Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales

(Supporting Information)

Sandro Azaele,[1] Amos Maritan,[2] Stephen J. Cornell,[3]
Samir Suweis,[2] Jayanth R. Banavar,[4] Doreen Gabriel,[5,6] William E. Kunin,[6]

[1]Department of Applied Mathematics, School of Mathematics,
University of Leeds, Leeds LS2 9JT, United Kingdom
[2]Dipartimento di Fisica ‘G. Galilei’ & CNISM, INFN,
Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[3]Institute of Integrative Biology,
University of Liverpool, Liverpool L69 7ZB, United Kingdom
[4]Department of Physics,
University of Maryland, College Park, MD 20742, USA
[5]Julius Kühn-Institut - Federal Research Centre for Cultivated Plants,
Bundesallee, 50, D-38116 Braunschweig, Germany
[6]School of Biology,
University of Leeds, Leeds LS2 9JT, United Kingdom

S1 The spatial variance

In this section we provide additional information about how to calculate the spatial variance from the pair correlation function (PCF).

Depending on the quality of the empirical data, the second-order summary statistics given by Eq.(4) of the main text might be quite variable. However, by using non-parametric methods or suitable envelope functions, one can smooth it out so as to obtain a non-parametric estimate of the PCF as a function of distance. Thus, we may explore the features of species' spatial distributions without resorting to any underlying model. In principle, we can directly use this non-parametric estimate of the PCF to calculate the quantities that we need in the mathematical framework, including the spatial variance. Specifically, because $\left\langle n_{\underline{x}} n_{\underline{y}} \right\rangle - \langle n \rangle^2 = \langle n \rangle^2 \left(g(|\underline{x} - \underline{y}|) - 1 \right)$ according to the definition in Eq. (1) in the main text, we can integrate this expression and obtain the variance of the population on circular areas of radius R through the following formula:

$$\sigma^2(R) = \langle n \rangle^2 \int_{|\underline{x}|<R} \mathrm{d}\underline{x} \int_{|\underline{y}|<R} \mathrm{d}\underline{y} \left[g(|\underline{x} - \underline{y}|) - 1 \right] \quad . \tag{1}$$

Therefore, with this expression one can calculate the spatial variance once the PCF is known even in non-parametric form. However, as the numerical evaluation of the integrals may be time-consuming for relatively large R and prone to rounding errors at some scales, we preferred to introduce a parametric estimate of the PCF. We found that the two-parameter function

$$g(r) = 1 + \frac{1}{2\pi} \left(\frac{\rho}{\lambda} \right)^2 K_0 \left(\frac{r}{\lambda} \right) \tag{2}$$

provides good fits to the empirical PCF as shown in Fig.S2 and is amenable to analytical calculations. $K_0(x)$ is a modified Bessel function of the second kind [1] with the following asymptotic behaviors

$$\frac{1}{2\pi} K_0 \left(\frac{r}{\lambda} \right) \sim \begin{cases} \dfrac{1}{2\pi} \ln \dfrac{\lambda}{r} & r \ll \lambda \\ \dfrac{1}{\sqrt{8\pi}} e^{-r/\lambda} & r \gg \lambda \end{cases} \tag{3}$$

Therefore ρ and λ play the role of two spatial scales: λ is a correlation length, above which the system becomes effectively uncorrelated, ρ gives an estimate of the degree of correlation (height of the PCF) at any given distance. This function has already been used to account for the spatial turnover of species [2, 3, 4]. Clearly, different empirical data sets might need other functions such as exponentials, logarithms, Gaussian functions or power laws, any of which our framework could easily accommodate.

In order to calculate $\sigma(R)$ analytically, we first deduce an equation satisfied by the PCF. If we use Eq. (1) of the main text as the definition of the PCF, we can write down the following differential equation

$$\lambda^2 \nabla_x^2 \langle n_x n_y \rangle + \langle n \rangle^2 - \langle n_x n_y \rangle + \rho^2 \langle n \rangle^2 \delta^2(x-y) = 0 \tag{4}$$

where $\delta^2(x-y)$ is Dirac's delta. This can also be verified by Fourier transforming Eq.(4). Let us now introduce the function $\gamma(x,R)$ defined as $\gamma(x,R) = \int_C \langle n_x n_y \rangle \, \mathrm{d}^2 y$ where C is a circular region of radius R. Note that $\int_C \gamma(x,R) \mathrm{d}^2 x = \langle N(R)^2 \rangle$ thereby allowing us to express the spatial variance as $\sigma^2(R) = \langle N(R)^2 \rangle - \langle N(R) \rangle^2$, being $\langle N(R) \rangle = \langle n \rangle \pi R^2$. All we need is to find out an equation for $\gamma(x,R)$ and solve it with appropriate conditions. If we integrate Eq.(4) with respect to y, we obtain the required equation:

$$\lambda^2 \nabla_x^2 \gamma(x,R) + \langle n \rangle^2 \pi R^2 - \gamma(x,R) + \rho^2 \langle n \rangle^2 \Theta(R - |x|) = 0 \tag{5}$$

where $\Theta(z) = 0$ for $z < 0$ and $\Theta(z) = 1$ for $z > 0$. This equation can be solved by exploiting the spherical symmetry of the problem. The solution is

$$\gamma(r = |x|, R) = \begin{cases} A I_0\left(\frac{r}{\lambda}\right) + \langle n \rangle^2 (\pi R^2 + \rho^2) & \text{for} \quad 0 \leq r \leq R \\ \\ B K_0\left(\frac{r}{\lambda}\right) + \langle n \rangle^2 \pi R^2 & \text{for} \quad r \geq R \end{cases} , \tag{6}$$

where $I_\nu(x)$ and $K_\nu(x)$ are modified Bessel functions of the first and second kind [1], respectively, and A, B are arbitrary constants to be determined by boundary conditions. The continuity of $\gamma(r,R)$ and $\partial_r \gamma(r,R)$ at $r = R$ provides the information needed to calculate the constants A, B. The final result can be expressed as

$$\sigma^2(R) = \langle n \rangle \rho^2 \langle N(R) \rangle \left(1 - \frac{2\lambda}{R} \frac{K_1\left(\frac{R}{\lambda}\right) I_1\left(\frac{R}{\lambda}\right)}{I_0\left(\frac{R}{\lambda}\right) K_1\left(\frac{R}{\lambda}\right) + I_1\left(\frac{R}{\lambda}\right) K_0\left(\frac{R}{\lambda}\right)} \right) . \tag{7}$$

This is the variance of the population at different scales implied by the PCF defined in Eq.(2). The formula has interesting regimes as the radius of a sampled area becomes small or large compared to the characteristic spatial length λ. If we set $\beta(R/\lambda) \equiv \sigma^2(R)/\langle N(R)\rangle$ and $x \equiv R/\lambda$, then when $x \gg 1$ we obtain $\beta(x) \simeq \langle n\rangle \rho^2(1-\frac{1}{x})$ whereas if $x \ll 1$ the expansion provides $\beta(x) \simeq \langle n\rangle \rho^2[(1-4\gamma+4\log(2)-4\log(x))\frac{x^2}{8}+O(x^4)]$, where γ is the Euler constant. Interestingly, the model exhibits two interesting regimes: $\sigma^2(A) \propto A$ for very large areas and $\sigma^2(A) \propto A^2$ (up to a logarithmic correction) for very small ones.

S2 The spatial species abundance distribution

Let us consider a general two-parameter function for the SAD, which we will express as $q(n|\alpha,\beta)$, where α and β are the two free parameters. The gamma and log-normal distributions are examples of such two-parameter distributions, widely used in the literature and mathematically flexible. In the present study, however, we will be using the gamma distribution for all the calculations.

We can make the SAD spatially dependent by substituting α and β with two appropriate functions, $\alpha(R)$ and $\beta(R)$, that depend on the radius, R, of the circular area we focus on, so that $q(n|\alpha(R),\beta(R))$ is our sSAD. The next problem is how to derive the functions $\alpha(R)$ and $\beta(R)$. Note that the first two moments of the sSAD have clear interpretations: the first is simply the mean density (number per unit area) of individuals per species, the second is related to the spatial variance that depends on the PCF as we have shown before. These two properties allow us to set the following pair of functional equations

$$\begin{aligned}\langle N(R)\rangle &= \int_0^\infty nq(n|\alpha(R),\beta(R))\mathrm{d}n \\ \sigma^2(R) &= \int_0^\infty (n-\langle n(R)\rangle)^2 q(n|\alpha(R),\beta(R))\mathrm{d}n \end{aligned} \quad (8)$$

that we have to solve to obtain $\alpha(R)$ and $\beta(R)$. Assuming that the average number of individuals per species is proportional to the sampled area, in our case the circular area πR^2, we have set $\langle N(R)\rangle = (N_0/S_0A_0)\pi R^2 = \langle n\rangle \pi R^2$. In Fig.(S4) we show an example of the agreement between empirical data and the parabolic curve without any fitting procedure. Instead, $\sigma^2(R)$ is given by Eq.(7) for the specific choice of the PCF defined in Eq.(2). The agreement between the empirical spatial variance and the theoretical one without best fitted parameters but using those from the PCF is shown in Fig.(S3). If we use a generic sSAD in Eqs.(8), the explicit expressions of $\alpha(R)$ and $\beta(R)$ may be quite cumbersome and in general an analytical expression may not be found so that one has to calculate them numerically. However, if we use a gamma distribution the expressions for $\alpha(R)$ and $\beta(R)$ can be achieved analytically, producing the following expression for the sSAD:

$$q(n|R) = \frac{1}{\beta(R)} \frac{(n/\beta(R))^{\alpha(R)-1}}{\Gamma(\alpha(R))} e^{-n/\beta(R)} \quad , \tag{9}$$

where $\Gamma(x)$ is a gamma function [1], $\alpha(R) = (\langle N(R) \rangle / \sigma(R))^2$ and $\beta(R) = \sigma(R)^2 / \langle N(R) \rangle$ are the functions defined explicitly in the main text. Note that $\alpha(R)$ is dimensionless and depends on λ and ρ only (apart from R), whereas $\beta(R)$ depends on λ, ρ and the mean density of individuals per species, i.e. $\langle n \rangle$.

S3 Extrapolation of species richness, SAR and SAD: detailed description

In this section we provide a detailed explanation of the methodology that was used to upscale the SAR and sSAD.

Eqs. (9) and (11) in the main body of the article can be used to estimate the sSAD and SAR at scale R if we know the total number of species $S_0(R_0)$ in the region of area $A_0 = \pi R_0^2$. If we only have fine-scale samples scattered across the region, we need a way to extrapolate to $S_0(R_0)$. To do so, we introduce the ‘disconnected’ SAR (dSAR) which gives the mean number of species found within a *disconnected* area A_{samp}. This latter is the total area corresponding to all the aggregated samples. Therefore, if we are given k samples distributed within the area A_0 and each with area a ($\ll A_0$), then $A_{samp} = ka$. In our case we used $a = 10 \times 10 m^2$ and took different values for k. The dSAR is in general different from the common SAR, because in this latter case the sampled area is meant to be ‘connected’ and not split into smaller samples. However, from the samples we can approximately calculate the PCF and then upscale the information collected at the sample level to the entire region with area A_0, where the dSAR and SAR must clearly coincide. From this large scale we can then calculate the common SAR at finer scales. Notice that, because samples are correlated in space owing to a non-trivial β-diversity, the spatial sampling is in general different from a random sampling.

Under the assumptions of the framework, the dSAR is always equal or higher than the SAR, because of the spatial turnover of species. In addition, all samples are spatially correlated and this correlation is still given by the PCF used for the SAR. Hence, one may think that the curve of the SAR can be used to obtain a first approximation of the dSAR, albeit with different areas, number of species and $\langle n \rangle$. Therefore, when sampling the (disconnected) area A_{samp}, we have used the following formula

$$S(R) = S_{emp}(R_{samp}) \frac{\int_1^\infty q(n|\alpha(R), \beta(R)) \mathrm{d}n}{\int_1^\infty q(m|\alpha(R_{samp}), \beta(R_{samp})) \mathrm{d}m} \tag{10}$$

to describe the dSAR at larger (disconnected) areas than A_{samp}, where $R_{samp} = \sqrt{A_{samp}/\pi}$ and $S_{emp}(R_{samp})$ is the total number of species found among the (disconnected) samples. The agreement of Eq. (10) with empirical data is shown in Fig.(S7). The quantities $\alpha(R)$ and $\beta(R)$ are again given by Eqs. (7) and (8) in the main body of the article. For this to be calculated, we need λ and ρ — which are estimated through a least-squares best fit of the empirical PCF as obtained from the samples — and the mean number of individuals per species $\langle n \rangle$, which we estimate as $\langle n_{emp}(R_{samp}) \rangle = N_{samp}/(A_{samp} S_{emp}(R_{samp}))$, where N_{samp} is the total number of individuals found in the area A_{samp}. The upscaled dSAR from the scattered samples is shown in Fig.(S7). Therefore the upscaled richness of the region is simply $S_{up}(R_0) := S(R_0)$. In Fig.(S8) we show how the predicted total number of species changes by varying the amount of sampled area. The predictions which we have shown in Fig.(4) of the main text refer to the case in which one samples 1.2% of the total area.

In order to calculate the SAR and SAD at scales $R < R_0$ we need to know the parameters at the scale R_0: λ, ρ are already known from the samples and they do not depend on the spatial scale in a first approximation. For the mean population density per species we have $\langle n_{up}(R_0) \rangle = S_{emp}(R_{samp}) \langle n_{emp}(R_{samp}) \rangle / S_{up}(R_0)$, where we have assumed that the density of individuals is scale independent. With these parameters we can calculate the SAR and the SAD at smaller scales with the following formulas

$$S(R) = S_{up}(R_0) \frac{\int_1^\infty q(n|\alpha(R), \beta(R)) \mathrm{d}n}{\int_1^\infty q(m|\alpha(R_0), \beta(R_0)) \mathrm{d}m} \tag{11}$$

and

$$sSAD(n|R, R_0) = S_{up}(R_0) \frac{q(n|\alpha(R), \beta(R))}{\int_1^\infty q(m|\alpha(R_0), \beta(R_0))\mathrm{d}m} \quad , \tag{12}$$

respectively. With such equations one can infer the mean number of species as well as the mean number of rare and common species in regions that have not been sampled yet.

References and Notes

[1] N.N. Lebedev and R.A. Silverman. *Special functions and their applications*. Dover Pubns, 1972.

[2] R. Condit, N. Pitman, E.G. Leigh, J. Chave, J. Terborgh, R.B. Foster, P. Núnez, S. Aguilar, R. Valencia, G. Villa, et al. Beta-diversity in tropical forest trees. *Science*, 295(5555):666, 2002.

[3] J. Chave and E.G. Leigh. A spatially explicit neutral model of beta-diversity in tropical forests. *Theoretical Population Biology*, 62(2):153–168, 2002.

[4] T. Zillio, I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan. Spatial scaling in model plant communities. *Phys. Rev. Lett.*, 95:098101, 2005.

[5] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan. Neutral theory and relative species abundance in ecology. *Nature*, 424(6952):1035–1037, 2003.

Figure S1: Pair correlation function calculated in two different ways. In the left panel, we selected two non-overlapping regions of the same area at distance r and then multiplied the number of individuals belonging to the same species present in the corresponding regions and, finally, averaged across species. Green, brown, purple and blue dots represent the empirical results corresponding to a pair of areas of $50^2, 31.25^2, 25^2$ and 20^2 m^2, respectively. Data taken from BCI. The right panel shows that the four curves collapse into one curve so long as we divide the previous results by the mean number of individuals found in the corresponding areas.

Figure S2: Pair correlation functions for the BCI and Pasoh forests (all possible pairs and average of pairs if they have the same distance). Black dots represent empirical data calculated according to the formula defined in Eq.(4) of the main text, whereas the solid curve is the best fit of the function in Eq.(2) to the data. The study plots in each case are 500×1000 m. We have restricted our analyses to a maximum distance of 500 m. This is because directional contrasts are constrained by plot shape at larger distances.

Figure S3: Spatial variance, $\sigma^2(r)$, as a function of the radius of the sampled area for the BCI and Pasoh forests. Black dots represent empirical data, whereas the solid curve is predicted by Eq.(7) as we use the best fitted parameters obtained from the PCF in Eq.(2) (see also Fig.(S2)). The scales are logarithmic in both axes.

Figure S4: Mean number of individuals per species (including all species) as a function of the radius of the sampled area. Black dots are empirical data from the BCI forest. The solid curve is given by $n(r) = (N_0/S_0A_0)\pi r^2$ where $N_0(= 21,204)$ is the total number of individuals in the study region, $S_0(= 228)$ the total number of species and $A_0(= 500,000\ m^2)$ the total area.

Figure S5: The thick solid curve represents the predicted Species Abundance Distribution (SAD) as predicted by the pair correlation function (PCF) and the mean population density per species. The parameters are those used in the predictions of Fig.(2) in the main text. Gray histograms represent empirical data from the Barro Colorado Island at the scale of the whole study region. Preston classes were calculated as in [5]. Note that this curve is not fit to the data, but rather was predicted by the PCF and the density, as outlined in the text.

Figure S6: Ecological patterns predicted by the theory at the scale of the study region. The thick solid curves represent the predicted Species Area Relationship (SAR, left panel) and Species Abundance Distribution (SAD, right panel) as given by the pair correlation function (PCF) and the mean population density per species. Black dots and gray histograms represent empirical data from the Pasoh forests at the scale of the whole study region. Preston classes were calculated as in [5]. Note that these curves are not fit to the data, but rather are predicted by the PCF and the density, as outlined in the text. The thin solid lines represent the SAR that would be predicted if individuals were randomly placed, i.e. spatial aggregation is negligible.

Figure S7: This figure shows the SAR (black dots) as a function of the radius R of the sampled area A, where the area is meant to be connected. The red dots, instead, show the mean number of species that one finds when using a *disconnected* area A (dSAR). In our case this means that we split the area $A_{samp} = \pi R_{samp}^2$ into samples of $10 \times 10m^2$ and then scattered them across the area A_0 of the Pasoh forest (spatial sampling). Therefore, the Radius in the x-axis represents the radius of a connected area for the SAR and the radius of a disconnected area for the dSAR in which we varied the number of samples (but not the size of one sample). The blue line is the upscaled dSAR given by Eq.(10) where $R_{samp} = \sqrt{A_{samp}/\pi}$ with $A_{samp} = ka$, $k = 60$ and $a = 10 \times 10m^2$ (1.2% of the total area). The gray band represents the 95% confidence intervals. The corresponding SAR is shown in Fig.S6.

Figure S8: Predicted total species richness as a function of the percentage of sampled area. Black dots represent the median whereas the error bars are the 90% confidence intervals. The number on the $x-$axis corresponds to the percentage of the total area sampled by using different numbers of $10m \times 10m$ non-overlapping samples. The two horizontal lines represent the true number of observed species in the corresponding 50 Ha plot forests.

Figure S9: Bilogarithmic plots of different SAR curves according to Eqs.10 and 11 of the main text. The parameter that sets the mean population density per species is $\langle n \rangle = 1000$ ind/(km^2·species) and ρ is 5,000 km for each of the three panels. The spatial length λ is 1 km (upper panel); 200 km (middle right panel); 700 km and (lower panel). Note that the resultant SAR is quite flexible in form, and the total number of species decreases when λ decreases.