are from table 1.



Appendix 2: Fits to species abundance distributions at intermediate areas

As discussed in the main text, we find that at intermediate to large areas that the SAD is characterized by the single parameter νAL^{-2}, and sufficiently large νAL^{-2}approaches the log-series $\phi_n \propto \frac{\alpha^n}{n}$. Both the negative binomial and the Gamma distribution reduce to the log-series when the shape parameter k is zero.

To investigate how the SADs approach the log-series, we used the procedure described in Appendix 1 to fit both Gamma and negative binomial distributions to our SAD data. Good fits (p>0.05) were found in 79% (Gamma) and 76% (negative binomial) of all cases, which shows that neither of these distributions is an exact representation of the true SAD because if they were we would expect 95% of all cases to have p>0.05.

We selected parameter combinations that satisfied the following criteria, in order to be in the scaling regime where the SAD is characterized by $A\nu L^{-2}$ only:

1. $L > 5$
2. $A|\log \nu| L^{-2} > 100$,
3. $\nu < 0.1$

In addition, we only used data sets where the p-value for the fit was greater than 0.05, to avoid cases where the trial distribution was not a good description of the data. We also did not use parameter values where no species had more than 20 individuals, to avoid cases where the fit only seemed good because there were too few points at which it was fitted.

Figure A2 shows the shape parameter k for a Gamma distribution fitted to the SADs. The results for fits to a negative binomial (not shown) display a similar pattern. The shape parameter for a particular kernel is mostly

described by $A\nu L^{-2}$, which is to be expected because we are in the parameter regime where the SAD itself is determined by $A\nu L^{-2}$ alone. The shape parameter decreases to zero as $A\nu L^{-2}$ increases, showing that the SAD approaching a log-series in this limit. For smaller values of $A\nu L^{-2}$ the shape parameter appears to approach 1, which would correspond to a purely exponential SAD, but our results do not extend far enough into this regime to explore further. The value of k also depends on the shape parameter η of the dispersal kernel.

We also fitted the analytically known SAD for Hubbell's spatially implicit local community Neutral model (**??**), known as the dispersal limited multinomial (DLM, which is here combined with the assumption that the metacommunity is infinite and has log-series abundances):

$$\phi_n = \theta \frac{J!\Gamma(\gamma+1)}{n!(J-n)!\Gamma(J+\gamma)} \int_0^1 \frac{\Gamma(n+\gamma z)\Gamma(J-n+\gamma(1-z))}{\Gamma(1+\gamma z)\Gamma(\gamma(1-z))}(1-z)^{\theta-1}dz, \tag{2}$$

where J is the local community size (here, the survey area A), θ is the fundamental biodiversity number of the metacommunity, and γ is related to the probability m that a recruit in the local community is an immigrant from the metacommunity by the following relationship

$$\gamma = \frac{m(J-1)}{1-m}.$$

Although the full sampling formula for this model (i.e. the joint probability that a sample contains a particular set of species abundances) is known (Etienne and Alonso 2005), we shall instead use same approximate Poisson likelihood (Equation 1) as was used for the other proposed SADs in order that the dispersal limited multinomial can be assessed on an equal footing. Performing a maximum likelihood fit using the DLM SAD is extremely computationally challenging, as the integral in equation (2) needs to by evaluated separately for each value of n and each candidate set of

parameters; for many of the parameter values encountered (J and/or θ large and/or $m \rightarrow 1$) the integrand becomes extremely sharply peaked, and the arguments of the Gamma functions become extremely large so that the final result became difficult to evaluate accurately (even when taking the differences of log Gamma functions rather than the ratio of Gamma functions). As a result, the code to perform this fitting is much slower than for the Gamma and Negative Binomial SADs, and we were unable to generate p-values in the same way as before (by re-fitting 1000 randomly generated samples from the fitted distribution). Instead, we used the log likelihood of the maximum likelihood fit to compare the goodness of fit.

Figure A3 compares the log likelihood of the maximum likelihood fit using the DLM with that of the Gamma and Negative Binomial distributions. The figure is color coded by the p-value of the fit by a Gamma distribution (panels A and C) and to a negative binomial distribution (panels B and D). It is clear that the DLM always fares worse than either of the other distributions in the case of a fat-tailed dispersal kernel (panels C and D). For a Gaussian dispersal kernel (panels A and B) It is clear that the DLM typically fares much worse than the Gamma distribution, except for some cases where the the Gamma distribution is itself a poor fit to the data (small p-value, points colored black). These cases lie in the region of larger values of $A\nu L^{-2}$. It should be noted that (i) in the majority of cases in this parameter region (98 out of 157) the Gamma distribution fits better than the DLM; and (ii) the fact that the DLM fares better than the Gamma distribution does not mean that the DLM itself is a good description. A similar pattern is seen when comparing the DLM with the negative binomial. A spreadsheet containing the fitted values and goodness of fit indicators for the other parameter combinations we simulated is included further as online material .

Figure A2

Shape parameter for a Gamma distribution fitted to the SAD data, in the regime where the SAD is a family of curves characterized by $A\nu L^{-2}$ and the kernel shape η only. Panel (A) shows data for Gaussian dispersal kernels (green circles); the black line is a loess smoothing of the data. In panel (B), the points correspond to: blue squares: $\eta = -4.0$; red circles: $\eta = -4.4$; green triangles: $\eta = -5$; magenta diamonds$\eta = -6$. The black curve is the same smoothed result for the Gaussian kernel as in panel (A).

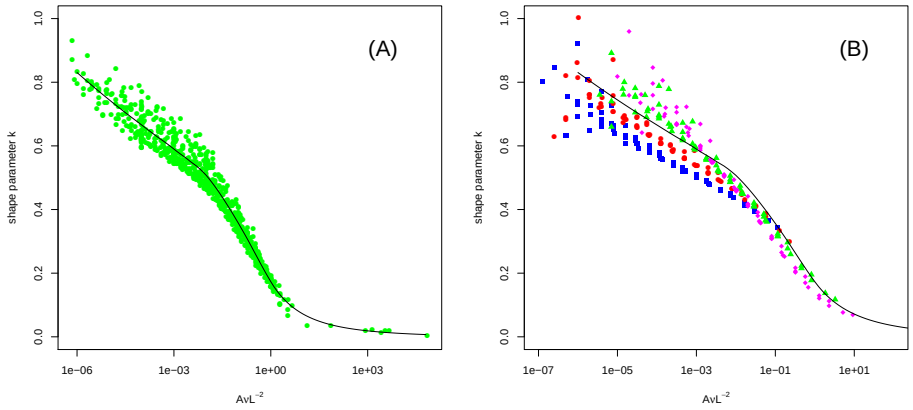
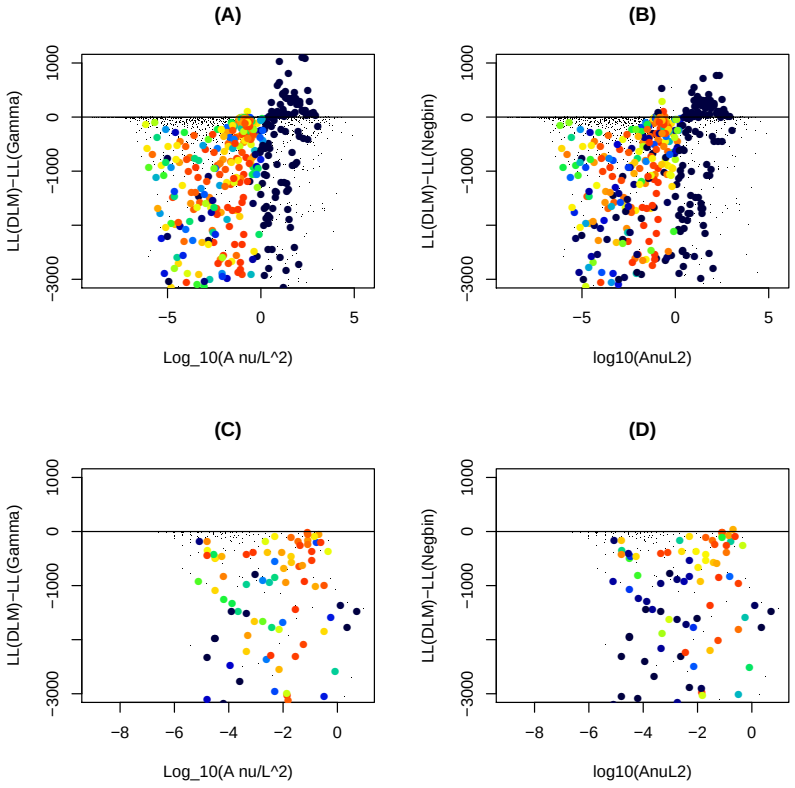


Figure A3

Comparison of the log likelihood for the fits with the DLM to the Gamma distribution (A and C) and the negative binomial distribution (B and D), for data from spatial neutral models with Gaussian dispersal kernels (A and B) and fat-tailed dispersal kernels (C and D), as a function of $A\nu L^{-2}$. Filled circles represent fits for parameter values deemed to be in the scaling regime ($L > 5$,$A|\log\nu|L^{-2} > 100$, and $\nu < 0.1$), and are colored according to the p-value for the fit by the Gamma distribution (A and C) or the negative binomial distribution (B and D), where cool colors represent low p-values (black: $p = 0$) and hot colors represent high p-values (red: $p = 1$). For reference, data for parameter values outside the scaling regime are included as grey dots.



Appendix 3: Supplementary figures

Figure A4: Fits for nearest neighbor dispersal at very large areas

SAD from spatially explicit neutral model with nearest neighbor dispersal, sample area $A = 2.56 \times 10^8$ and speciation rate $\nu = 0.003$. The black line represents the best fit to the log series formula $s_n = \frac{\theta}{n}(1-\nu)^n$, the best fitting parameters being $\theta = 1.156 \times 10^6$ and $\nu = 0.00322$. Note that the fitted value of ν is close to the true value (0.003), but the fit is visibly poor. Comparing this result to Figure 3 of the main text, where a log-series fits the spatially explicit simulation data extremely well, demonstrates the fundamental difference between the nearest neighbor dispersal model and the further neighbor dispersal model.

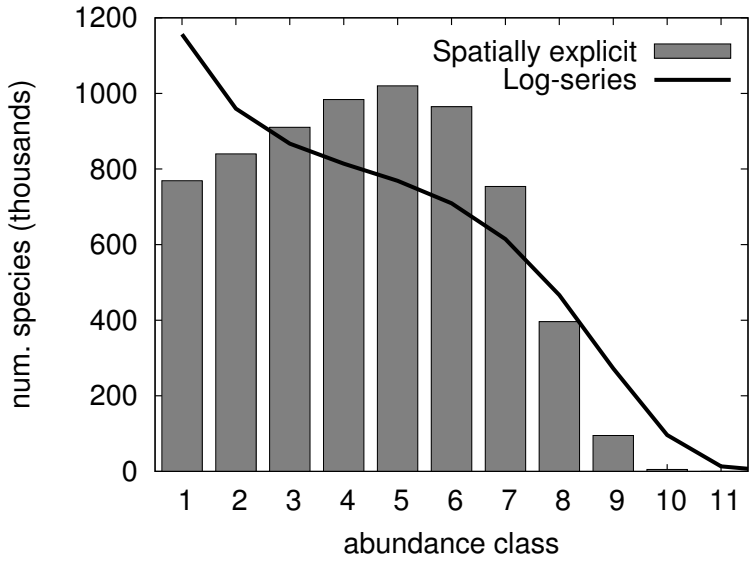
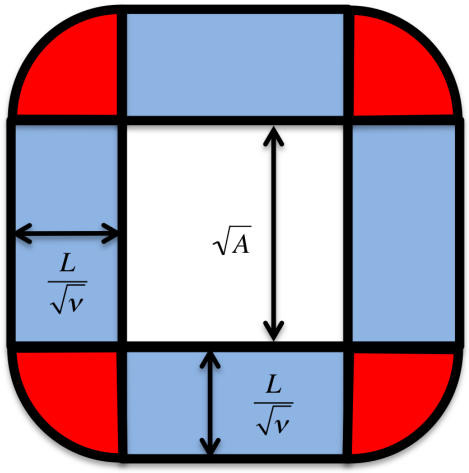


Figure A5: Calculating the effective metacommunity size for large scale log-series fitting

We wish to calculate the combined area of the sampling region A and all points around it that are within a distance of the correlation length $\frac{L}{\sqrt{\nu}}$. The diagram here shows that this area is made up of 9 component parts: the area A (white), four rectangles each of size $\sqrt{A} \times \frac{L}{\sqrt{\nu}}$ (blue) and four quarter circles of radius $\frac{L}{\sqrt{\nu}}$ (red). The four quarter circles together make a full circle of area $\pi\frac{L^2}{\nu}$ and so the total desired area J_{M} is given by $J_{\mathrm{M}} = A + 4\sqrt{A}\frac{L}{\sqrt{\nu}} + \pi\frac{L^2}{\nu}$. Consequently $J_{\mathrm{M}} = A + e$ where $e = 4\sqrt{A}\frac{L}{\sqrt{\nu}} + \pi\frac{L^2}{\nu}$ so that $e \propto \frac{L}{\sqrt{\nu}}$ as claimed in the main text.



References

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

Etienne, R. S. 2005. A new sampling theory for neutral biodiversity. Ecology Letters 8:253–260.

Etienne, R. S. 2007. A neutral sampling formula for multiple samples and an 'exact' test of neutrality.

Ecology Letters 10:608–618. Fisher, R. A., A. S. Corbet, and C. B. 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.

Houchmandzadeh, B., and M. Vallade. 2003. Clustering in neutral ecology. Physical Review E 68:061912.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.