# A geometric approach to scaling individual distributions to macroecological patterns

Nao Takashina[1]*, Buntarou Kusumoto[2], Yasuhiro Kubota[3], Evan P. Economo[1]

[1] *Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Okinawa 904-0495, Japan*
[2] *Center for Strategic Research Project, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan*
[3] *Faculty of Science, University of the Ryukyus, Nishihara, Okinawa, 903-0213, Japan*

*Keywords:* Endemic area relationship, relative species abundance, spatially explicit model, species area relationship

**Abstract**

Understanding macroecological patterns across scales is a central goal of ecology and a key need for conservation biology. Much research has focused on quantifying and understanding macroecological patterns such as the species-area relationship (SAR), the endemic-area relationship (EAR) and relative species abundance curve (RSA). Understanding how these aggregate patterns emerge from underlying spatial pattern at individual level, and how they relate to each other, has both basic and applied relevance. To address this challenge, we develop a novel spatially explicit geometric framework to understand multiple macroecological patterns, including the SAR, EAR, RSA, and their relationships. First, we provide a general theory that can be used to derive the asymptotic slopes of the SAR and EAR, and demonstrates the dependency of RSAs on the shape of the sampling region. Second, assuming specific shapes of the sampling region, species geographic ranges, and individual distribution patterns therein based on theory of stochastic point processes, we demonstrate various well-documented macroecological patterns can be recovered, including the tri-phasic SAR and various RSAs (e.g., Fisher's logseries and the Poisson lognormal distribution). We also demonstrate that a single equation unifies RSAs across scales, and provide a new prediction of the EAR. Finally, to demonstrate the applicability of the proposed model to ecological questions, we provide how beta diversity changes with spatial extent and its grain over multiple scales. Emergent macroecological patterns are often attributed to ecological and evolutionary mechanisms, but our geometric approach still can recover many previously observed patterns based on simple assumptions about species geographic ranges and the spatial distribution of individuals, emphasizing the importance of geometric considerations in macroecological studies.

*Electronic address: nao.takashina@gmail.com; Corresponding author

# 1 Introduction

The problem of pattern and scale is central in ecology, and critical insights into conservation biology emerge as we observe different aspects of ecosystems across scales [1]. The species area relationship (SAR), endemic area relationship (EAR), and relative species abundance (RSA) are such examples characterizing macroecological and community patterns of ecosystem across scales. Disentangling each of these patterns has a long history and ample literature (e.g., [2–6]), and there are still active discussions over fundamental patterns (e.g., [7, 8]). In particular, recent research has focused on the quantitative investigation of scaling issues of macroecological patterns [9–12]. However, little is known about how individual-level spatial distributions scale up to aggregate patterns such as the SAR, EAR, and RSA. Also, most existing models provide predictions of only one or two of these patterns (but see [11]), and therefore understanding these emergent patterns within the framework of a single model as well as from the individual level provides us a holistic insight into ecosystem structure.

A model linking individual distributions with aggregate macroecological patterns would be useful for other ecological questions in community ecology and conservation biology. For example, to investigate the scale-dependence of beta diversity across multiple scales and how it relates to the SAR [13], one needs species abundance information at finer scales and, in a theoretical framework, a consistent scale-change operation must be defined. These macroecological patterns are also critical and increasingly necessary for implementing effective ecosystem management, given accelerating loss of biodiversity worldwide [14, 15]. For example, the SAR and EAR provide information about how many species will be affected and be lost from the landscape when a certain region is degraded, respectively [16]. The RSA provides more detailed information on community structure such as composition of rare species, and it holds broad utility [17]. Practically, biodiversity conservation has finite resources available, and thus spatial prioritization of areas is a key step to maximize return on investment [18–20]. A high degree of spatial information is necessary for optimal allocation, such as individual distributions and the number of individuals across multiple scales (e.g., [21, 22]). These issues are tackled more easily by individual based models defined in continuous space that meet the above-mentioned requirements. Above all, an individual based model that describes multiple macroecological features across scales would advance the theoretical bases of these fields. For example, Takashina et al. developed a spatially explicit framework for ecosystem assessment [21] and population estimates combined with a specific sampling strategy [22], but their model was limited to a single population because a model that can be applied to community-level structure was not available.

Here, we propose a novel geometric approach that provides a bridge between configurations of geographic ranges of species and individual distribution patterns on one hand and emergent macroecological patterns on the other hand. We refer to our approach as "geometric" because we put a particular emphasis on how geometric characteristics of the sampling region, geographic ranges, and the distribution of individuals within the range put constraints on common aggregate macroecolgical patterns across scales. It enables us to explicitly discuss the effect of particular sampling scheme and its scale dependence, and the SAR, EAR, and RSA across scales are derived from these geometric considerations. Some

of the notable properties of our model are (i) this approach can be applied to any shapes of sampling and species geographic ranges and any individual distribution patterns, (ii) one can easily incorporate variations of biological parameters (e.g., density, dispersal length) between species, and (iii) no preceding knowledge of macroecological pattern is required to calculate above-mentioned patterns and, therefore, we can discuss in what conditions the model recovers emergent macroecological patterns. These properties also provides a significant applicability to conservation practices, since the model provides macroecological information on arbitrary spatial scales.

Conceptually similar approaches have been investigated in previous studies without individual distributions [9,10,23], and with explicitly integrating individual distributions [11,24]. Allen and White [23] derived an upper bound of the SAR, by calculating an overlapping area between the sampling region and the randomly placed geographic range without considering individual distributions therein, leading to a biphasic SAR on a log-log plot. Although this model provides the well-known asymptotic slope of 1 at very large sampling scales, it did not capture the sampling process on smaller scales responsible for the triphasic curve. Plotkin et al. [24] applied the Thomas process, a point process model, to generate aggregated individual distribution patterns in tropical forest plots (50-ha). They independently estimated a dispersal distance and density from spatial distribution of individual trees of each species, and generated superposition of multiple species in a forest. The fitted Thomas process model very precisely recovers SARs in the plot. Grilli et al. [11] also applied a point process model to account for individual distributions. With aggregated communities generated by the Poisson cluster model, they derived the SAR, EAR, and RSA. In doing so, the model requires a neutral assumption and a specific input of the RSA form in the whole system.

The paper contains three major parts, and some of key findings are briefly summarized here. First, we develop a general theory based on our geometric approach to link the SAR, EAR, and RSA across scales to spatial point patterns, without assuming specific spatial structures. This part provides core ideas of the proposed model. As this is a rather general framework, any spatial configurations defined by geographic ranges and individual distribution patterns can be applied. Even with this general discussion, many important implications can be derived; the asymptotic slope of SAR and EAR that are the same value 1 on a log-log plot, and potential unification of RSAs and its artifact effect of the shape of sampling regions. However, readers interested in more concrete results may skip section 2.2 – 2.4 and refer to these sections when it is required. Second, we discuss emergent macroecological patterns of the geometric approach by assuming specific patterns of species distribution and individual distributions therein. We use spatial point processes to generate individual distribution patterns and introduce stochasticity in realized geometric patterns. We show that the model produces a well-documented macroecological patterns such as the tri-phasic SAR with its asymptotic slope 1 on a log-log plot, with limited effect of underlying individual distributions (random or clustering). We discuss how Fisher's logseries, the negative binomial distribution, and the Poisson lognormal distributions are obtained as an RSA at small sampling scales. Then we demonstrate how different forms of the RSA across spatial scales emerge from a single equation (Eq. 43). We also demonstrate, in addition to sampling

schemes, size distribution of geographic ranges affects the RSA. In addition, our model also provides new potential form of EARs with asymptotic slope 1 on a log-log plot. Third, we demonstrate potential applicability of the geometric model to ecological questions. As an example, we provide the scale-dependence of beta diversity over tri-phasic scales of SAR, in which a theoretical approach was challenging since it demands a model describing detailed ecological structures (e.g., the number of individuals) across the scales.

It is worth noting that the emergent macroecological patterns are often attributed to ecological and evolutionary mechanisms, but our geometric approach still can recover many previously observed patterns based on simple assumptions about geographic ranges and individual distributions. Moreover, from our geometric considerations, some important implications are obtained such as necessity to specify sampling area and its shape to discuss a general ecological pattern and potential simplification of the assumption of individual distribution patterns.

# 2 General theory

We develop here a general theory of our geometric approach, an attempt to provide a general framework for calculations of the SAR, EAR, and RSA without assuming any specific geometric patterns. More specifically, here we do not assume individual distribution patterns, shapes of the sampling region and the geographic range. However, it is worth noting that the central assumptions of the theory are that we assume no interaction between species and, therefore, geographic ranges of each species are chosen independently in a homogeneous environment. Once ecological properties and the sampling scheme are specified, the SAR, EAR, and RSA can be obtained via the general framework. We will discuss some specific situations in Results section. Central parameters used throughout the paper are summarized in Table 1.

## 2.1 Basic notation

### 2.1.1 Intensity and intensity measure

We introduce some basic quantities of ecosystem which will be used in the following discussions. Each of these are generally used in the field of point process, and we borrow some ideas in this paper. Interested readers may refer to literature in the field (e.g., [25, 26]) or applications to ecological studies (e.g., [11,21,22,24,27]). Later, we will combine the general theory with spatial point processes to examine specific individual distribution patterns such as random or clustering distributions.

The intensity, $\lambda_i$, and the intensity measure, $\mu_i$, of species $i$ are the central quantities of our model characterizing the ecological community: the average density of individuals of species $i$, and the average number of individuals of species $i$ in region $A$, respectively. Given the area $\nu(A)$ and the number of individuals of species $i$, $N_i(A)$ in $A$, the intensity (density)

Table 1: Definition of parameters

| Symbol | Parameter |
|---|---|
| $S$ | Sampling region |
| $G$, $(G_i)$ | Geographic range of species $(i)$ |
| $\nu(A)^\dagger$ | Area of $A$ (km$^2$) |
| $\mathbf{x}_c^A$ | Geometric center of $A$ |
| $l_A$ | Shortest distance in $A$ from the geometric center (km) |
| $L_A$ | Largest distance in $A$ from the geometric center (km) |
| $r_a{}^\ddagger$ | Radius of $A$ (km) |
| $X$ | Number of individuals |
| $N_i(A)$ | Number of individuals of species $i$ in $A$ |
| $N_s(A)$ | Number of species in $A$ |
| $\lambda_i$ | Intensity of species $i$ |
| $\lambda_s$ | Intensity of species |
| $\mu_i(A)$ | Intensity measure of species $i$ in $A$ |
| $\mu_s(A)$ | Intensity measure of species in $A$ |

$^\dagger$ Replace $A$ with $S$ or $G$ depending on context
$^\ddagger$ When $A$ is a circle

of species $i$ in $A$ is defined as [26]

$$\lambda_i \coloneqq \frac{N_i(A)}{\nu(A)}. \tag{1}$$

The intensity measure of species in $A$ is given by multiplying the intensity by area $\nu(A)$

$$\mu_i(A) \coloneqq \lambda_i \nu(A). \tag{2}$$

The same quantities for species assembly are also defined within this framework. Let $N_s(A)$ be the number of species in $A$. Then, the intensity and intensity measure in $A$ of the species assembly are defined as $\lambda_s \coloneqq N_s(A)/\nu(A)$ and $\mu_s \coloneqq \lambda_s \nu(A)$, respectively. Then the vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_s)$ holds the information of community abundance. Let us define the intensity measure of species, $\mu_s(A)$ as

$$\mu_s(A) \coloneqq \int_A f(\mathbf{x}) d\mathbf{x}, \tag{3}$$

where, $f(\mathbf{x})$ is a point field holding information on all individual locations in plane $\mathbf{x} \in \mathbf{R}^2$, and $f(\mathbf{x})d\mathbf{x}$ provides the number of points in region $d\mathbf{x}$. Let us assume the homogeneous environment, and then we set

$$f(\mathbf{x}) = \lambda_s, \tag{4}$$

where, $\lambda_s$ is the intensity of species, and we obtain the following relationship

$$\mu_s = \lambda_s \nu(A). \tag{5}$$

### 2.1.2 Sampling region and endemic region

In this paper, we refer to $S$ as the *sampling region* which defines the spatial scales discussed. Similarly, we refer to $G_i$ as the *geographic range of species* $i$. We often use $G$ when $G_i$ is identical in size and shape to $G_j$ for all $j \neq i$ or species identity is not important, and we call $G$ the *geographic range of species* or simply *geographic range* when no confusion occurs. Throughout the analysis, the geographic range $G$ is used as an endemic region of species where the individual distributions of the species are restricted within the region. The regions $S$ and $G$ hold multiple information in configuration space such as location, area, and shape. For convenience, regarding to relative sizes of geometries $S$ and $G$, we use the symbols $>$ and $<$. For example, $S > G$ means that there exists a shift that the entire geographic range $G$ is included in the sampling region ($G \subset S$) without rotating these geometries. When we mention the areas, we represent these by $\nu(S)$ and $\nu(G)$, respectively. As noted above, the location of each geographic range $G_i$ has a certain stochasticity, and therefore its location and properties of set theory such as the overlap area, $\nu(S \cap G_i)$, are random variable. An example of a configuration of the sampling region and geographic ranges is shown in Fig. 1a.

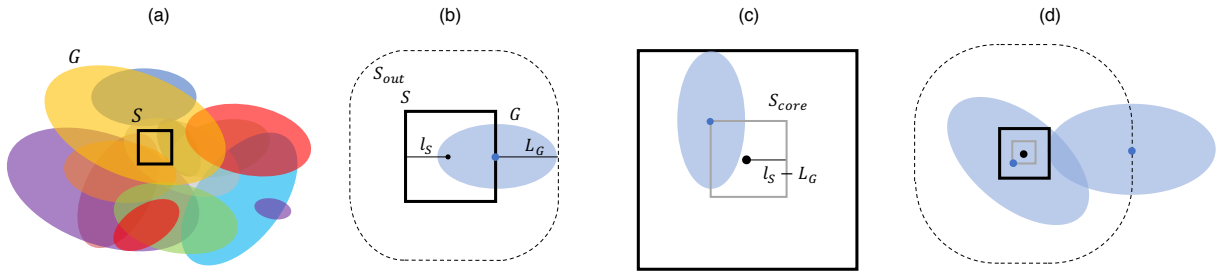


Figure 1: Schematic representations of (a) the sampling region and geographic ranges of species, and typical geometric relationships used in derivations of the (b) SAR, (c) EAR, and (d) RSA. For each panel, the sampling region $S$ (black square) and the geographic ranges $G$ (ellipse filled by color and geographic range of each species is represented by different color) are presented with its geometric center (black and blue dots). $S_{out}$ is the region described with dotted line (b and d), where if the geometric center of a range $\mathbf{x}_c^G$ falls in this region, then the probability of overlapping $S$ and $G$ is not zero: there is a chance individuals of this species is found in $S$. $S_{core}$ is the region described by gray line (c and d), where provided $S > G$ ($S < G$) if the geometric center of the geographic range $\mathbf{x}_c^G$ falls in this region, then the entire geographic range $G$ ($S$) is included in $S$ ($G$). For the derivation of the EAR, only the case $S > G$ is used, but both $S > G$ and $S < G$ are used for the derivation of the RSA. $l_A$ and $L_A$ are the shortest and largest distance of $A$ from its geometric center. See further explanations in main text and Appendix A.

## 2.2 Species area relationship

Here we derive a method to calculate the species area relationship (SAR). Often it exhibits a tri-phasic form on a log-log plot [3,5] characterized a fast increase on small scales, a slowdown on intermediate scales, and again an increase asymptotically toward linearity for the slope on very large scales where sampling region exceeds correlation distance of biogeographic process with separate evolutionary histories [5, 6]. As the simplest example, let us first assume a homogeneous where biological properties, such as intensity (density) of each species, geographic range size, and average dispersal distance are identical between species. However, as we will see below, it is straightforward to incorporate species variations. To calculate the species number in a given area, we define the nonzero probability $p^{nz}(\mathbf{x}_c^S, \mathbf{x}_c^G)$ inside $S_{out}$: the probability that a species with the geometric center of its geographic range $\mathbf{x}_c^G \in S_{out}$ provides at least one individual to the sampling region $S$ with its geometric center $\mathbf{x}_c^S \in S$. $S_{out}$ is the region that if the geometric center of the range of a species belongs to this region, $\mathbf{x}_c^G \in S_{out}$, then the probability of overlapping sampling region and geographic range is not zero, $P(\nu(S \cap G) = 0) \neq 0$. Therefore, the species provides its individuals to the sampling region $S$ at a non-zero probability (Fig. 1b). See Appendix A for further explanations for $S_{out}$.

In this simplest example, the average number of species given sampling area $\nu(S)$ is described:

$$\begin{aligned} N_s(\nu(S)) &= \int_{S_{out}} f(\mathbf{x}) p^{nz}(\mathbf{x}_c^S, \mathbf{x}) d\mathbf{x}, \\ &= \lambda_s \nu^{nz}(S_{out}), \end{aligned} \tag{6}$$

where $\nu^{nz}(S_{out})$ is the area of $S_{out}$ weighted by the non-zero probability $p^{nz}(\mathbf{x}_c^S, \mathbf{x}_c^G)$ over the region $S_{out}$. The SAR is obtained by calculating Eq. (6) across sampling areas $\nu(S)$. From Eq. (6), we can show that the slope of the SAR on a log-log plot approaches 1 at large scales. First, the slope is calculated by

$$\frac{d \log N_s(\nu(S))}{d \log \nu(S)} = \frac{\nu(S)}{N_s(\nu(S))} \frac{dN_s(\nu(S))}{d\nu(S)}. \tag{7}$$

Second, when the sampling region becomes significantly larger than the geographic range $S \gg G$, $S_{out}$ is nearly equivalent to $S$. In this limit, $p^{nz}(\mathbf{x}_c^S, \mathbf{x}_c^G) \simeq 1$ is satisfied, providing

$$\begin{aligned} N_s(\nu(S)) &\simeq \int_S f(\mathbf{x}) d\mathbf{x}, \\ &= \lambda_s \nu(S). \end{aligned} \tag{8}$$

The slope 1 is obtained by substituting this into Eq. (7). Due to the prerequisite relationship $S \gg G$, this asymptotic slope is achieved faster if the range sizes are smaller.

As noted above, it is straightforward to incorporate species variations by assuming that some biological properties follow certain probability distribution functions (pdfs). It is sufficient to consider biological parameters that affect the non-zero probability $p^{nz}(\mathbf{x}_c^S, \mathbf{x}_c^G)$

(e.g., range size and intensity of each species). We define such parameters by a vector $\mathbf{b} \coloneqq (b_1, b_2, \cdots, b_k) \in \mathbf{B}$, where $\mathbf{B}$ is a state space created by biologically plausible parameter sets. When the range size $G$ varies between species, and hence it generates a probability distribution of $G \in \mathcal{G}$, $p(G)$, where $\mathcal{G}$ is the set of geographically plausible range sizes $G$. For each $G$, the $S_{out}$ is unequally determined. When it is necessary to specify $G$ of each $S_{out}$, we write $S_{out}^G$. Here, let us assume the independence of biological parameters $\mathbf{b}$ from the geographic ranges $G$. Therefore, the pdfs of biological parameters and the range are described by $p(\mathbf{b}, G) = p(\mathbf{b})p(G)$. Now, Eq. (6) becomes

$$\begin{aligned} N_s(\nu(S)) &= \int_{\mathbf{b}\in\mathbf{B}} p(\mathbf{b}) \int_{G'\in\mathcal{G}} p(G') \int_{S_{out}^{G'}} f(\mathbf{x}) p^{nz}(\mathbf{x}_c^S, \mathbf{x}) d\mathbf{x} dG' d\mathbf{b}, \\ &= \lambda_s \mathrm{E}_{\mathbf{b}} \mathrm{E}_{S_{out}}[\nu^{nz}(S_{out})], \end{aligned} \tag{9}$$

where, $\mathrm{E}_{S_{out}}$ and $\mathrm{E}_{\mathbf{b}}$ are the average over the geographic ranges and biological parameters, respectively. As $\nu^{nz}(S_{out})$ is the average area to find at least one individual, $\mathrm{E}_{\mathbf{b}}\mathrm{E}_{S_{out}}[\nu^{nz}(S_{out})]$ is the such area averaged over the pdfs $p(\mathbf{b})$ and $p(G)$. Eq. (6) is simply a special case of Eq. (9) by setting $p(\mathbf{b}) = p(G) = 1$. We can make intuitive discussions about the effect of biological parameters from Eq. (9). For example, if the intensity of each species is high, it is likely to increase the area $\nu^{nz}(S_{out})$ and it increases the number of species in given area, and vice versa. In this situation, the same discussion about the slope of SAR curve Eq. (8) can be made as long as the sampling region is significantly larger than geographic range of each species, $S \gg G_i$ for all $i$.

## 2.3 Endemic area relationship

Next we discuss the derivation of the endemic area relationship (EAR). Despite its importance to conservation biology, understanding of the empirical shape across scales is largely deficient [10]. Here, we define the endemic species, with explicitly introducing geographic range with an arbitrary shape, as the species with its geographic range $G$ is completely included in the sampling region $G \subset S$. This definition is different from the commonly used definition (e.g., [10, 11, 16]) where the endemic species is defined, using the SAR, as the average number of species whose population is completely covered by a region $A$ with area $\nu(A)$: $N_{en}(\nu(A)) = N_s(\nu(A')) - N_s(\nu(A' \setminus A))$, where $N_{en}(\nu(A))$ is the number of endemic species in area $\nu(A)$, and $\nu(A')$ is the system size. This definition relies on the occurrence probability of individuals of a species, and hence its individual distribution pattern. Typically SAR curves show $N_s(\nu(A)) > 0$ with positive area $\nu(A) > 0$, and therefore $N_{en}(\nu(A)) > 0$ for any small area $\nu(A)$. On the other hand, our definition is based on an explicit form of the endemic region, and the number of the endemic species is 0 if the sampling region is smaller than geographic range of all species: $S < G_i$ for all $i$. In this framework, individual distribution patterns do not affect the EAR. This definition may be more straightforward to discuss an extinction of species, since habitable regions of all species are explicitly taken into account. However, as we will see below, the two definitions become equivalent in the large limit of $A$.

In our framework, the number of endemic species in a given area is calculated by the endemic probability $p^{en}(\mathbf{x}_c^S, \mathbf{x}_c^G)$: the probability that a geographic range with its geometric

center $\mathbf{x}_c^G \in S \setminus S_{core}$ becomes subset of the sampling region with the geometric center $\mathbf{x}_c^S$, $G \subset S$. $S_{core}$ is the subset of the sampling region $S$ where if the center of a geographic range $\mathbf{x}_c^G$ falls in this regions, then the entire geographic range $G$ is always included in $S$: $P(G \subset S) = p^{en}(\mathbf{x}_c^S, \mathbf{x}_c^G) = 1$ provided that $\mathbf{x}_c^G \in S_{core}$ and $S > G$ (Fig. 1c). See Appendix A for further explanation of $S_{core}$.

The number of endemic species within the sampling region $S$ is calculated as

$$\begin{aligned} N_{es}(\nu(S)) &= \int_S f(\mathbf{x}) p^{en}(\mathbf{x}_c^S, \mathbf{x}) d\mathbf{x}, \\ &= \int_{S_{core}} f(\mathbf{x}) d\mathbf{x} + \int_{S \setminus S_{core}} f(\mathbf{x}) p^{en}(\mathbf{x}_c^S, \mathbf{x}) d\mathbf{x}, \\ &= \lambda_s(\nu(S_{core}) + \nu^{en}(S \setminus S_{core})), \end{aligned} \tag{10}$$

where, $\nu^{en}(S_{core})$ is the area of $S \setminus S_{core}$ weighted by the endemic probability $p^{en}(\mathbf{x}_c^S, \mathbf{x}_c^G)$ over this region. As in the case of the SAR, we can show the slope of the EAR on a log-log plot approaches 1 when $S \gg G$. In this limit, $S_{core}$ is nearly equivalent to $S$, and hence Eq. (10) becomes

$$\begin{aligned} N_{es}(\nu(S)) &\simeq \int_S f(\mathbf{x}) d\mathbf{x}, \\ &= \lambda_s \nu(S). \end{aligned} \tag{11}$$

This is equivalent to Eq. (8), and hence the slope approaches 1 at a large limit of sampling area. As we noted, the same result can be obtained in another definitions using the SAR, since $N_{en}(\nu(A)) \simeq N_s(\nu(A))$ when $A$ approaches $A'$.

We now introduce the EAR in the situation where species have different geographic ranges $G \in \mathcal{G}$. As in the case of $S_{out}$, $S_{core}$ is also uniquely determined by each $G$. By introducing the variation in geographic range, Eq. (10) now becomes

$$\begin{aligned} N_{es}(\nu(S)) &= \int_{G' \in \mathcal{G}} p(G') \left( \int_{G'} f(\mathbf{x}) d\mathbf{x} + \int_{S \setminus S_{core}^{G'}} f(\mathbf{x}) p^{en}(\mathbf{x}_c^S, \mathbf{x}) d\mathbf{x} \right) dG', \\ &= \lambda_s \mathrm{E}_{S_{core}}[\nu(S_{core}) + \nu^{en}(S \setminus S_{core})]. \end{aligned} \tag{12}$$

As in the case of the SAR, we obtain Eq. (10) by setting $p(G) = 1$ in Eq. (12). In addition, the same discussion about the slope of the EAR can be made in this situation, if the sampling region is significantly larger than the geographic ranges: $S \gg G_i$ for all $i$.

## 2.4 Relative species abundance

Here, we develop a method to derive the relative species abundance (RSA). RSA, or SAD, a histogram of species count, can be affected by multiple ecological mechanisms such as species interactions and demographic stochasticity [7]. RSAs derived here are the expected value of assemblage of community with an identical area, and we do not consider the effect of the undersampling. For convenience, we use the notation $P_{\nu(S)}(X = x)$ to describe the

probability of finding a species with abundance $x$ in a sampling region with the area $\nu(S)$, and $P_{\nu(S)}(x)$ and $P(x)$ are also used when no confusion occurs. Derivation of the RSA is more cumbersome than the SAR and EAR since we use all probabilities of $P_{\nu(S)}(X = x)$. In our framework, the average abundance of species $i$ is proportional to the area where the sampling region $S$ and the geographic range of a species $G_i$ are overlapped, $\nu(S \cap G_i)$. Each sampled species may have different overlapped area ($\nu(S \cap G_i) \neq \nu(S \cap G_j),\ (i \neq j)$), and its area changes as sampling and range scales as well as these shapes change. Therefore, to calculate the RSA across scales we need a pdf of $\nu(S \cap G)$, provided $\nu(S \cap G) \neq \phi$, which describes the variation of $\nu(S \cap G)$ between sampled species.

Roughly speaking, there may exist two regimes to determine the maximum value of $\nu(S \cap G)$ among sampled species: $S > G$ and $S < G$, provided distance relationships $l_S > L_G$ and $L_S < l_G$, respectively. Namely, the maximum overlapped area is either $\nu(S)$ or $\nu(G)$ depending on $S < G$ (Fig. 2a, b) or $S > G$ (Fig. 2c, d), respectively: $\max\{\nu(S \cap G)\} = \min\{\nu(S), \nu(G)\}$. This maximum value in the overlapping area occurs if the geometric center

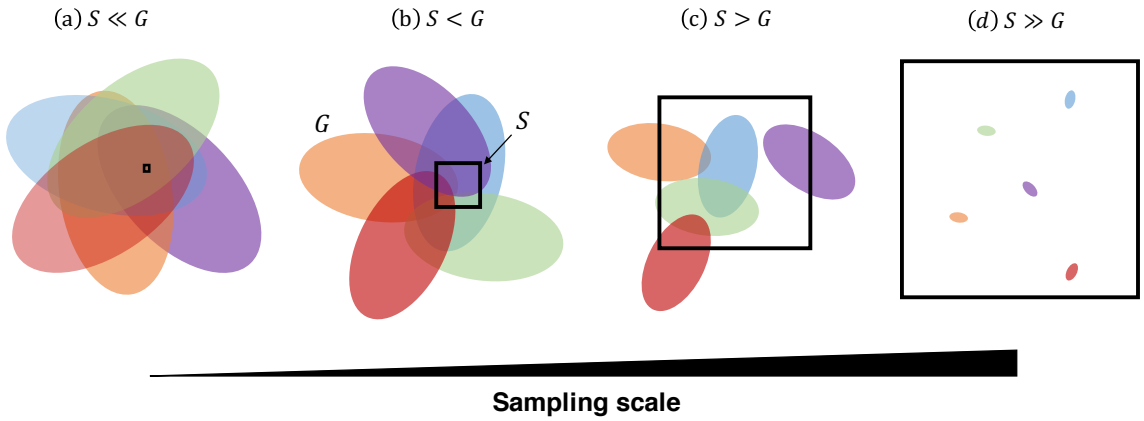


Figure 2: Four different sampling phases across sampling region (black square). As an example, the sampling region $S$ and species geographic range $G$ are depicted by squares and colored ellipses, and geographic ranges of five species are shown in different colors. In each panel, different sampling scales are shown, but the size of ranges remain the same and configuration of the geographic ranges is changed for a presentation purposes. (a) and (d): when the sampling scale is sufficiently small or large, all sampled species have ranges completely overlapping with the the sampling region ($\nu(S)$ and $\nu(G)$, respectively). (b) and (c): in intermediate sampling scales, this overlapping region is different between sampled species.

of the geographic range $\mathbf{x}_c^G$ falls in $S_{core}$: $P(\nu(S \cap G) = \min\{\nu(S), \nu(G)\}) = 1$ provided $\mathbf{x}_c^G \in S_{core}$ (Fig. 1d). As the assumption of random placement of the center of geographic range, the probability that the overlapped area is its maximum value is proportional to the

area of the $S_{core}$ in the region $S_{out}$. Therefore, we have the following relationships:

$$\begin{cases} P(\nu(S \cap G) = \min\{\nu(S), \nu(G)\}) = \frac{\nu(S_{core})}{\nu(S_{out})}, \\ P(0 \leq \nu(S \cap G) < \min\{\nu(S), \nu(G)\}) = \frac{\nu(S_{out} \setminus S_{core})}{\nu(S_{out})}. \end{cases} \tag{13}$$

As implied in Fig. 2a and d, when $S \ll G$ or $S \gg G$ is satisfied, $P(\nu(S \cap G) = \min\{\nu(S), \nu(G)\}) \simeq 1$ is satisfied. The abundance of a species within overlapping area $\nu(S \cap G)$ shows variation, and its pdf corresponds to the relative species abundance in area $\nu(S \cap G)$. Let $P_{\nu(S)}(x \mid \nu(S \cap G))$ be the probability of finding a species with $x$ individuals in the sampling region $S$ given the overlapped area $\nu(S \cap G)$. Then the RSA provided an arbitrary sampling area $\nu(S)$ is described as

$$\begin{aligned} P_{\nu(S)}(X = x) \quad = \quad & \frac{\nu(S_{core})}{\nu(S_{out})} P_{\nu(S)}(x \mid \nu(S \cap G) = \min\{\nu(S), \nu(G)\}) + \\ & \frac{\nu(S_{out} \setminus S_{core})}{\nu(S_{out})} P_{\nu(S)}(x \mid 0 \leq \nu(S \cap G) < \min\{\nu(S), \nu(G)\}), \end{aligned} \tag{14}$$

where, $P_{\nu(S)}(x \mid \nu(S \cap G) = \min\{\nu(S), \nu(G)\})$ and $P_{\nu(S)}(x \mid 0 \leq \nu(S \cap G) < \min\{\nu(S), \nu(G)\})$ correspond to two pdfs. Namely Eq. (14) is sam of two pdfs weighted by Eq. (13). Especially, when $S \ll G$ or $S \gg G$ is satisfied, Eq. (14) is simplified to

$$P_{\nu(S)}(X = x) \simeq P_{\nu(S)}(x \mid \nu(S \cap G) = \min\{\nu(S), \nu(G)\}). \tag{15}$$

It turns out that in this limit (Fig. 2a and d), the RSAs share the same form of a pdf with different parameter values due to a change in $\min\{\nu(S), \nu(G)\}$. Therefore, it provides a potential upscaling (downscaling) framework within sampling scales at small/large limit, as well as between sampling scales at small and large limits as long as the schemes in Fig. 2a and d hold.

In line with the discussion above, it is straightforward to incorporate the species variations such as biological parameters $\mathbf{b}$ and the geographic range $G \in \mathcal{G}$, where $G$ uniquely determines $S_{core}$ and $S_{out}$. By introducing pdfs of these parameters $p(\mathbf{b})$ and $p(G)$, and averaging Eq. (14) by the pdfs, we obtain

$$\begin{aligned} P_{\nu(S)}(x) \quad = \quad & \int_{\mathbf{b} \in \mathbf{B}} p(\mathbf{b}) \int_{G' \in \mathcal{G}} p(G') \{ \frac{\nu(S_{core})}{\nu(S_{out})} P_{\nu(S)}(x \mid \min\{\nu(S), \nu(G')\}) + \\ & \frac{\nu(S_{out} \setminus S_{core})}{\nu(S_{out})} P_{\nu(S)}(x \mid 0 \leq \nu(S \cap G') < \min\{\nu(S), \nu(G')\}) \} \, dG' d\mathbf{b}. \end{aligned} \tag{16}$$

It is worth noting that, as Fig. 2 and Eq. (14) imply, sampling schemes affect the RSA by two ways. First, different sampling scales cause different overlapping patterns with the geographic range of each sampled species. Second, there are sampling schemes that does not have phases shown in Fig. 2, such as line transect, and pooling data. In that case, sampling regions with the same area may produce different RSA patterns.

# 3 Emergent macroecological patterns of geometric model

We use the general theory developed above to acquire new insights into effects of scale and space on biodiversity pattern. First, we need to introduce a point field $f(\mathbf{x})$ defined above that holds information of individual distributions of all species. For this purpose, we make use of spatial point processes [25,26], a set of spatially explicit stochastic models that generate various point distribution patterns such as random and clustering patterns. One of the advantages of this model is that spatial point processes are amenable to mathematical analysis and there have been a number of applications to ecological studies [11,21,22,24,27]. Although the underlying assumptions are simple the models can provide consistent patterns with observed SARs or a population occupancy probability [11,24,27]. See some properties of spatial point processes used in this paper in Appendix B.

Here, we assume that sampling region and the species geographic ranges are described by circles (Fig. 3a). Also, individual distributions therein are described by the homogeneous Poisson process or Thomas process, showing random and clustering distribution patterns, respectively. The assumption of the shapes makes mathematical analysis rather simple and transparent as we can omit the effect of rotation of the geographic ranges. As we will see below, the second term of the EAR (Eqs. 10, 12) disappears under this assumption, since the endemic probability $p^{en}(\mathbf{x}_c^S, \mathbf{x}_c^G) = 1$, if $\mathbf{x}_c^G \in S_{core}$ is satisfied. For convenience, all the analyses below are conducted in the polar coordinate. The schematic diagrams under this situation corresponding to Fig. 1 are shown in Fig. 3.

Let the intensity measure of species, $\mu_s$, be a function $f(r, \theta)$ defined in the polar coordinate and as we assume the homogeneous environment and no interaction between species (see General Theory), the intensity measure of species in the circle with a radius $R$ is

$$\begin{aligned} \mu_s &= \int_0^{2\pi} \int_0^R f(r,\theta) r dr d\theta, \\ &= 2\pi \int_0^R f(r) r dr. \end{aligned} \tag{17}$$

If we define $\mu_s := \lambda_s \pi R^2$, where $\lambda_s$ is the intensity of species, $f(r)$ is calculated as

$$f(r) = \lambda_s. \tag{18}$$

## 3.1 Species area relationship

Let us consider the situation where each species has the area of geographic range $\nu(G) = \pi r_g^2$ characterized by its radius $r_g$. We assume that centers of geographic ranges are randomly distributed across space, and individual distributions therein show either random or clustering pattern. Then our sampling regime is introduced in such a way that we randomly place the sampling region with radius $r_s$ and count the number of species across spatial scales by changing the size of sampling region. By doing so, we obtain the SAR (Fig. 3a). From Eqs

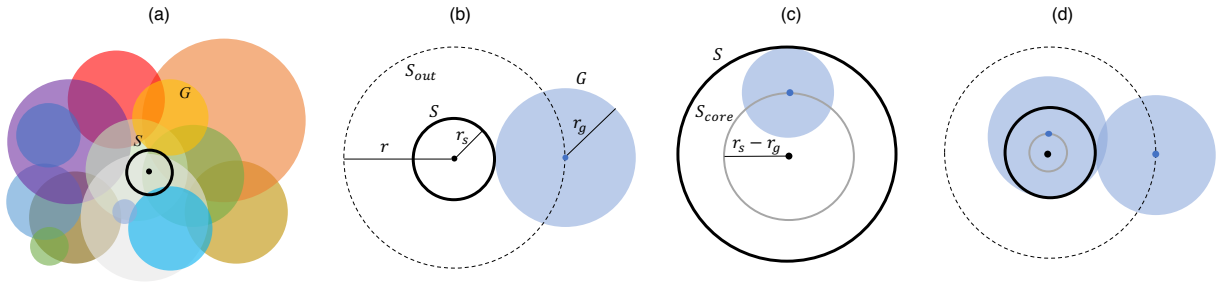


Figure 3: Schematic representations of (a) the sampling region and geographic ranges, and typical geometric relationships used in derivations of the (b) SAR, (c) EAR, and (d) RSA. $S$ is the sampling region with radius $r_s$, $G$ is the geographic range of a species with radius $r_g$. The distance between the centers of $S$ and $G$ is represented by $r$ (in this example, $r = r_s + r_g$). See Fig. 1 for more explanations.

(17, 18) and with the aid of Fig. 3b, we calculate the number of species $N_s$ found within a sampling unit with area $\pi r_s^2$, as

$$\begin{aligned} N_s(\pi r_s^2) &= \int_0^{2\pi} \int_0^{r_s+r_g} f(r,\theta) p^{nz}(r, r_s \mid r_g) dr d\theta, \\ &= 2\pi\lambda_s \int_0^{r_s+r_g} r p^{nz}(r, r_s \mid r_g) dr, \\ &= 2\pi\lambda_s \mathrm{E}_{nz}[r], \end{aligned} \tag{19}$$

where, $p^{nz}(r, r_s \mid r_g)$ is the non-zero probability, provided $r_s$ and $r_g$, that at least one individual of a specie is found given an $r$-distance separation of the centers of sampling and geographic range. Note that $2\pi \int_0^{r_s+r_g} r p^{nz}(r, r_s \mid r_g) dr$ in the second line gives the area corresponding to $\nu^{nz}(S_{out})$ in Eq. (6). Eq. (19) implies that the number of species within the sampling region $S$ is proportional to the average distance that at least one individuals are found.

We can generalize Eq. (19) to a situation where each species has a different geographic range size, therefore different radius of the geographic range $r_g$, and biological parameter $\mathbf{b}$. Then $r_g$ and $\mathbf{b}$ follow probability distribution functions, $p(r_g)$ and $p(\mathbf{b})$, respectively. In that situation, Eq. (19) becomes

$$\begin{aligned} N_s(\pi r_s^2) &= \int_{\mathbf{b}\in\mathbf{B}} p(\mathbf{b}) \int_0^{\infty} p(r_g') \int_0^{2\pi} \int_0^{r_s+r_g'} r p^{nz}(r, r_s \mid r_g') dr d\theta dr_g' d\mathbf{b}, \\ &= 2\pi\lambda_s \mathrm{E}_{\mathbf{b}} \mathrm{E}_{r_g} \mathrm{E}_{nz}[r]. \end{aligned} \tag{20}$$

Since this is a special case of the general theory developed above, the discussion about asymptotic slope (Eq. 8) holds.

Here, to show some numerical results in the situation where the geographic range and/or biological parameters differ between species, we examine a situation where the radius of

geographic range $r_g$ follows the exponential distribution with the parameter $\lambda_e$

$$p(r_g) = \lambda_e e^{-\lambda_e r_g}. \tag{21}$$

For the biological parameter, we assume that the intensity of each species $\lambda_i$ follows the gamma distribution:

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}, \tag{22}$$

where, $\alpha$ and $\beta$ are shape and rate parameters of the gamma distribution respectively. We replace $\lambda$ by $\lambda_i$ or $\lambda_{th,i}$ depending on the underlying individual distribution process. In addition, we define $\bar{c}_i$ and $\lambda_i^p$ as $\sqrt{\lambda_{th,i}}$ to satisfy $\lambda_i = \lambda_{th,i} = \bar{c}_i \lambda_i^p$. Fig. 4 shows numerically calculated values under cases where species parameters are identical (Eq. 19) and variable (Eq. 20) situations, showing the tri-phasic feature on a log-log plot with different geographic range and individual distribution patterns. The third phase of the SAR appears around the average area of the geographic range. As discussed above, the slope of each curve asymptotically approaches 1 at large scales. Also, if the geographic range is smaller, it approaches the asymptotic value faster as discussed above. Moreover, it shows that the differences between the two individual distribution patterns are relatively small and slight deviations appear only on small sampling scales. The differences where species have either equivalent or variable intensity $\lambda_i$ are negligibly small when individual distributions is described by the homogeneous Poisson process with the radius of the geographic range $r_g = 10$km. We checked this holds true for the other curves, and also for different parameter sets of gamma distribution.

## 3.2 Endemic area relationship

When both the sapling region and endemic region are circle, the geographic range of an endemic species satisfies the condition shown in Fig. 3c: the sampling region must be larger than geographic range (i.e., $r_s > r_g$), and the distance between $S$ and $G$ must be small enough (i.e., $r \leq r_s - r_g$). Therefore, the number of endemic species $N_{es}$ in area $\pi r_s^2$ is described as

$$\begin{aligned} N_{es}(\pi r_s^2) &= \int_0^{2\pi} \int_0^{r_s - r_g} f(r,\theta) dr d\theta, \\ &= \pi \lambda_s (r_s - r_g)^2. \end{aligned} \tag{23}$$

By applying Eq. (7) to calculate the slope of the EAR, it is easily shown provided $r_s > r_g$

$$\frac{d \log N_{es}(\pi r_s^2)}{d \log(\pi r_s^2)} = \frac{r_s}{(r_s - r_g)^2} \left(1 - \frac{r_g}{r_s}\right). \tag{24}$$

As discussed above, it approach 1 as the sampling region becomes significantly larger than the geographic range: $r_s \gg r_g$.

As above, we can generalize the EAR to incorporate variations in geographic range sizes of species by introducing a pdf of the radius of the geographic range, $p(r_g)$. However, as

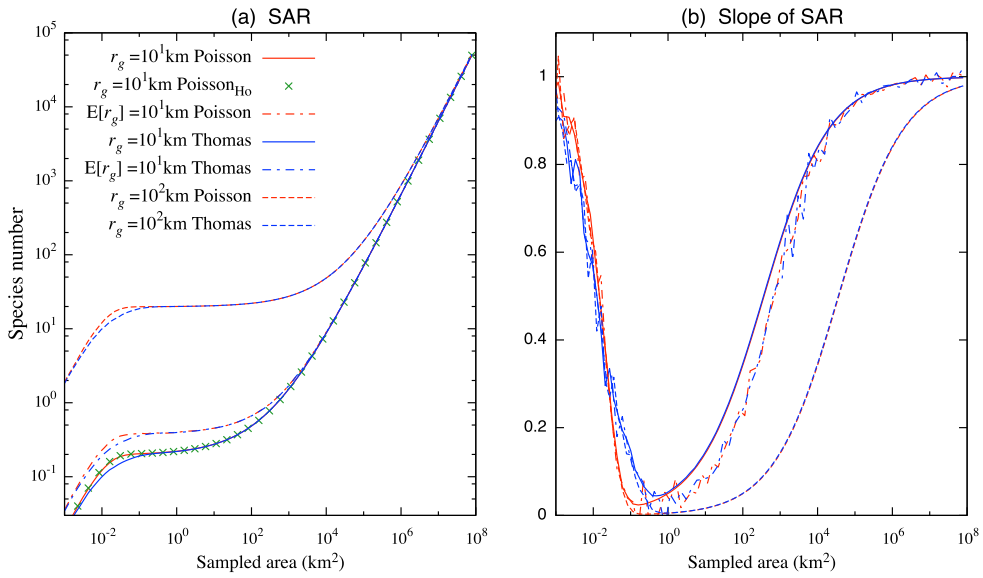


Figure 4: The species area relationship [SAR; (a)] and its slope (b) under the homogeneous Poisson process (red and green) and Thomas process (blue), obtained using Eqs. (19) and (20). The intensity $\lambda_i$ varies between species according to the gamma distribution except for the homogeneous situation (cross; labeled Poisson$_{\rm Ho}$). The lines labeled E[$r_g$] represents the situation where the radius of geographic ranges follows the exponential distribution. Otherwise, the radius of geographic range is identical between species (line; $r_g$=10km line, and dashed; $r_g$=100km). For the Thomas process, the parameters are $\lambda_s = 0.00064$ (50000 species/$\pi \times 5.0 \times 10^8$km$^2$), $\lambda_{th,i} = 100$, $\bar{c}_i = 10$, $\lambda_i^p = 10$, $\sigma_p = 0.1$. For the gamma distribution, we used $\alpha = 10$ and $\beta = 0.1$ (E$[\lambda_i] = 100$, Var$[\lambda_i] = 1000$), and, for the exponential distribution, $\lambda_e$=0.1 (E$[r_g] = 10$, Var$[r_g] = 100$).

noted above, biological parameters do not affect the EAR. For the sake of comparison with the case of the constant $r_g$, let us assume the average value satisfies E$[r_g] = r_g$. Eq. (23) now becomes

$$\begin{aligned} N_{es}(\pi r_s^2) &= \int_0^{2\pi} \int_0^{r_s} p(r_g') \int_0^{r_s - r_g'} f(r,\theta) dr dr_g' d\theta, \\ &= 2\pi\lambda_s \int_0^{r_s} p(r_g') \int_0^{r_s - r_g'} r dr dr_g', \\ &= \pi\lambda_s \int_0^{r_s} p(r_g')(r_s - r_g')^2 dr_g'. \end{aligned} \tag{25}$$

As before, let us assume the pdf of the radius $r_g$ follows the exponential distribution $p(r_g) =$

$\lambda_e e^{-\lambda_e r_g}$. Then, the number of endemic species is

$$\begin{aligned} N_{es}(\pi r_s^2) &= \pi\lambda_s \int_0^{r_s} \lambda_e e^{-\lambda_e r_g}(r_s - r_g)^2 dr_g, \\ &= \frac{2\lambda_s\pi}{\lambda_e^2}\left(1 - r_s\lambda_e + \frac{(r_s\lambda_e)^2}{2} - e^{-r_s\lambda}\right), \\ &= \frac{2\lambda_s\pi}{\lambda_e^2}\sum_{n=3}(-1)^{n+1}\frac{(r_s\lambda_e)^n}{n!}. \end{aligned} \tag{26}$$

Since $dN_{es}/\pi r_s^2 = \lambda_s \sum_{n=2}(-1)^n \frac{(r_s\lambda_e)^{n-1}}{n!}$, we calculate the slope as

$$\frac{d\log N_{es}(\pi r_s^2)}{d\log(\pi r_s^2)} = -\frac{r_s\lambda_e}{2}\frac{1 - r_s\lambda_e - e^{-r_s\lambda_e}}{1 - r_s\lambda_e + \frac{(r_s\lambda_e)^2}{2} - e^{-r_s\lambda_e}}. \tag{27}$$

Fig. 5 shows the EAR calculated by Eqs. (23) and (25), and its slope defined by Eqs. (24) and (27). As discussed in the General Theory section, the EAR on a log-log plot shows asymptotic slope 1, equivalent to the SAR when the sampling region is significantly larger than the geographic range sizes $S \gg G$. By the same discussion above, it occurs earlier when the range sizes are small. The slope of the EAR becomes infinitely large as soon as the sampling size becomes equivalent to single-sized geographic range. On the other hand, if radii of the geographic range are exponentially distributed, the slope approaches 1 from a finite value (Fig. 5b), and it also shows that, at small scales, the endemic species is proportional to the sampling area.

### 3.3 Relative species abundance

To derive RSAs across sampling scales, we extensively use mixed probability distributions, especially mixed Poisson distributions [28]. The mixed Poisson distribution was first introduced in ecological study by Fisher et al. [2] in which the celebrated Fisher's logseries was derived. In our geometric approach, mixed Poisson distributions appear naturally by the nature of multiple stochasticity to determine the number of individuals observed. Namely, there exist intra-species (variations from the expected number of individuals) and inter-species variations (variations of the expected number of individuals itself) of individuals within an overlapped area between the sampling and geographic region $\nu(S \cap G)$. In addition, the overlapped area also varies between species, and the probability distribution of the area $P(\nu(S \cap G))$ depends on sampling scale as shown in Fig 2. Hence, we need to resolve the scale effect to derive the RSA across scales.

Mixed Poisson distributions are known to produce a variety of probability functions [28], and it is hard to make an exhaustive list of potential RSAs. However, we can consistently obtain RSAs in arbitral sampling scales $S$ by scaling up/down, once all the parameters are provided. Therefore, we will first focus to recover some well documented RSAs, when the sampling region is much smaller than the geographic range size $S \ll G$, probably the most common situation in practice. Second, we will scale up the sampling region $S$, by keeping the

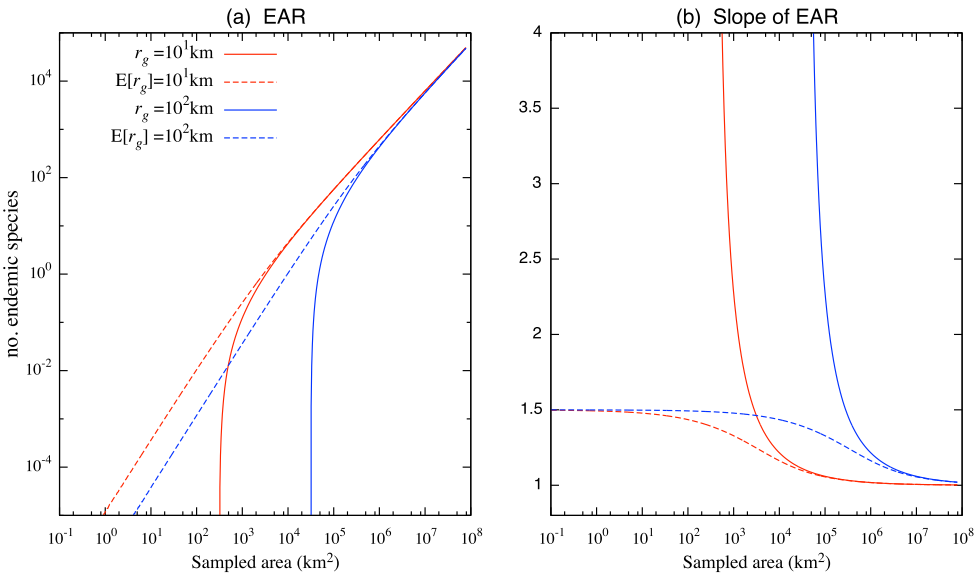


Figure 5: (a) The endemic-area relationship (EAR) and (b) its slope. Color corresponds to geographic range size (red; $r_g$ = 10km and blue; $r_g$ = 100km). Solid lines represent the situation where no variation in geographic range size occurs (Eq. 23). Dotted lines correspond to the situation where the radius of the geographic range varies between species following an exponential distribution (Eq. 26). To obtain the average value $\mathrm{E}[r_g] = \{10, 100\}$, we set the parameter $\lambda_e = \{0.1, 0.01\}$. For other parameters, the same values are used as in Fig. 4.

same assumption, to see how RSA will change. In the first step, we will recover the negative binomial distribution [29], Fisher's logseries [2], and Poisson lognormal distribution [30].

In the following analysis, we assume that individual distributions are random: it is described by the homogeneous Poisson process. However, as we will see below, the difference of RSAs between the homogeneous Poisson process and Thomas process are small to discuss qualitative features of the RSA, except for the small scales where the deviations of two processes appear in the SAR (Fig. 4).

### 3.3.1 Some basic properties of mixed Poisson distribution

Here, we introduce some basic properties of mixed Poisson distributions used in the following analysis. For a mixed Poisson distribution, the Poisson intensity itself also follows a probability distribution $g(\lambda)$, and the probability is described [28]

$$P(X = x) = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} g(\lambda) d\lambda. \tag{28}$$

As in [28], we denote this mixture as

$$f(x \mid \lambda) \underset{\lambda}{\wedge} g(\lambda). \tag{29}$$

Note that in our application of the homogeneous Poisson process where the probability variable is the number of individuals, the intensity $\lambda$ in Eq. (28) is replaced by the intensity measure $\mu_i = \lambda_i \nu(A)$ where $\lambda_i$ and/or $\nu(A)$ vary independently. To specify the probability variable, we use the following expressions:

$$f(x \mid \lambda_i \nu(A)) \underset{\lambda_i}{\wedge} g(\lambda_i), \tag{30}$$

$$f(x \mid \lambda_i \nu(A)) \underset{\lambda_i}{\wedge} g(\lambda_i) \underset{\nu(A)}{\wedge} h(\nu(A)), \tag{31}$$

where, the first expression is the case for only $\lambda_i$ varies and the second expression is the case for both $\lambda_i$ and $\nu(A)$ vary.

From the Proposition 1 and 2 in Appendix E, the probabilities of the mixed Poisson distribution shown in Eqs. (30) and (31) are described

$$P(X = x) = \frac{1}{x!} \sum_{r=0}^{\infty} \nu(A)^{x+r} \frac{(-1)^r}{r!} \mu_{x+r}(\lambda_i), \tag{32}$$

$$P(X = x) = \frac{1}{x!} \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} \mu_{x+r}(\lambda_i) \mu_{x+r}(\nu(A)), \tag{33}$$

respectively, where $\mu_r(x)$ is the $r$th moment of $x$ about the origin. As we see below, $\nu(A) = \min\{\nu(S), \nu(G)\}$ in our application, and these are RSAs we explore.

### 3.3.2 Sampling region is much smaller than geographic range: $S \ll G$

In our framework, if species $i$ is sampled, the range of species $i$ must be overlapping with the sampling region, $S \cap G_i \neq \phi$. Provided this property and if the sampling region is much smaller than the geographic range, the sampling region is completely included in the geographic range $P(\nu(S \cap G_i) = \nu(S)) \simeq 1$ as in Fig. 2a. In that situation, we can neglect the variations of the overlapped region between species, and consider only the abundance variations of inter- and intra-species within the the sampling region $S$. Namely, this is the situation described by Eq. (30) and the RSA under this condision is obtained by calculating Eq. (32).

#### Negative binomial (Poisson-gamma) distribution and Fisher's logseries

It is well known that the mixing the Poisson distribution with the gamma distribution produces the negative binomial distribution [28]. By taking certain limits of the negative binomial distribution, Fisher et al. [2] obtained the Fisher's logseries. Here, we follow the same

idea to derive these two distributions. First, by Eq. (28), mixing the Poisson distribution with the gamma distribution $g(\lambda)$, Eq. (22) gives the following probability function

$$P(X=x) = \frac{\beta^\alpha}{x!\Gamma(\alpha)} \int_0^\infty \lambda^{x+\alpha-1} e^{-\lambda(1+\beta)} d\lambda, \tag{34}$$

On the other hand, the mixture in our situation is described by Eq. (30), not by Eq. (29), and the probability of finding a species with $x$ individuals are described

$$\begin{aligned} P(X=x) &= \int_0^\infty \frac{(\lambda_i \nu(S))^x}{x!} e^{-\lambda_i \nu(S)} \frac{\beta'^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\lambda_i \beta'} d\lambda_i, \\ &= \frac{\beta'^\alpha \nu(S)^x}{x!\Gamma(\alpha)} \int_0^\infty \lambda_i^{x+\alpha-1} e^{-\lambda_i(\beta'+\nu(S))} d\lambda_i, \end{aligned} \tag{35}$$

where, we set $\beta' = \beta\nu(S)$ and changing the variable as $\nu(S)\lambda_i = \lambda_i'$, we recover the form of Eq. (34). Eq. (34) is well known to produce the negative binomial distribution

$$P(X=x) = \left( \begin{array}{c} x+\alpha-1 \\ x \end{array} \right) p^x (1-p)^\alpha, \tag{36}$$

where, in our case $p = \nu(S)/(\nu(S)+\beta)$.

To obtain Fisher's logseries from Eq. (36), we need to further assume the sampling effect: the number $s$ of individuals are sampled. By taking the Fisher's limit [31], we obtain:

$$P(X=x) = \lim_{\substack{s\to\infty,\, \alpha\to 0 \\ s\alpha\to\gamma}} s \left( \begin{array}{c} x+\alpha-1 \\ x \end{array} \right) p^x (1-p)^\alpha = \frac{\gamma p^x}{x}. \tag{37}$$

**Poisson-lognormal distribution**

The Poisson-lognormal distribution is obtained by mixing the Poisson distribution with the lognormal distribution $g(\lambda) = 1/(\lambda\sqrt{2\pi\sigma^2}) e^{-(\log\lambda-\mu)^2/2\sigma^2}$ [30]:

$$P(X=x) = \frac{1}{x!\sqrt{2\pi\sigma^2}} \int_0^\infty \lambda_i^{x-1} e^{-\lambda_i - \frac{(\log\lambda_i-\mu)^2}{2\sigma^2}} d\lambda_i \tag{38}$$

Instead, in our situation, the mixture is as in Eq. (30), and it is described

$$\begin{aligned} P(X=x) &= \int_0^\infty \frac{(\lambda_i\nu(S))^x}{x!} e^{-\lambda_i\nu(S)} \frac{1}{\lambda_i\sqrt{2\pi\sigma^2}} e^{-\frac{(\log\lambda_i-\mu)^2}{2\sigma^2}} d\lambda_i, \\ &= \frac{\nu(S)}{x!\sqrt{2\pi\sigma^2}} \int_0^\infty (\lambda_i\nu(S))^{x-1} e^{-\lambda_i\nu(S) - \frac{(\log\lambda_i-\mu)^2}{2\sigma^2}} d\lambda_i, \end{aligned} \tag{39}$$

where, by setting $\mu = \mu' - \log\nu(S)$ and $\nu(S)\lambda_i = \lambda_i'$, we recover the form of Eq. (38). Therefore, if the parameter $\lambda_i$ follows the lognormal distribution and sampling region is much smaller than the geographic range, we expect to observe an RSA curve that follows the Poisson-lognormal distribution.

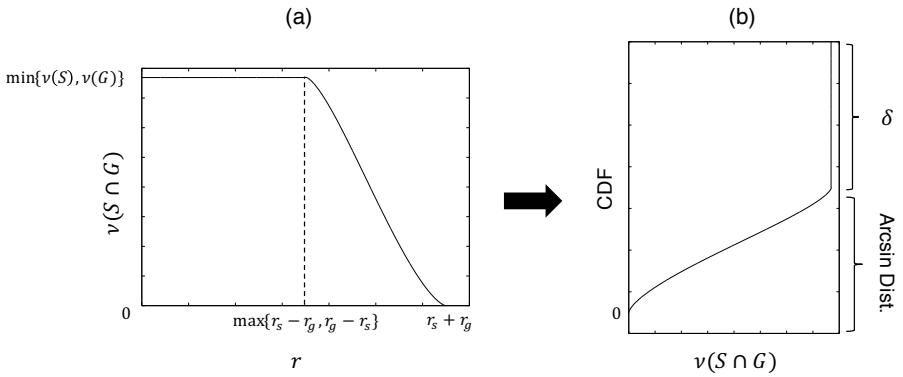


Figure 6: An example form of (a) the overlapped area $\nu(S \cap G)$ given an $r$-distance separation of geometric between the circle sampling region and the geographic range of the species (both have a circle shape), and (b) the cumulative distribution functions of overlapped area.

### 3.3.3 General situation

With the results of the analysis above, we can explore more general situations where the sampling region does not necessarily satisfy $S \ll G$. In these situations, the probability function of finding a species with $x$ individuals has the form of Eq. (14), in which as in Eq. (15), the above-mentioned situation $S \ll G$ (and $S \gg G$) is treated as a special situation. The weighting terms Eq. (13) in this example are easily calculated with the aid of Fig. 3:

$$\begin{cases} \frac{\nu(S_{core})}{\nu(S_{out})} = \frac{\max\{r_g - r_s, r_s - r_g\}}{r_g + r_s}, \\ \frac{\nu(S_{out} \backslash S_{core})}{\nu(S_{out})} = \frac{\min\{2r_s, 2r_g\}}{r_g + r_s}, \end{cases} \tag{40}$$

where, one can easily see that the second weighting term disappears when $S \ll G$ and $S \gg G$, namely $r_s \ll r_g$ and $r_s \gg r_g$, respectively. The first probability function of Eq. (14) corresponds to the situation where we can neglect the variation of overlapped region, and this situation was already discussed above. Namely, it is described by Eq. (32). The second probability of Eq. (14) corresponds to the situation where the overlapped area varies between species. As we discussed above, the mixed probability distribution has the relationship of Eq. (31), and the probability of finding a species with $x$ individuals is described by Eq. (33).

For simplicity, we focus here on the case where the $\lambda_i$ follows the gamma distribution as in the first example above (Poisson-gamma), since this case is mathematically more amenable than the case of Poisson-lognormal, and therefore it allows us to discuss results in a mathematically more transparent way. This sole assumption, however, produces a variety of RSAs including similar forms to the negatively skewed lognormal distribution [6], or a left-skewed bell shape curve on a logarithmic scale, resembling the Poisson-lognormal distribution.

To make use of Eq. (33), we need the $r$th moment of a pdf of the overlapped area $\nu(S \cap G)$. Fig. 6a shows, provided $S$ and $G$ are circle, an example of the relationship of the overlapped area $\nu(S \cap G)$ given an $r$-distance separation of geometric centers between the sampling region

and the geographic range of species. To infer the pdf, we use Fig. 6b, obtained by rotating Fig. 6a and normalizing its maximum value to 1, as a cumulative distribution function of $P(\nu(S \cap G))$. Since this function has two different phases, we decompose it into two parts. The region described with $\delta$ corresponds to the first probability function in Eq. (14) in which the probability is described by a delta distribution and has a peak at $\min\{\nu(S), \nu(G)\}$. The rest corresponds to the second probability function in Eq. (14), and it has a similar shape to an arcsin function. Because we require the moment of a probability distribution function, $P(\nu(S \cap G))$, it may be a reasonable to approximate it by an existing pdf. Therefore, we apply the arcsin distribution, which is a special case of the beta distribution [32]

$$f(x) = \frac{1}{\pi\sqrt{x(\min\{\nu(S), \nu(G)\} - x)}}, \qquad x \in [0, \min\{\nu(S), \nu(G)\}] \tag{41}$$

where, the support is defined as $x \in [0, \min\{\nu(S), \nu(G)\}]$. The $r$th moment of Eq. (41) about the origin is, by the Proposition 3 in Appendix, described by $\mu_r^{arc} = \frac{\nu(S)^r}{\pi}\mathrm{B}(\frac{1}{2}, \frac{1}{2} + r)$ where, $\mathrm{B}(x, y) = \int_0^1 u^{x-1}(1-u)^{y-1}du$ is the beta function. By substituting this equation, together with the $r$th moment of the gamma distribution Eq. (22), $\mu_r^{gam} = \frac{\Gamma(r+\alpha)}{\Gamma(\alpha)}\frac{1}{\beta^r}$, into Eq. (33), we obtain the following probability function (see Appendix C for the detailed derivation):

$$P(X = x) = \gamma \begin{pmatrix} x + \alpha - 1 \\ x \end{pmatrix} p^x (1-p)^\alpha \, {}_2F_1(\frac{1}{2} + x, x + \alpha; 1 + x; -\frac{\min\{\nu(S), \nu(G)\}}{\beta}), \tag{42}$$

where, $\gamma$ is the coefficient $\gamma = \Gamma(1/2+x)(\sqrt{\pi}\Gamma(x+1))^{-1}(1-p)^{-\alpha-x}$ and this part is simplified for $x \geq 1$ as $\gamma = x^{-1}\mathrm{B}(1/2, x)^{-1}(1-p)^{-\alpha-x}$, and $p$ is now becomes $p = \min\{\nu(S), \nu(G)\}$ $/(\min\{\nu(S), \nu(G)\} + \beta)$. The second to fourth factors correspond to the negative binomial distribution the form corresponds to Eq. (36), and ${}_2F_1$ is the Gauss hypergeometric function. We call this distribution the Poisson-gamma-arcsine distribution. From, Eqs. (36), (40), and (42), we derive the general form as in Eq. (14):

$$\begin{aligned} P(X = x) = &\begin{pmatrix} x + \alpha - 1 \\ x \end{pmatrix} p^x (1-p)^\alpha \times \\ &\left(\frac{\max\{r_g - r_s, r_s - r_g\}}{r_g + r_s} + \gamma\frac{\min\{2r_s, 2r_g\}}{r_g + r_s} \, {}_2F_1(\frac{1}{2} + x, x + \alpha; 1 + x; -\frac{\min\{\nu(S), \nu(G)\}}{\beta})\right), \end{aligned} \tag{43}$$

where, the probability distribution is a weighting sum of the negative binomial distribution and the Poisson-gamma-arcsine distribution. As already discussed above, $S \ll G$ ($S \gg G$) is the special case of Eq. (43). That is, by taking $r_s \ll r_g$ ($r_s \gg r_g$), Eq. (43) becomes

$$P(X = x) \simeq \begin{pmatrix} x + \alpha - 1 \\ x \end{pmatrix} p^x (1-p)^\alpha. \tag{44}$$

Namely, the RSAs follow the same probability distribution when the area of sampling region and geographic range are significantly different. It is worth emphasizing that to derive

Eq.(44), the term approximated by the arcsine distribution disappears. And, as noted in the General Theory section, this provides a potential upscaling (downscaling) framework between scales that are applicable the relationship of Eq. (44). Since all the parameters except for $\nu(S)$ are consistent across spatial scales, the difference caused by a scale change is only in the parameter value $p = \min\{\nu(S), \nu(G)\}/(\min\{\nu(S), \nu(G)\} + \beta)$. It is expected that $p$ is close to 1 if $\nu(G) = \min\{\nu(S), \nu(G)\}$, namely $\nu(G) \gg 1$.

The same discussion can be made when the intensity, $\lambda$, follows the lognormal distribution, hence the situation as in Eq. (38), and the RSA under this case is derived in Appendix C, but with a mathematically less tractable and computationally more intensive form. However, we numerically found that similar RSA patterns may be observed as in the case of the Poisson-gamma-arcsine distribution (Fig. 4 below). It may reflect the fact that the lognormal and the gamma distributions have a rather similar shape and the difference may not play a major role in fitting ecological data [31]. In addition, using Eq. (16) we can formally write down RSAs when the biological parameters and/or geographic range size differ between species. To see this, let us assume that the radii of geographic ranges across species follows an exponential distribution with a parameter $\lambda_e$ as above, and denote the right-hand side of Eq. (43) by $P(x, r_g)$. Then, the RSA corresponding to Eq. (16) has the following form

$$P(X = x) = \int_0^\infty \lambda_e e^{-\lambda_e r_g} P(x, r_g) dr_g. \tag{45}$$

However, this has a rather complicated form. Below we discuss numerical results where distribution patterns are generated by the above-mentioned manner.

Examples of RSAs derived by Eq. (43) across scales (the radius of the sampling region is $r_s = 0.05, 0.1, 0.5, 10, 100$ and 1000km, respectively) are shown in Fig. 7 associated with numerically obtained RSAs provided that the underlying individual distributions are the homogeneous Poisson or Thomas processes. Since numerical calculations with large geographic ranges are computationally expensive, we set $r_g = 10$km and this is suffice to check our analysis. Here, we examined the zero-truncated form of Eq. (43), since in practice we do not observe the event that 0 individual is found. To do this, we multiply each distribution by $(1 - P(X = 0))^{-1}$, where $P(X = 0)$ is $(1 - p)^\alpha$ for Eq. (36) and ${}_2F_1(\frac{1}{2}, \alpha; 1; -\frac{\min\{\nu(S),\nu(G)\}}{\beta})$ for Eq. (42), respectively. In general, the homogeneous Poisson process and Thomas processes show qualitatively similar curves except for Fig. 7b where deviations between two processes in the SAR appear (Fig. 4). As expected, when the spatial scales of sampling and geographic ranges are significantly different (i.e., Figs. 7a and b; $S \ll G$; the weighting term is $(r_g - r_s)/(r_g + r_s) < 0.02$) the analytical results show good agreement with numerical results, as Eq. (43) approaches the non-approximated RSA (i.e., Poisson-gamma distribution; Eq. 44). Outside this region, the effect of approximation appears (Figs. 7c-f), showing deviations from the numerical results especially for small $x$, but it still describes qualitative aspects of each RSA. The tail on small $x$ (e.g. Fig. 7f) disappears in the large limit of the sampling region. Fig. A.1 in Appendix is in such a situation, showing a left-skewed bell shape on a logarithmic scale with four different parameter sets. The effect of different values of $r_g$ (producing 20% or 40% larger area of $G$) is provided via Eq. (43) in Fig. A.2.

We also numerically performed RSAs when the radius of geographic range varies according to the exponential distribution with the parameter $\lambda_e = 0.1$ as in the above analysis (Fig. 8). We found consistent patterns with Fig. 7 when the sampling area is small (Fig. 8a-c). However, patterns are rather different especially for larger scales (Fig. 8e, f), and these also show an inconsistent pattern between the homogeneous Poisson process and Thomas process. This inconsistency between two individual distributions may be attributed to an effect of small scales: at large sampling scales, a number of species with small geographic range sizes are sampled due to a nature of the exponential distribution (monotonically decreasing function), and those showing clustering distribution patterns (Thomas process) can provide a larger number of individuals than species with the homogeneous Poisson process within a small geographic range even with the same intensity. These suggest that different assumptions of geographic range size, on top of sampling shape, causes a different shape of RSAs even though the same sampling area is applied. See also Fig. A.2.

# 4 Application: spatial scaling of $\beta$ diversity

$\beta$ diversity is an important concept in community ecology and conservation biology that describes variations of species compositions between multiple assemblies across spatial scales [13, 33]. The spatial variations inherently include the scaling effect of the size of concerned region (spatial extent) and its subregion (spatial grain; Fig. 9), and its scale effect and relationship with other macroecological properties is often of interest to community ecologists [13]. However, data are often not sufficiently available across scales to evaluate the beta-diversity the relationship across scales empirically. Further, as far as we know, there have been limited theoretical attempts to explore these scaling issues. Plotkin and Muller-Landau [34] discussed the effect of clustering of conspecific species on similarity between two subregions with combining a species abundance distribution patterns in a spatially implicit framework. Barton et al. [13] provided a potential relationship between spatial extent and multiple communities within each spatial grain, but it was based on a conceptual discussion.

Here, to demonstrate potential uses of the geometric model developed, we apply the model to the scaling issue of $\beta$ diversity between multiple communities in a spatial extent. For this purpose, our analysis is minimal and restricted to two situations that was discussed in the previous section. Namely, individual distributions are described by the homogeneous Poisson (random) or Thomas (clustering) process with its radius of the geographic range $\mathrm{E}[r_g] = 10^1$km, and the same parameter values are used as shown in Fig. 4. Since $\beta$ diversity depends on the number of spatial grains, we need to use a normalized $\beta$ diversity defined on $[0, 1]$ to make a meaningful comparison [33]. Here we adopt the approach developed in Jost [33] but there are several different (normalized) diversity indices that, e.g., assign different weights to each community (e.g., [35]). In Appendix D, we summarize scaling issues of $\beta$ diversity and the diversity index used in our analysis.

To see how the normalized $\beta$ diversity changes across spatial extents and spatial grains, we compute Eq. (A.11). We choose the range of spatial extent to cover three different phases in the SAR (Fig. 4) ($2^{-6}\mathrm{km}^2$ – $2^{16}\mathrm{km}^2$), and we divided the spatial extent into equal-

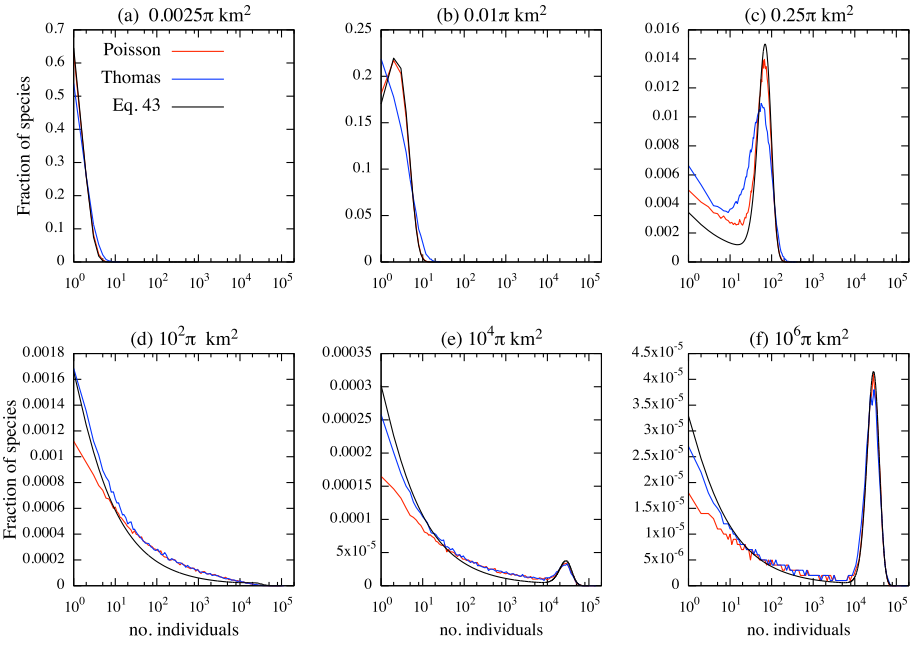


Figure 7: Relative species abundance across sampling regions with constant geographic range sizes across species ($r_g = 10$km). The radius of the sampling region in each panel (a)-(f) is $r_s = 0.05, 0.1, 0.5, 10, 100$ and 1000km, respectively, and these values are chosen so as to make transitions of the RSA tractable. Each panel shows three different curves regarding to the situations where underlying individual distributions are the homogeneous Poisson process and its theoretical form (Eq. 43), and the Thomas process. The theoretical form agrees well with its fully simulated values when the spatial scales of sampling and geographic range are significantly different (i.e., $S \ll G$ and $S \gg G$), in which Eq. (43) approaches to Eq. (44) where no approximation is made. Outside these regions, the theoretical curve still captures qualitative aspects of the numerical results, but with different degrees. For other parameters, the same values are used as in Fig. 4. See the main text for further explanations.

sized subregions with spatial grain size $2^{-8}$km$^2$ – $2^{s-2}$km$^2$, where $s$ determines the size of spatial extent, so that each scenario has minimum 4 subregions. To compute the normalized $\beta$ diversity across spatial extent and spatial grains, we used a single realization of point patterns by taking the following three steps: (i) define point patterns in the maximum spatial extent (e.g., defined on the plane $[0, 2^8] \times [0, 2^8]$); (ii) define a spatial extent (e.g., on the plane $[0, 2^{-1}] \times [0, 2^{-1}]$); and (iii) calculate the normalized $\beta$ diversity for each spatial grain with area $(2^{-8}, 2^{-6}, 2^{-4})$. In the step (ii), we started from the minimum spatial extent $2^{-8}$km$^2$, and repeated the step (ii) and (iii) until spatial extent reached the maximum extent $2^{16}$km$^2$. We eliminate the situation where no individual exists when the minimum spatial grain size is $2^{-8}$km$^2$ to make sure each scenario contains at least one individual. When all individuals are situated in one subregion, we define the normalized $\beta$ diversity as 0.

Fig. 10 shows the numerically calculated normalized $\beta$ diversity averaged over 500 simu-

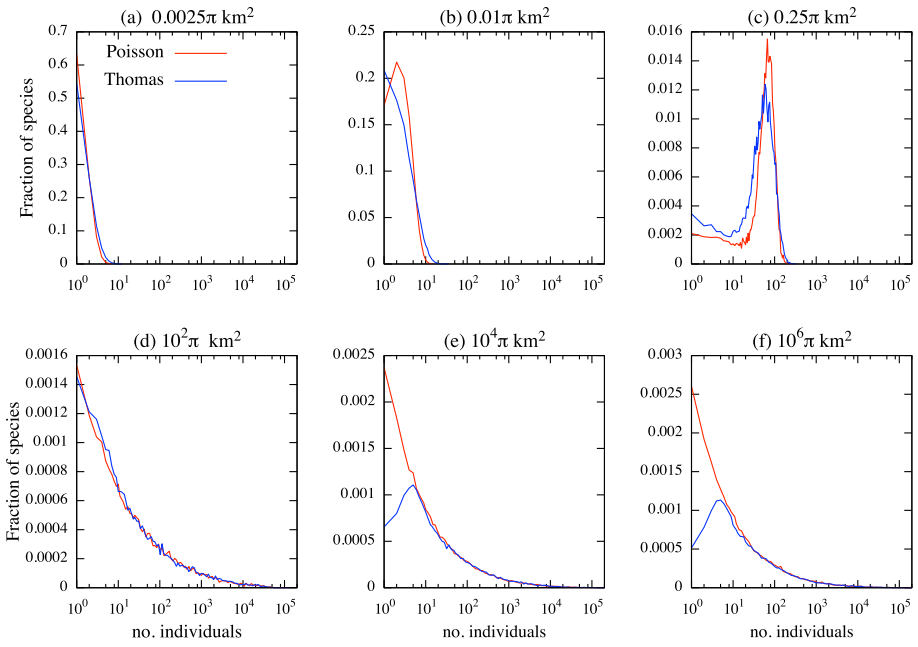


Figure 8: Relative species abundance across sampling regions with variable geographic ranges. The radius of the sampling region in each panel (a)-(f) is $r_s = 0.05, 0.1, 0.5, 10, 100$ and 1000km, respectively. Each panel shows two numerical results where underlying individual distributions are the homogeneous Poisson process and Thomas process. Radii of the geographic range follow the exponential distribution with the parameter $\lambda_e$=0.1 ($\mathrm{E}[r_g] = 10, \mathrm{Var}[r_g] = 100$). For other parameters, the same values are used as in Fig. 4. See the main text for further explanation.

lation trials, associated with the SAR curves of underlying processes. Overall, the underlying processes under concern do not cause a prominent qualitative effect. The top two figures show a heat map of the normalized $\beta$ diversity under the (a) homogeneous Poisson and (b) Thomas processes. In both panels, the normalized $\beta$ diversity shows a higher value as the spatial extent increases. Under this operation, the effect of applying different spatial grain becomes small, especially the SAR is in the third phase (the bottom two figures with fixed spatial grains $2^{-8}\mathrm{km}^2$, $2^{-6}\mathrm{km}^2$). Conversely, the normalized $\beta$ diversity shows a small value when spatial extent underlies in the left-side of the second phase of of the SAR, and its spatial grain is large.

# 5 Discussion

We develop a novel framework to derive macroecological patterns across scales by explicitly linking the distribution of individuals within ranges, the size of ranges, and the spatial extent of the sampling region. The model phenomenologically describes species and individual dis-

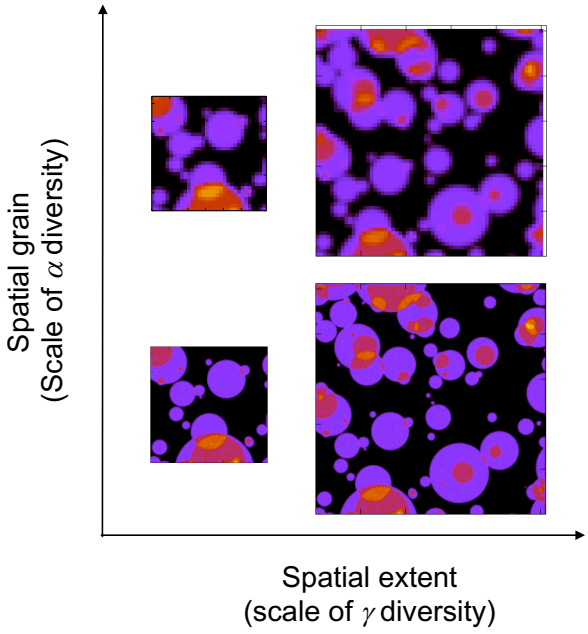


Figure 9: Schematic figure of spatial extent and spatial grain. Blighter colors indicate higher abundance of all the species within the spatial extent 128km×128km (left) and 256km×256km (right). Two examples of spatial grain are used: 1km×1km (bottom; fine) and 4km×4km (top; coarse).

tributions and in the derivations the SAR, EAR, and RSA, no preceding assumptions about these relationships are required. Rather, the model requires a single set of parameters to generate macroecological patterns across scales. Although the model does not explicitly assume specific biological mechanisms such as population or community dynamics, dispersal, and speciation, it still recovers several well-known macroecological patterns including the tri-phasic SAR and Fisher's logseries [2], the Poisson gamma distribution (negative binomial distribution) [29], Poisson lognormal distribution [30], and forms similar to a negatively skewed lognormal distribution [6] as RSAs. This finding does not imply that the biological mechanisms shaping biodiversity pattern are irrelevant or uninteresting. Rather, our theory demonstrates the minimum assumptions are sufficient to recover such ubiquitous ecological patterns by linking pattern in individual and species distribution to aggregate macroecological patterns. That said, it is clear the general forms of commonly-studied macroecological patterns investigated here are general features of biodiversity pattern emerging from the most basic assumptions, and are not indicative of specific ecological processes to the exclusion of others.

We presented that the tri-phasic SAR with its asymptotic slope 1 on a log-log plot is generally observed in the presented geometric model (Fig. 4a) under both random and clustered individual distribution patterns, and identical and non-identical ecological parameters across

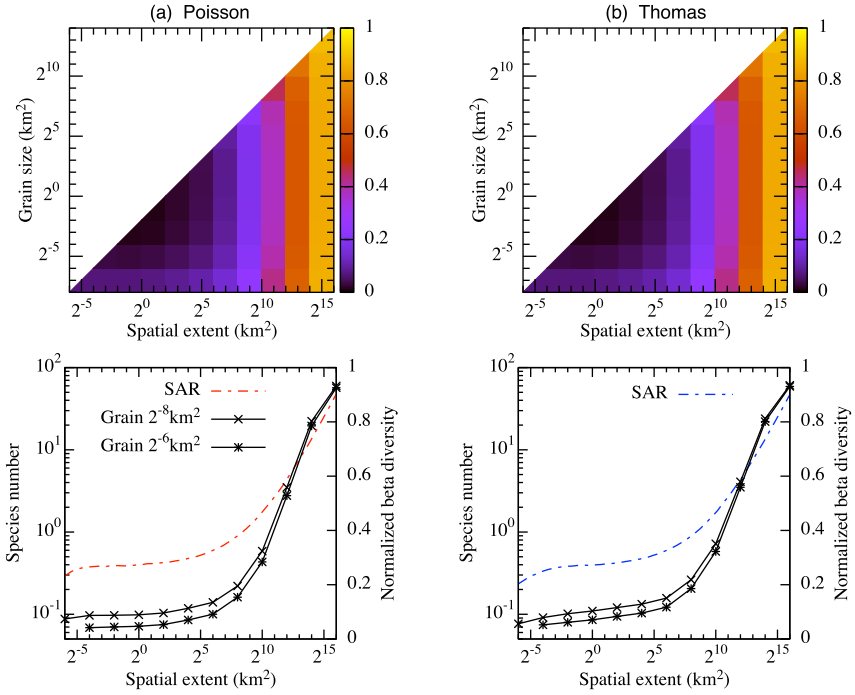


Figure 10: The normalized $\beta$ diversity averaged over 500 simulation trials, associated with the SAR curves of underlying (a) homogeneous Poisson processes (random) and (b) Thomas process (clustering). Top two figures is the heat map of the normalized $\beta$ diversity across spatial extent and spatial grain. Bottom two figures show some slices of the top figures, where fixed spatial grains ($2^{-8}$km$^2$, $2^{-6}$km$^2$) are applied while spatial extent varies. The same parameter values are used as in the results of the SARs (Fig. 4) under the homogeneous Poisson or Thomas process with a radius of the geographic range $\mathrm{E}[r_g] = 10^1$km.

species. Our results are in line with the findings of Grilli et al. [11] that the tri-phasic SAR is the outcome associated with two bending points in the SAR produced by local and large sampling area that scales the species number provided by a simple geometric realization. In particular, we showed that the third phase in the SAR appears around the average area of the species geographic ranges. This may correspond to the biological interpretations that sampling area exceeds correlation distance of biogeographic process [5,6]. Notably, we found a negligible difference in the SAR between the homogeneous situation (species parameters are identical) and a situation where only the individual intensity ($\lambda_i$) varies between species. However, this variation is necessary in our model to derive well-documented RSAs discussed above. In addition, we demonstrated that the spatial scaling of beta diversity can be considered in the context of the tri-phasic features of the SAR curve. Typically the normalized beta diversity increases with the spatial extent, and it shows largest value in the third phase

of the SAR.

In our framework, the EAR is calculated by explicitly taking geographic ranges into account. Under this definition, the EAR on a log-log plot has the same asymptotic slope 1 as the SAR. This result cannot be directly compared to previous studies [10, 11, 16] where the number of endemic species is calculated based on the probability of finding a species in a given region where the probability asymptotically approaches 0 as the sampling region approaches 0. On the other hand, we apply a more straightforward definition: the expected number of geographic ranges completely enclosed by the sampling region. Therefore, our definition gives 0 endemic species unless the scale of sampling region exceeds the scale of species geographic range. However, asymptotic behavior at small sampling scales can also occur in our definition by introducing variations in range size that rather small ranges allow to exist. The differences between the two definitions especially appear in small sampling scales. In fact, Grilli et al. [11] showed that the slope of the EAR converges to the slope of SAR at large scales, and this agrees with our discussion at large limit of sampling scales.

Importantly, we provide a single equation that unifies RSAs across sampling scales (Eqs. 14 and 43). The equation is composed of two pdfs where their weights are determined by scales of the sampling region and species geographic range. Specifically, one of the pdfs accounts effects of intra- and inter- species variations in an equal-sized region, and the other distribution accounts, on top of these variations, variations of sampled area of geographic range, $\nu(S \cap G)$. In the latter distribution, the effect of variations in sampled area of geographic range (i.e., variations of $\nu(S \cap G)$) is explicitly considered: given an identical range size, a species partly sampled ($S$ partly overlaps with $G$) shows smaller abundance than a species entirely sampled ($S$ completely overlaps with $G$). With an assumption of identical sampling region, this latter variation causes left-tailed RSAs as some species are sampled only small regions and others are fully sampled (Fig. 7c-f). Once the sampled region becomes sufficiently large compared to the geographic range, the number of partially sampled species becomes negligible and the left tail vanishes, resulting in the lognormal-like distribution (Fig. A.1), provided each species has sufficiently large abundance. Lognormal or lognormal-like distributions have been claimed as potential RSA (SAD) at global scales in previous work [36–38]. Rosindell et al. [38] incorporated a mode of speciation (protracted speciation) into Hubbell's neutral theory that counts species that a certain generations has existed, in contrast to the point (singleton) mutation as in the original neutral model [6]. This assumption may exclude species that have narrow geographic ranges and is prone to extinction, and be comparable to our assumption that all species has large enough geographic range to persist ($100\pi\text{km}^2$ in Fig. A.1). The assumption that the range sizes follow an exponential distribution may be comparable to the assumption of the point speciation, since it produces a number of extinction-prone species with narrow geographic ranges: it leads to RSAs that the majority of species are singletons when spatial structure is ignored (i.e., homogeneous Poisson process, Fig. 8), and RSAs with a peak at a small number of individuals when individuals tend to be aggregated.

More specifically, Eq. (43) is the weighted sum of the Poisson-gamma distribution and Poisson-gamma-arcsine distribution, and this single equation does not provide a single form

of pdf corresponding to the RSA, but generates multiple forms. Also, we showed that the above pattern may not be only one explanation of RSAs but can change with patterns of geographic ranges as well as the shape of the sampling region. These results provide a new insight into a longstanding discussion about a universal RSA pattern sampled in biological communities, where a number of authors have attempted to fit a single pdf to RSAs across scales (e.g., [8, 29]). Also, we emphasize that the shape of sampling region alters the expected RSA patterns and, hence, this information must be provided with its area to interpret and compare RSA patterns. In practice, we need further information of the geographic range of each species to test our prediction regarding the variation of overlapping regions. However, this information is not usually available [39], except for some species in several localities [40, 41], and an assumption of its probability distribution is required until we find general insight.

Another notable finding is that the prominent variations of different individual distribution patterns (random and clustering) occur only a limited spatial scales in the SAR. This occurs around the transition points between the first and second phases of the tri-phasic curve (Fig. 4), and agrees with the results obtained by Plotkin et al. [24], in which the authors examined the random-placement model and Thomas process to fit observed SAR patterns up to 50-ha in tropical forests. This finding suggests that some biological processes such as dispersal dominate community pattern formations within the scope of small scales. However, once the sampling area becomes large enough, e.g., significantly larger than the area covered by the dispersal kernel, each component of a cluster plays the same role as randomly placed individuals (Note that each parent location placed randomly in a generation of Thomas process, see Appendix B). Intuitively speaking, if two geometric patterns generated each by the homogeneous Poisson process or Thomas process are observed by the scope of such a large scale, it is no longer easy to distinguish these two patterns as long as the expected number of individuals is the same and it is sufficiently large. In addition to the SAR, some qualitative features are shared between the two distribution patterns in RSAs. The similar discussion may apply to RSA, but as we see in Fig. 8e and f, inconsistency of qualitative feature may also occur at large scales when stochasticity plays a role as discussed in Results. This property suggests that random individual distributions can be used to explore the macroecological patterns with these spatial scales, and it makes analysis significantly more accessible. Nonetheless, even without such detailed spatial information, our framework suggests that upscaling (downscaling) of RSAs is possible within and between small and large sampling scales as long as the situation of Fig. 2a and d hold. In these limits, the effect of spatial overlap between the sampling region and geographic range (Fig. 2c and d) is omitted and, hence, RSAs share a same pdf in these scales with different parameter values that is caused by changes in $\min\{\nu(S), \nu(G)\}$ (Eq. 14). A method to change the scale in RSAs that contains a single pdf was previously developed by Azaele et al. [12] where the parameters of a potential RSA (they applied a gamma distribution) vary with the sampling scale. In contrast, our approach explicitly evaluate when this scale change (with a single pdf) is feasible provided small/large limit of sampling region.

In this study, we developed a theoretical framework to derive macroecological patterns

across scales, and demonstrated its applicability using an example problem: understanding the scaling issues of $\beta$ diversity. Other promising applications are for biodiversity conservation and ecosystem management, where spatially integrated approaches such as spatial design of reserve networks [18, 42], estimation of biodiversity loss after habitat fragmentation [39], and scale dependence of management decision making [20, 43] have been widely discussed. Our framework may help contribute toward a theoretical basis to problems facing these fields. We minimized the assumptions of the model for reasons of parsimony and analytical tractability. Nonetheless, the general theory developed to derive the SAR, EAR, and RSA in an arbitrary situation can be flexibly extended and examined various ecological assumptions, at least numerically; for example, point processes provide frameworks to discuss environmental heterogeneity (e.g., habitat quality) and interaction of each individuals (e.g., [26]). Furthermore, one can use any mechanistic or phenomenological population/community model to generate a point field $f(\mathbf{x})$, and use the general theory here to examine the ramifications for emergent macroecological patterns. Such experiments would provide further insights into both utility of these analytical approaches and enhance quantitative understanding the macroecological patterns that are currently of wide interest to researchers.

# Acknowledgements

We would like to thank Y. Iwasa, T. Fung, A. Yamauchi, Y. Tachiki, A. Sasaki and W. Godsoe for their thoughtful comments. NT was supported by Grant-in-Aid for the Japan Society for the Promotion of Science (JSPS) Fellows. Financial support was provided by JSPS (No. 15K14607 to YK and No. 17K15180 to EPE) and Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers, JSPS (to YK). EPE and NT were additionally supported by subsidy funding to OIST.

# References

[1] S. A. Levin, The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture, Ecology 73 (6) (1992) 1943–1967.

[2] R. A. Fisher, A. S. Corbet, C. B. Williams, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, J. Anim. Ecol. 12 (1) (1943) 42. doi:10.2307/1411.
URL http://www.jstor.org/stable/1411?origin=crossref

[3] F. Preston, Time and space and the variation of species, Ecology 41 (4) (1960) 611–627.

[4] R. M. May, Patterns of Species Abundance and Diversity, in: M. J. Cody, J. M. Diamond (Eds.), Ecol. Evol. Communities, The Belknap Press of Harvard Univ. Press, Cambridge, MA and London, 1975, pp. 81–120.

[5] M. Rosenzweig, Species diversity in space and time, Cambridge University Press, Cambridge, UK, 1995.

[6] S. P. Hubbell, The Unified Neutral Theory of Biodiversity and Biogeography, Princeton University Press, Princeton, NJ, 2001.

[7] B. J. McGill, R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas, B. J. Enquist, J. L. Green, F. He, A. H. Hurlbert, A. E. Magurran, P. A. Marquet, B. A. Maurer, A. Ostling, C. U. Soykan, K. I. Ugland, E. P. White, Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework, Ecol. Lett. 10 (10) (2007) 995–1015.
URL http://doi.wiley.com/10.1111/j.1461-0248.2007.01094.x

[8] E. Baldridge, D. J. Harris, X. Xiao, E. P. White, An extensive comparison of species-abundance distribution models, PeerJ 4 (2016) e2823. doi:10.7717/peerj.2823.
URL https://peerj.com/articles/2823

[9] D. Storch, A. L. Šizling, J. Reif, J. Polechová, E. Šizlingová, K. J. Gaston, The quest for a null model for macroecological patterns: Geometry of species distributions at multiple spatial scales (2008). doi:10.1111/j.1461-0248.2008.01206.x.

[10] D. Storch, P. Keil, W. Jetz, Universal speciesarea and endemicsarea relationships at continental scales, Nature 488 (7409) (2012) 78–81. doi:10.1038/nature11226.
URL http://www.nature.com/doifinder/10.1038/nature11226

[11] J. Grilli, S. Azaele, J. R. Banavar, A. Maritan, Spatial aggregation and the species-area relationship across scales, J. Theor. Biol. 313 (2012) 87–97. doi:10.1016/j.jtbi.2012.07.030.

[12] S. Azaele, A. Maritan, S. J. Cornell, S. Suweis, J. R. Banavar, D. Gabriel, W. E. Kunin, Towards a unified descriptive theory for spatial ecology: Predicting biodiversity patterns across spatial scales, Methods Ecol. Evol. 6 (3) (2015) 324–332. doi:10.1111/2041-210X.12319.

[13] P. S. Barton, S. A. Cunningham, A. D. Manning, H. Gibb, D. B. Lindenmayer, R. K. Didham, The spatial scaling of beta diversity, Glob. Ecol. Biogeogr. 22 (6) (2013) 639–647. doi:10.1111/geb.12031.

[14] S. L. Pimm, C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, J. O. Sexton, The biodiversity of species and their rates of extinction, distribution, and protection, Science 344 (6187) (2014) 1246752–1246752. doi:10.1126/science.1246752.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.1246752

[15] J. E. Watson, D. F. Shanahan, M. Di Marco, J. Allan, W. F. Laurance, E. W. Sanderson, B. Mackey, O. Venter, Catastrophic Declines in Wilderness Areas Undermine Global

Environment Targets, Curr. Biol. 26 (21) (2016) 2929–2934. doi:10.1016/j.cub.2016.08.049.

[16] F. He, S. P. Hubbell, Speciesarea relationships always overestimate extinction rates from habitat loss, Nature 473 (7347) (2011) 368–371.

[17] T. J. Matthews, R. J. Whittaker, Fitting and comparing competing models of the species abundance distribution: assessment and prospect, Front. Biogeogr. 6 (2) (2014) 67–82. arXiv:0608246v3, doi:10.5811/westjem.2011.5.6700.

[18] H. Possingham, I. Ball, S. Andelman, Mathematical methods for identifying representative reserve networks, in: S. Ferson, M. Burgman (Eds.), Quant. methods Conserv. Biol., Springer-Verlag New York, New York, USA, 2000, pp. 291–305. doi:10.1007/0-387-22648-6_17.

[19] C. R. Margules, R. L. Pressey, Systematic conservation planning., Nature 405 (6783) (2000) 243–53. doi:10.1038/35012251.

[20] N. Takashina, M. Baskett, Exploring the effect of the spatial scale of fishery management, J. Theor. Biol. 390 (2016) 14–22. doi:10.1016/j.jtbi.2015.11.005.

[21] N. Takashina, M. Beger, B. Kusumoto, S. Rathnayake, H. Possingham, A theory for ecological survey methods to map individual distributions, Theor. Ecol. 11 (2017) 213–223. doi:https://doi.org/10.1007/s12080-017-0359-7.

[22] N. Takashina, B. Kusumoto, M. Beger, S. Rathnayake, H. P. Possingham, Spatially explicit approach to estimation of total population abundance in field surveys, J. Theor. Biol. 453 (2018) 88–95. doi:10.1016/j.jtbi.2018.05.013.

[23] A. P. Allen, E. P. White, Effects of range size on species-area relationships, Evol. Ecol. Res. 5 (4) (2003) 493–499. doi:10.1073/pnas.0500424102.

[24] J. B. Plotkin, M. D. Potts, N. Leslie, N. Manokaran, J. Lafrankie, P. S. Ashton, Species-area curves, spatial aggregation, and habitat specialization in tropical forests., J. Theor. Biol. 207 (1) (2000) 81–99. doi:10.1006/jtbi.2000.2158.

[25] J. Illian, A. Penttinen, H. Stoyan, D. Stoyan, Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley & Sons, Ltd, Chichester, UK., 2008. doi:10.1002/9780470725160.

[26] S. N. Chiu, D. Stoyan, W. S. Kendall, J. Mecke, Stochastic Geometry and Its Applications, John Wiley & Sons, New York, 2013.

[27] S. Azaele, S. J. Cornell, W. E. Kunin, Downscaling species occupancy from coarse spatial scales, Ecol. Appl. 22 (3) (2012) 1004–1014. doi:10.1890/11-0536.1.

[28] D. Karlis, E. Xekalaki, Mixed poisson distributions, Int. Stat. Rev. 73 (1) (2005) 35–58. doi:10.1111/j.1751-5823.2005.tb00250.x.

[29] S. R. Connolly, M. A. MacNeil, M. J. Caley, N. Knowlton, E. Cripps, M. Hisano, L. M. Thibaut, B. D. Bhattacharya, L. Benedetti-Cecchi, R. E. Brainard, A. Brandt, F. Bulleri, K. E. Ellingsen, S. Kaiser, I. Kroncke, K. Linse, E. Maggi, T. D. O'Hara, L. Plaisance, G. C. B. Poore, S. K. Sarkar, K. K. Satpathy, U. Schuckel, A. Williams, R. S. Wilson, Commonness and rarity in the marine biosphere, Proc. Natl. Acad. Sci. 111 (23) (2014) 8524–8529. doi:10.1073/pnas.1406664111.
URL http://www.pnas.org/cgi/doi/10.1073/pnas.1406664111

[30] M. G. Bulmer, On Fitting the Poisson Lognormal Distribution to Species-Abundance Data, Biometrics 30 (1) (1974) 101. doi:10.2307/2529621.
URL http://www.jstor.org/stable/2529621?origin=crossref

[31] B. Dennis, G. P. Patil, Applications in ecology, in: E. L. Crow, K. Shimizu (Eds.), Lognormal Distrib. theory Appl., Markel Dekker, Inc, New York, USA, 1988, pp. 303–330.

[32] Arcsine distribution. Encyclopedia of Mathematics. URL: http://www.encyclopediaofmath.org/index.php?title=Arcsine_distribution&oldid=33530 Data Accessed: Nov 26 2017.
URL http://www.encyclopediaofmath.org/index.php?title=Arcsine{_}distribution{&}oldid=33530

[33] L. Jost, Partitioning diversity into independent alpha and beta components, Ecology 88 (10) (2007) 2427–2439. doi:10.1890/06-1736.1.

[34] J. B. Plotkin, H. C. Muller-Landau, Sampling the species composition of a landscape, Ecology 83 (2002) 3344–3356. doi:10.1890/0012-9658(2002)083[3344:STSCOA]2.0.CO;2.

[35] C. H. Chiu, L. Jost, A. Chao, Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers, Ecol. Monogr. 84 (1) (2014) 21–44. doi:10.1890/12-0960.1.

[36] R. D. Gregory, Abundance patterns of European breeding birds, Ecography (Cop.). 23 (2) (2000) 201–208. doi:10.1111/j.1600-0587.2000.tb00276.x.

[37] B. J. McGill, Does Mother Nature really prefer rare species or are log-left-skewed SADs a sampling artefact?, Ecol. Lett. 6 (8) (2003) 766–773. doi:10.1046/j.1461-0248.2003.00491.x.

[38] J. Rosindell, S. J. Cornell, S. P. Hubbell, R. S. Etienne, Protracted speciation revitalizes the neutral theory of biodiversity, Ecol. Lett. 13 (6) (2010) 716–727. doi:10.1111/j.1461-0248.2010.01463.x.

[39] R. A. Chisholm, F. Lim, Y. S. Yeoh, W. W. Seah, R. Condit, J. Rosindell, Speciesarea relationships and biodiversity loss in fragmented landscapes (2018). doi:10.1111/ele.12943.

[40] K. J. Gaston, Species-range size distributions: Products of speciation, extinction and transformation, Philos. Trans. R. Soc. B Biol. Sci. 353 (1366) (1998) 219–230. doi:10.1098/rstb.1998.0204.

[41] C. D. L. Orme, R. G. Davies, V. A. Olson, G. H. Thomas, T. S. Ding, P. C. Rasmussen, R. S. Ridgely, A. J. Stattersfield, P. M. Bennett, I. P. Owens, T. M. Blackburn, K. J. Gaston, Global patterns of geographic range size in birds, PLoS Biol. 4 (7) (2006) 1276–1283. doi:10.1371/journal.pbio.0040208.

[42] M. Beger, J. McGowan, E. A. Treml, A. L. Green, A. T. White, N. H. Wolff, C. J. Klein, P. J. Mumby, H. P. Possingham, Integrating regional conservation priorities for multiple objectives into national policy, Nat. Commun. 6. doi:10.1038/ncomms9208.

[43] M. Bode, J. N. Sanchirico, P. R. Armsworth, Returns from matching management resolution to ecological variation in a coral reef fishery, Proc. R. Soc. B Biol. Sci. 283 (2016) 1826. doi:10.1098/rspb.2015.2828.

[44] B. D. Coleman, On random placement and species-area relations, Math. Biosci. 54 (3-4) (1981) 191–215. doi:10.1016/0025-5564(81)90086-9.

[45] M. J. Anderson, T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. Kraft, J. C. Stegen, N. G. Swenson, Navigating the multiple meanings of $\beta$ diversity: A roadmap for the practicing ecologist, Ecol. Lett. 14 (1) (2011) 19–28. doi:10.1111/j.1461-0248.2010.01552.x.

[46] L. Jost, Entropy and diversity, Oikos 113 (2) (2006) 363–375. doi:10.1111/j.2006.0030-1299.14714.x.

[47] A. Chao, C. H. Chiu, T. C. Hsieh, B. D. Inouye, Proposing a resolution to debates on diversity partitioning, Ecology 93 (9) (2012) 2037–2051. doi:10.1890/11-1817.1.

[48] A. Chao, C. H. Chiu, Bridging the variance and diversity decomposition approaches to beta diversity via similarity and differentiation measures, Methods Ecol. Evol. 7 (8) (2016) 919–928. doi:10.1111/2041-210X.12551.

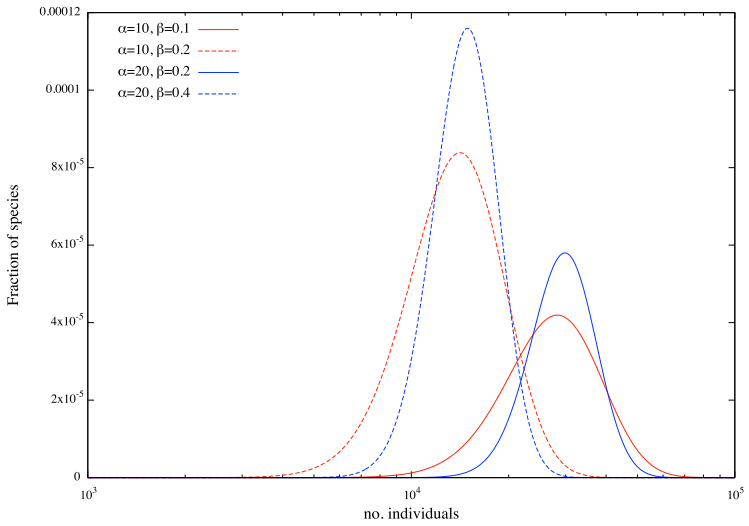


Figure A.1: Relative species abundance in the large limit of the sampling region (Eq. 44) with four different parameter sets $(\alpha, \beta)$ of the gamma distribution. $\alpha = 10$, $\beta = 0.1$ is used in Fig. 7. For other parameters, the same values are used as in Fig. 7.

# Appendix

## A $S_{out}$ and $S_{core}$

For each set of the sampling region $S$ and species geographic range $G$, we uniquely otain $S_{out}$ and $S_{core}$, which are used in calculation of SAR, EAE, and RSA. See the main text for specific calculations.

$S_{out}$ is the region that the probability of overlapping the sampling region and the geographic range is not zero, $P(\nu(S \cap G) = \phi) \neq 0$, if the geometric center of the geographic range falls in this region, $\mathbf{x}_c^G \in S_{out}$. $S_{out}$ includes $S$ as a subset, $S \subset S_{out}$, in the region that the shortest distance between the boundary of sampling region and $S_{out}$ equals to the largest distance of the geometric center of $G$ and its boundary: $L_G = \inf\{\|\mathbf{x} - \mathbf{x}'\|; \mathbf{x} \in \partial S, \mathbf{x}' \in \partial S_{out}\}$, where $\partial A$ denotes the boundary of $A$.

$S_{core}$ is the region, provided the relative size of geometries $S > G$ $(S < G)$, if the center of a geographic range falls in the region, then the region $G$ $(S)$ is entirely included in $S$ $(G)$: Namely $P(G \subset S) = 1$ $(P(S \subset G) = 1)$ provided $\mathbf{x}_c^G \in S_{core}$. $S_{core}$ is the closed region and typically a subset of the sampling region $S_{core} \subset S$, in which any points on this boundary, $\mathbf{x} \in \partial S_{core}$ holds the following relationship: when $S > G$, $\inf\{\|\mathbf{x} - \mathbf{x}'\|; \mathbf{x} \in \partial S, \mathbf{x}' \in \partial S_{core}\} = \sup\{\|\mathbf{x}_c^G - \mathbf{x}\|; \mathbf{x} \in \partial G\}$; when $S < G$, $\inf\{\|\mathbf{x}_c^G - \mathbf{x}\|; \mathbf{x} \in \partial G\} = \sup\{\|\mathbf{x} - \mathbf{x}'\|; \mathbf{x} \in \partial S, \mathbf{x}' \in \partial S_{core}\}$. Each region is determined by the minimum distance $l_S$, $l_G$ and maximum distance $L_S$, $L_G$ between the sampling center and its boundary. For the calculation of the EAR, only the case $S > G$ is used, but both $S > G$ and $S < G$ are used for the calculation of the RSA.

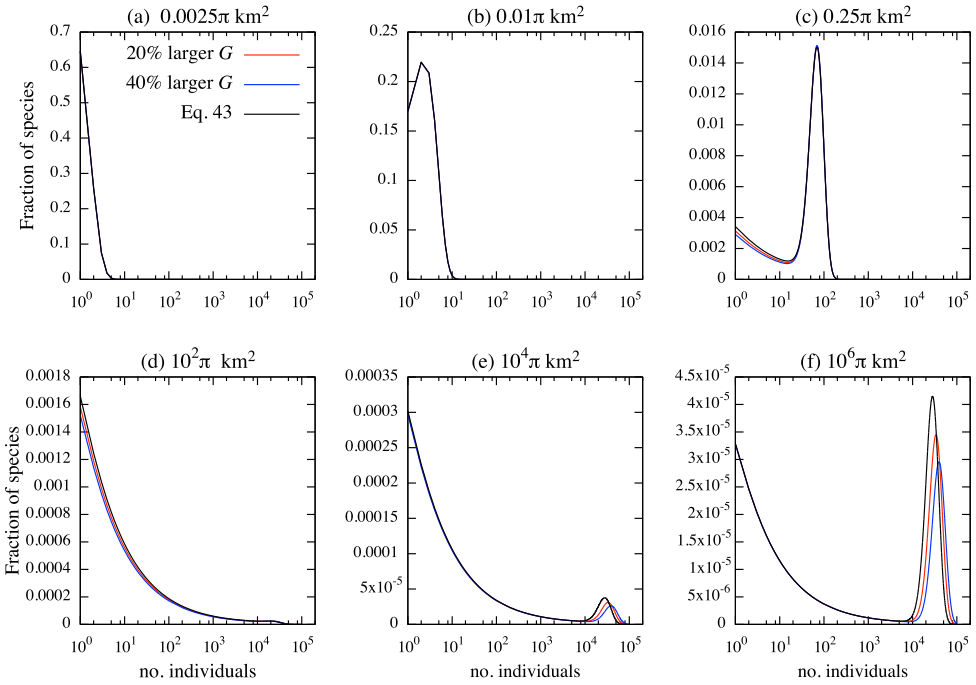


Figure A.2: Relative species abundance across sampling regions derived by Eq. (43). The radius of the sampling region in each panel (a)-(f) is $r_s = 0.05, 0.1, 0.5, 10, 100$ and 1000km, respectively. Three lines in each panel corresponding to the base value $r_g = 10$km (as in Fig. 7), 10.95km (red line; 20% larger than the base area $\nu(G) = 10^2\pi\text{km}^2$), and 11.83km (blue line; 40% larger than the base area). For other parameters, the same values are used as in Fig. 7.

# B Generating species distribution patterns

To examine the general theory of the geometric approach developed above, we need to introduce a point field $f(\mathbf{x})$ defined above that holds information of individual distributions of all species. For this purpose, we make use of point processes [25, 26], a set of spatially explicit stochastic models that generate various point distribution patterns such as random and clustering patterns. One of the advantages of these models is that point processes are amenable to mathematical analysis and easy to implement numerical simulations. In addition, there are a number of applications to ecological studies [11, 21, 22, 24, 27], and models can provide consistent patterns with observed SARs or a population occupancy probability regardless of simple assumptions [11, 24, 27].

Here, we examine the theory developed using two individual distribution patterns: random and clustering distribution patterns, described by the homogeneous Poisson process and the Thomas process, respectively. The former is often used to develop a simplest possible model (e.g., [44]), or to see how the simple assumption deviates from more biologically reliable models (e.g., [16, 21, 24]). As we will see below, deviations of these two distributions in

the SAR and RSA are relatively small to make qualitative discussions and these appear in small scales.

The homogeneous Poisson process and Thomas process are defined using the intensity and intensity measure of species $i$, defined above (Eqs. 1 and 2). Also, as mentioned above, individuals of species $i$ is restricted within its range $G_i$. For example, if species $i$ is distributed randomly in region $A$, the probability to find $x$ individuals within the region with area $\nu(A)$ follows the Poisson distribution with average $\mu_i(A) = \lambda_i \nu(A)$:

$$P(X = x) = \frac{\mu_i(A)^x}{x!} e^{-\mu_i(A)}. \tag{A.1}$$

This process generates the homogeneous Poisson process.

On the other hand, the Thomas process is described by the following three steps:

1. Parents of species $i$ are randomly placed according to the homogeneous Poisson process with intensity $\lambda_i^p$.
2. Each parent of species $i$ produces a random discrete number $c_i$ of daughters, realized independently and identically.
3. The daughters are scattered around their parents independently with an isotropic bivariate Gaussian distribution with the variance $\sigma_i^2$, and all the parents are removed in the realized point pattern.

The intensity of individuals of species $i$ for the Thomas process is [26]

$$\lambda_{th,i} = \bar{c}_i \lambda_i^p, \tag{A.2}$$

where, $\bar{c}_i$ is the average number of daughters per parent. To guarantee a consistent number of total expected individuals given area between the two processes, we set $\lambda_i^p$ and $\bar{c}_i$ as

$$\lambda_{th,i} = \bar{c}_i \lambda_i^p = \lambda_i. \tag{A.3}$$

We also assume that the number of daughters per parents $c_i$ follows the Poisson distribution with the average number $\bar{c}_i$.

By superimposing distributions of all species generated either by the homogeneous Poisson process or the Thomas process, we obtain individual distributions of all species in the whole ecosystem. With these specific individual distribution patterns, we can examine calculations presented in General Theory to obtain the SAR, EAR, and RSA across scales.

# C Derivation of the Poisson-gamma-arcsine distribution and Poisson-lognormal-arcsine distribution

## C.1 Poisson-gamma-arcsine distribution (Eq. 42)

Now, using Proposition 3 and the $r$th moment of the gamma distribution $f(\lambda) = \beta^\alpha / \Gamma(\alpha) \lambda^{\alpha-1} e^{-\lambda\beta}$

$$\mu_r^{gam} = \frac{\Gamma(r+\alpha)}{\Gamma(\alpha)} \frac{1}{\beta^r}, \tag{A.4}$$

we calculate the probability of mixing distribution Eq. (A.13) with properties of the beta function $\mathrm{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$ and the Gamma function $\Gamma(1/2)=\sqrt{\pi}$:

$$\begin{aligned}
P(X=x) &= \frac{1}{x!}\sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\frac{\Gamma(r+x+\alpha)}{\Gamma(\alpha)}\frac{1}{\beta^{x+r}}\frac{\nu(A)^{r+x}}{\pi}\mathrm{B}(\frac{1}{2}+r+x,\frac{1}{2}),\\
&= \frac{1}{x!}\left(\frac{\nu(A)}{\beta}\right)^x\frac{1}{\sqrt{\pi}\Gamma(\alpha)}\sum_{r=0}^{\infty}\frac{(-\nu(A)/\beta)^r}{r!}\frac{\Gamma(r+x+\alpha)\Gamma(\frac{1}{2}+r+x)}{\Gamma(1+r+x)},\\
&= \frac{1}{x!}\left(\frac{\nu(A)}{\beta}\right)^x\frac{1}{\sqrt{\pi}\Gamma(\alpha)}\frac{\Gamma(x+\alpha)\Gamma(\frac{1}{2}+x)}{\Gamma(1+x)}\,{}_2F_1(\frac{1}{2}+x,x+\alpha;1+x;-\nu(A)/\beta),\\
&= a\begin{pmatrix}x+\alpha-1\\ x\end{pmatrix}p^x(1-p)^\alpha\,{}_2F_1(\frac{1}{2}+x,x+\alpha;1+x;-\nu(A)/\beta), \qquad \text{(A.5)}
\end{aligned}$$

where, $a$ is $\Gamma(1/2+x)(\sqrt{\pi}\Gamma(x+1))^{-1}(1-p)^{-\alpha-x}$ and is simplified for $x\geq 1$ as $x^{-1}\mathrm{B}(1/2,x)^{-1}(1-p)^{-\alpha-x}$, the second to forth factors correspond to the negative binomial distribution with $p=\nu(A)/(\nu(A)+\beta)$, and ${}_2F_1$ is the Gauss hypergeometric function. $\alpha$ and $\beta$ are the shape and rate parameters of the gamma distribution. Note the negative binomial distribution here is equivalent to Eq. (36).

## C.2 Poisson-lognormal-arcsine distribution

By substituting the $r$th moment about the origin of the arcsine distribution Eq. (A.15), and the Poisson-lognormal distribution $\mu_r^{pl}=\exp(r\mu+r^2\sigma^2/2)$ [30] into Eq. (33), we obtain the RSA of an arbitrary form

$$P(X=x)=\frac{\min\{\nu(S),\nu(G)\}^x}{x!\sqrt{\pi}}\sum_{r=0}^{\infty}\frac{(-\min\{\nu(S),\nu(G)\})^r}{r!}\frac{\Gamma(\frac{1}{2}+r+x)}{\Gamma(1+r+x)}e^{(x+r)\mu+\frac{(x+r)^2}{2}\sigma^2}, \qquad \text{(A.6)}$$

where as in the case of the Poisson-gamma-arcsine distribution, we call this pdf the Poisson-lognormal-arcsine distribution. Using Eqs. (38), (40), and (A.6), we obtain the full form of the RSA across scales

$$\begin{aligned}
P(X=x) = {}& \frac{\max\{r_g-r_s,r_s-r_g\}}{r_g+r_s}\frac{1}{x!\sqrt{2\pi\sigma^2}}\int_0^\infty\lambda_i^{x-1}e^{-\lambda_i-\frac{(\log\lambda_i-\mu)^2}{2\sigma^2}}d\lambda_i+ \qquad \text{(A.7)}\\
&\frac{\min\{2r_s,2r_g\}}{r_g+r_s}\frac{\min\{\nu(S),\nu(G)\}^x}{x!\sqrt{\pi}}\sum_{r=0}^{\infty}\frac{(-\min\{\nu(S),\nu(G)\})^r}{r!}\frac{\Gamma(\frac{1}{2}+r+x)}{\Gamma(1+r+x)}e^{(x+r)\mu+\frac{(x+r)^2}{2}\sigma^2}.
\end{aligned}$$

Yet this has a still intricate form, Eq. (A.7) becomes the Poisson-lognormal distribution in the limits of $S\ll G$ and $S\gg G$ as above

$$P(X=x)=\frac{1}{x!\sqrt{2\pi\sigma^2}}\int_0^\infty\lambda_i'^{x-1}e^{-\lambda_i'-\frac{(\log\lambda_i'-\mu')^2}{2\sigma^2}}d\lambda_i', \qquad \text{(A.8)}$$

where, by setting $\mu=\mu'-\log(\min\{\nu(S),\nu(G)\})$ and $\min\{\nu(S),\nu(G)\}\lambda_i=\lambda_i'$, we recover the form of Eq. (38). Therefore, if the parameter $\lambda_i$ follows the Log-normal distribution and sampling region is much smaller than the geographic range, we expect to observe an RSA curve that follows Poisson-lognormal distribution.

# D Scaling issue of $\beta$ diversity

## D.1 Spatial grain and spatial extent

To quantify $\beta$ diversity it requires us to define two spatial scales; *spatial extent* and *spatial grain* [13] Fig. 9. Spatial extent is the scope of our observation, and spatial grain is the unit of sampling within the extent. Once we define the spatial extent, spatial grain, and point patterns, we obtain the matrix $P$: provided that there are $s$ species in the spatial extent, and arbitrarily dividing the community into $n$ assemblies, the occurrence probability of species $i$ in community $j$, $p_{ij}$ ($\sum_i p_{ij} = 1$), creates the following matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{s1} & \cdots & p_{sn} \end{pmatrix}. \tag{A.9}$$

Changing the size of spatial extent and/or spatial grain changes the size of the matrix $P$, since a smaller spatial extent may hold a fewer number of species, and a fine spatial grain increases the number of sampling patches. Therefore, if we change either one or both of these scales, the matrix $P$ is changed into another matrix $P'$. We describe this operation as $P \mapsto P'$. For example, if Eq. (A.9) is the $s \times n$ matrix with equal-sized patches and $n$ is an even number, and if its spatial grain is doubled, the operation $P \mapsto P'$ gives a $s \times n/2$ matrix.

In practice, each patch has a different significance on a diversity index, and this is described by the weight vector $\mathbf{w} = (w_1, w_2, \cdots, w_n)$. We can also define the same operation for the weight vector associated with a scale change.

## D.2 Diversity indices

Statistical discussions over reliable diversity metrics have led to a number of definitions in the literature [45], and bridges between different definitions has been actively discussed [33, 35, 46–48]. Here we adapt the definition of Jost [33]. Jost [33] showed that when weights in diversity indices are unequal, the only meaningful diversity index the Shannon measure since this only satisfies five requisite conditions for diversity indices. The Shannon measure is described by

$${}^1D_\alpha = \exp\{-w_1 \sum_i^s p_{i1} \log(p_{i1}) - w_2 \sum_i^s p_{i2} \log(p_{i2}) \cdots - w_n \sum_i^s p_{in} \log(p_{in})\}, \tag{A.10a}$$

$${}^1D_\gamma = \exp\{-\sum_i^s (w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in}) \log(w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in})\}, \tag{A.10b}$$

$${}^1D_\beta = \frac{{}^1D_\gamma}{{}^1D_\alpha}, \tag{A.10c}$$

In order to discuss the effect of a scale change of the spatial grain consistently, we further need a condition for independence of a choice of weights on the gamma diversity Eq. (A.10b). That is to say, scale changes of the spatial grain, provided a consistent spatial extent, should not affect the gamma diversity. It is easy to see that this condition is satisfied if we define the weight vector by population abundance of each patch, $w_j = \sum_i N_{ij} / \sum_{i,j} N_{ij}$, where $N_{ij}$ is the abundance of species $i$ in patch $j$. Furthermore, this is the only choice to hold this condition (Theorem 1 in Appendix E).

Since $\beta$ diversity depends on the number of spatial grains, we require normalizing the $\beta$ diversity onto $[0, 1]$ to make a meaningful comparison of spatial variations [33]. When all the weighting terms are not identical, the regional homogeneity measure $(1/{}^1D_\beta - 1/{}^1D_w)/(1 - 1/{}^1D_w)$ [33] may be used for the normalized measure of $\beta$ diversity, where ${}^1D_w$ is the Shannon measure of weighting term ${}^1D_w = \exp(-\sum_j w_j \log(w_j))$. We use its complement

$$1 - \frac{1/{}^1D_\beta - 1/{}^1D_w}{1 - 1/{}^1D_w}, \tag{A.11}$$

as a relative inhomogeneity measure. This measures is 0 if all the communities are identical and 1 if all the communities are distinct.

# E Proofs of propositions and theorem

Here we describe some propositions which are used in the main text and Appendix D.2. In the main text, the description $\nu(A)$ used below is replaced by $\min\{\nu(S), \nu(G)\}$.

**Proposition 1.** *Provided that the moments of the mixing distribution in a mixed Poisson model exist, the probability function of the mixture distribution can be written as*

$$P(X = x) = \frac{1}{x!} \sum_{r=0}^{\infty} \nu(A)^{x+r} \frac{(-1)^r}{r!} \mu_{x+r}(\lambda), \tag{A.12}$$

*where, $\mu_r(\lambda)$ are the $r$th moment of $\lambda$ about the origin.*

*Proof.* The proof is straightforward from the definition and similar result is found in [28]:

$$\begin{aligned}
P(X = x) &= \int_0^\infty \frac{(\lambda\nu(A))^x}{x!} e^{-\lambda\nu(A)} g(\lambda) d\lambda, \\
&= \frac{1}{x!} \int_0^\infty \left( \sum_{r=0}^{\infty} (-1)^r \frac{\lambda^r \nu(A)^r}{r!} \right) \lambda^x \nu(A)^x g(\lambda) d\lambda, \\
&= \frac{1}{x!} \sum_{r=0}^{\infty} \nu(A)^{x+r} \frac{(-1)^r}{r!} \int_0^\infty \lambda^{x+r} g(\lambda) d\lambda, \\
&= \frac{1}{x!} \sum_{r=0}^{\infty} \nu(A)^{x+r} \frac{(-1)^r}{r!} \mu_{x+r}(\lambda).
\end{aligned}$$

□

**Proposition 2.** *Provided that the moments of the mixing distribution in a mixed Poisson model exist, the probability function of the mixture distribution can be written as*

$$P(X=x)=\frac{1}{x!}\sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\mu_{x+r}(\lambda)\mu_{x+r}(\nu(A)), \tag{A.13}$$

*where, $\mu_r(\lambda)$ and $\mu_r(\nu(A))$ are the $r$th moment of $\lambda$ and $\nu(A)$ about the origin, respectively.*

*Proof.* The proof is just an extension of the Proposition 1.

$$\begin{aligned}
P(X=x) &= \int_0^{\infty}\int_0^{\infty}\frac{(\lambda\nu(A))^x}{x!}e^{-\lambda\nu(A)}g(\lambda)h(\nu(A))d\lambda d\nu(A),\\
&= \frac{1}{x!}\int_0^{\infty}\int_0^{\infty}\sum_{r=0}^{\infty}(-1)^r\left(\frac{\lambda^r\nu(A)^r}{r!}\right)\lambda^x g(\lambda)\nu(A)^x h(\nu(A))d\lambda d\nu(A),\\
&= \frac{1}{x!}\sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\int_0^{\infty}\lambda^{x+r}g(\lambda)d\lambda\int_0^{\infty}\nu(A)^{x+r}h(\nu(A))d\nu(A),\\
&= \frac{1}{x!}\sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\mu_{x+r}(\lambda)\mu_{x+r}(\nu(A)).
\end{aligned}$$

□

**Proposition 3.** *The $r$th moment of the arcsine distribution with support $x\in[0,\nu(A)]$*

$$f(x)=\frac{1}{\pi\sqrt{x(\nu(A)-x)}}, \tag{A.14}$$

*is described by*

$$\mu_r^{arc}=\frac{\nu(A)^r}{\pi}\mathrm{B}(\frac{1}{2}+r,\frac{1}{2}), \tag{A.15}$$

*where, $\mathrm{B}(x,y)=\int_0^1 u^{x-1}(1-u)^{y-1}du$ is the beta function.*

*Proof.* Using the substitution $w=x/\nu(A)$,

$$\begin{aligned}
\mu_r &= \frac{1}{\pi}\int_0^{\nu(A)}x^r x^{-\frac{1}{2}}(\nu(A)-x)^{-\frac{1}{2}}dx,\\
&= \frac{\nu(A)^r}{\pi}\int_0^1 w^{r+\frac{1}{2}-1}(1-w)^{\frac{1}{2}-1}dw,\\
&= \frac{\nu(A)^r}{\pi}\mathrm{B}(\frac{1}{2}+r,\frac{1}{2}).
\end{aligned}$$

□

**Threorem 1.** *The gamma diversity defined by Eq. (A.10b) that is independent of the choice of the number of patches or its size is uniquely determined, and it is when the weight vector is proportional to the population abundance of each patch.*

*Proof.* Let us assume there exists a vector $\mathbf{w}' = (w'_1, w'_2, \cdots, w'_n)$ that satisfies ${}^1D_\gamma^{\mathbf{w}} = {}^1D_\gamma^{\mathbf{w}'}$, where superscripts $\mathbf{w}$ and $\mathbf{w}'$ indicate weight vector used. By the assumption, we have $\sum_i^s (w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in}) \log(w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in}) = \sum_i^s (w'_1 p_{i1} + w'_2 p_{i2} + \cdots + w'_n p_{in}) \log(w'_1 p_{i1} + w'_2 p_{i2} + \cdots + w'_n p_{in})$ and arranging this expression, it becomes $\sum_{i,j} p_{ij} \{ w_j \log(w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in}) - w'_j \log(w'_1 p_{i1} + w'_2 p_{i2} + \cdots + w'_n p_{in}) \} = 0$. Since $p_{ij}$ may or may not be zero we must have $w_j \log(w_1 p_{i1} + w_2 p_{i2} + \cdots + w_n p_{in}) = w'_j \log(w'_1 p_{i1} + w'_2 p_{i2} + \cdots + w'_n p_{in})$ for all $i$ and $j$. This is clearly $w_j = w'_j$. □