Jaccard dissimilarity in stochastic community models based on the species-independence assumption

Ryosuke Iritani[1,2,*], Vicente J. Ontiveros[3,4], David Alonso[4], José A. Capitán[5], William Godsoe[6], and Shinichi Tatsumi[7]

[1]RIKEN Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS), RIKEN, Wako, Saitama, Japan
[2]Department of Biology, Faculty of Sciences, The University of Tokyo
[*]`lambtani@gmail.com`
[3]Institute of Aquatic Ecology, University of Girona, Girona, Spain
[4]Theoretical and Computational Ecology, Center for Advanced Studies of Blanes (CEAB-CSIC), Spanish Council for Scientific Research, Acces Cala St. Francesc 14, Blanes E-17300 Spain
[5]Complex Systems Group, Department of Applied Mathematics, Universidad Politécnica de Madrid, Avenida Juan de Herrera, 6, Madrid E-28040 Spain
[6]BioProtection Research Centre, Lincoln University, Lincoln, New Zealand
[7]Hokkaido Research Center, Forestry and Forest Products Research Institute, Sapporo, Hokkaido, Japan

January 1, 2023

- Conception and design: RI, VJO, WG, DA, and ST;
- Acquisition of data: not applicable;
- Analysis: RI, WG, DA, JAC, and ST;
- Interpretation: RI and ST;
- Drafting: all;
- Revision: all;
- All authors approved the content of the manuscript for submission.
- ORCID iD:
  RI 0000-0002-2396-1109
  VJO 0000-0001-8477-2574
  DA 0000-0002-8888-1644
  JAC 0000-0002-6245-0088
  WG 0000-0003-1697-6916
  ST 0000-0002-1789-1685
- No conflict of interest to be declared exists.
- Running head: *Beta-diversity in a heterogeneous world*
- Keywords: Beta-diversity; Heterogeneity; Jaccard index; Null models; Similarity coefficient; Species Distribution Models; Stochasticity.

Abstract

Understanding how the spatial variation in species composition (beta-diversity) varies with biotic and abiotic conditions is one of the ultimate goals in biology. Theory predicts that beta-diversity is a consequence of two factors, species-level differences (defined as the variations among species in the probabilities that species are present in the landscape) and spatial heterogeneity (defined as the difference, between two sites, in the probabilities with which species are present). At present, however, the importance of each factor is unclear. Here, we take a probabilistic and combinatorial approach to examine the effects of species differences and spatial heterogeneity on the degree to which species assemblages in two spatial locations differ in species compositions. We first derived analytical and approximation formulae of the expectation and variance of the pairwise beta-diversity, based on the assumption that the presence probabilities of species are independent of each other. We found that, contrary to the intuitive claim that differences among species lead to greater beta-diversity, our method predicts that the reverse is also likely under some, though not all, circumstances. Strikingly, when space is homogeneous, beta-diversity decreases with species differences. This suggests that policy making for increasing species differences would, without the effort to maintaining environmental heterogeneity, induce biotic homogeneization. These findings suggest that maintaining beta-diversity as a conservation target may lead to undesired outcome unless species traits and environmental homogenization are monitored and adequately managed. Second, we illustrate our method using data from five woodpecker species in Switzerland, showing that the woodpecker species' joint distributions change considerably with time, and also that such changes are basically explained by the changes in the incidences of some of the species. The new framework can improve our understanding of how pairwise beta-diversity responds to species differences and spatial heterogeneity.

1 Introduction

Beta-diversity (the spatial variation in species compositions) links diversity across scales (Whittaker 1972; Anderson *et al.* 2010; Chase *et al.* 2019; Poggiato *et al.* 2021). Beta-diversity varies with fundamental processes such as dispersal, environmental filtering and species interactions (Vellend 2010; Anderson *et al.* 2010; Socolar *et al.* 2016; Maynard *et al.* 2017; Legendre 2019; Thompson *et al.* 2020). Understanding the patterns of beta-diversity is thus considered as one of the ultimate goals in ecology. Reductions in beta-diversity, known as biotic homogenization (Olden & Poff 2003; Olden & Rooney 2006; Olden *et al.* 2018), have been caused by various types of global changes, such as urbanization (McKinney 2006), species invasion (Powell *et al.* 2013), climate change (Veech & Crist 2007), land-use change (Vellend *et al.* 2007), and interactions thereof (Karp *et al.* 2017). A consequence of biotic homogenization is the reduction of ecosystem functioning across the globe (Hautier *et al.* 2017; Mori *et al.* 2018; Albrecht *et al.* 2021; Wang *et al.* 2021). Understanding how beta-diversity changes in response to variations in biotic and abiotic conditions should lead to better management, conservation, and urban planning in our modern society (Crowther *et al.* 2015).

Beta-diversity is often measured by using pairwise indices based on empirical presence-absence (incidence) data (Koleff *et al.* 2003), but even for such simple, incidence-based beta-diversity, how biotic and abiotic variables influence the beta-diversity has not been conceptually established. That is, we know little about "beta-diversity patterns", the dependence of beta-diversity on biotic and abiotic factors. Indeed, results from previous work on beta-diversity patterns have been mixed. For example, theory shows that dispersal, which is one of the fundamental processes, tends to homogenize local communities and thereby reduce beta-diversity (Loreau 2000; Mouquet & Loreau 2003; Thompson *et al.* 2020), or to have rather opposing effects (Lu *et al.* 2019; Lu 2021). Meanwhile, experimental work suggests that dispersal may promote beta-diversity (Vannette & Fukami 2017). Disturbance like fire is also suggested to have opposing effects on beta-diversity (Altermatt *et al.* 2011; Myers *et al.* 2015; Ojima & Jiang 2016). To better interpret complicated beta-diversity patterns, it is necessary to establish a general framework that can incorporate variable biotic and abiotic conditions.

Among other factors, species differences and spatial heterogeneity both have profound effects on beta-diversity (Peres-Neto *et al.* 2001). As well as spatial heterogeneity (Veech & Crist 2007; Báldi 2008;

Jankowski *et al.* 2009; Allouche *et al.* 2012; Bar-Massada & Wood 2013; Heino *et al.* 2013; Pomara *et al.* 2013; Astorga *et al.* 2014; Heino *et al.* 2014; Bar-Massada 2015a,b; Royan *et al.* 2015; Bar-Massada & Belmaker 2017; Zorzal-Almeida *et al.* 2017; Ben-Hur & Kadmon 2020; Shinohara *et al.* 2022), species differences can strongly influence beta-diversity. For example, dispersal mode (passive vs. active) and body size (small vs. large) are both strong determinants of beta-diversity patterns as revealed by meta-analyses (Soininen *et al.* 2007; De Bie *et al.* 2012). Similarly, Cao *et al.* (2021) showed that beta-diversity may co-vary with niche-related characteristics. It is of great importance to take species differences and spatial heterogeneity both into account (Peres-Neto *et al.* 2001; Gotelli & Ulrich 2011). However, there is no established theoretical prediction for the effects of species differences and spatial heterogeneity.

To discern beta-diversity patterns under variable conditions, it is of great use to regard incidence-data and resulting beta-diversity both as stochastically changing quantities, because by doing so, we can assess how the probability distributions of beta-diversity varies with mechanistic factors that determine species presence probabilities. (Baroni-Urbani 1980; Real & Vargas 1996; Gotelli & Ulrich 2011; Hui & McGeoch 2014; Chung *et al.* 2019). In this article, we explore the details of pairwise compositional dissimilarity between two species assemblages (Jaccard dissimilarty) under the influence of species differences and spatial heterogeneity. The logical starting point is akin to one of the most successful null models in ecology, the Theory of Island Biogeography (TIB; MacArthur & Wilson 1963; MacArthur & Wilson 1967). Following recent work on TIB (Alonso *et al.* 2015; Ontiveros *et al.* 2021), our work assumes that species incidences are independent from each other both within and between sites, which nullifies any correlation between them, the so-called *species independence assumption*. We thereby examine the effect of species differences and spatial heterogeneity on the expectation of pairwise dissimilarity of species compositions (Jaccard 1908, 1912; Veech 2012; Arita 2017; Keil *et al.* 2021). This allows for the consideration of various realistic factors that drive local biotas dramatically different, such as the distance to the mainland, and specific spatial niche partitioning.

We first obtain the exact and approximated expression for the expectation and variance of Jaccard dissimilarity under the species independence assumption. Second, we examine how the expectation and variance in beta-diversity respond to species differences and spatial heterogeneity. Specifically, we numerically generate species presence probabilities for a pair of sites, and assess how varying species-differences and spatial heterogeneity influence beta-diversity patterns. We find that species differences can have opposing effects on beta-diversity. Counterintuitively, we find that species differences result in lower beta-diversity when site heterogeneity is scarce. While traditional analyses of beta-diversity focus on sites censused completely, many contemporary data-sets are based on statistical models such as Species Distribution Models (SDMs; Elith & Leathwick 2009; Guisan *et al.* 2017; Zurell *et al.* 2020). Indeed, recent work has attempted to improve the statistical power of local species richness estimation in SDMs under heterogeneity (Calabrese *et al.* 2013). We therefore examine how our approach can be applied to such statistical models. The temporal Jaccard dissimilarity (Legendre 2019; Magurran *et al.* 2019; Figure 1) is designed to project how the local species composition across regions is expected to change with time, particularly in response to global environmental changes. We combined our method with SDMs and assessed the expected changes in the compositions of woodpecker species across Switzerland. We believe that the approach be a starting point to provide further insights to the extent and intensity of future compositional change, and to help us allocate resources for tasks such as monitoring, conservation, or restoration.

Table 1: Summary of notation used in the main text.

Notation	**Definition**	**Note**
i	Species label	$i = 1, 2, ..., \text{ or } S$
S	The total number of species in the mainland	"species pool size"
j	Site label, with $j = 1$ or 2	"sites" may be spatial or temporal
$x_{i,j}$	Incidence of species i in site j	0 (absence) or 1 (presence)
$\mathbf{X}_{(S,2)}$	Incidence table of size S-by-2	Abbreviated to $\mathbf{X}$
$:=$	Defining a quantity	
$\equiv$	Identity	usually with respect to "all i"
Ω	Set of incidences (absence 0 and presence 1)	$:= \{ 0, 1 \}$
$\mathbf{x}_{\circ,j}$	Column vector of configuration in site j	
$\lvert \mathbf{x}_{\circ,j} \rvert$	The number of species present in site j	$= \sum_{i=1}^{S} x_{i,j}$
$\mathbf{x}_{i,\circ}$	Row vector of species i's incidence in space	
$p_{i,j}$	Probability of i present in j (i.e., Prob $\left[x_{i,j} = 1\right]$)	$a_{i,j} = 1 - p_{i,j}$ for probability of absence
$b_{i,\circ}$	Probability of i present in both sites 1 and 2	Prob $[x_{i,1} = x_{i,1} = 1]$; b for "both"
$d_{i,\circ}$	Probability of i absent from both sites 1 and 2	Prob $[x_{i,1} = x_{i,1} = 0]$; d for "double-absence"
$P_{\mathbf{X}}$	Probability that a table $\mathbf{X}$ is observed	$= \prod_{i=1}^{S} \prod_{j=1}^{2} p_{i,j}^{x_{i,j}} a_{i,j}^{1-x_{i,j}}$
$\beta_{\mathbf{X}}^{\mathrm{J}}$	Jaccard dissimilarity for an incidence table $\mathbf{X}$	
$\gamma_{\mathbf{X}}$	The total number of species present in the landscape for table $\mathbf{X}$	"Gamma-diversity"
$\mathbf{E}\left[\beta^{\mathrm{J}}\right]$	Expectation of Jaccard dissimilarity	$\mathbf{E}\left[\beta^{\mathrm{J}} \middle\vert \gamma > 0\right]$ for conditional expectation
$\mathbf{V}\left[\beta^{\mathrm{J}}\right]$	Variance of Jaccard dissimilarity	$\mathbf{Std}\left[\beta^{\mathrm{J}}\right] := \sqrt{\mathbf{V}[\beta^{\mathrm{J}}]}$
$\beta_{\mathrm{heur}}^{\mathrm{J}}$	Approximation of $\mathbf{E}\left[\beta^{\mathrm{J}} \middle\vert \gamma > 0\right]$	"Heuristic approximation"
w	Species difference	$0 \leq w < 2$; Eqn (5)
h	Spatial heterogeneity	$0 \leq h \leq 1$; Eqn (6)
μ_j	Average presence probability in site j	

2 Methods and Results

Model

Jaccard dissimilarity is defined as the proportion of the number of unique species to that of present species in, at least, one of the sites (Jaccard 1908, 1912), which measures a compositional difference between two sites (Box 1). Since Jaccard dissimilarity is designed to measure the difference in the compositions of two sites, we consider a landscape (metacommunity) consisting of two spatially or temporally segregated sites (Figure 1). We consider a matrix of binary elements, in which species are listed from top to bottom (labelled $i = 1, 2, ..., S$), and sites are listed from left to right (from $j = 1, 2$). That is, an incidence table of S species with 2 sites is written as $\mathbf{X}$. Following Chase *et al.*'s (2011) definition, we define "species pool" as the collection of species that can possibly inhabit either site within a reasonable time period during which the incidence data census is performed, and therefore defined S as the maximum number of species that may be present in at least one of the sites (with a positive probability; we refer to S as "species pool size"). We do not consider a species that can never be present.

We write $x_{i,j}$ for (i, j)-th element of an incidence table $\mathbf{X}$, with $x_{i,j} = 1$ if species i is present in j, and $x_{i,j} = 0$ otherwise absent. The formal approach to derive Jaccard dissimilarity and their moments is encapsulated in Box 1. We write $p_{i,j}$ for the probability that species i is present in site j (and this probabilistic approach is well-established; MacArthur & Wilson 1963; Real *et al.* 2016; Carmona & Pärtel 2020). The key assumption is that $p_{i,j}$ may take different values depending on species i and site j but are independent of each other (the species independence assumption). The sum $\sum_{i=1}^{S} x_{i,j}$, which represents the species richness (alpha-diversity), follows the Poisson-Binomial distribution with parameters $\left(p_{1,j}, p_{2,j}, ..., p_{S,j}\right)$. For brevity, we write: (i) $a_{i,j} := 1 - p_{i,j}$ for the probability that species i is absent in j; (ii) $b_{i,\circ} := p_{i,1}p_{i,2}$ for the probability that species i is present in both sites 1 and 2 (probability of "commonness"); (iii) $d_{i,\circ} := a_{i,1}a_{i,2}$ for the probability that species i is absent in both sites 1 and 2 (probability of "double-absence"). As a result, $1 - b_{i,\circ} - d_{i,\circ}$ represents the probability that species i is unique to either site 1 or 2. Note that we do not distinguish in which sites (1 or 2) the unique species are present. For example, colonization ability of species, stress-tolerance of species, distance from continents, and disturbance frequency all jointly influence presence probabilities.

To analyze the probability distribution of beta-diversity, we present a novel derivation using tools from stochastic analyses and combinatorics. These tools allow us to exactly compute the expectation and variance of Jaccard dissimilarity expected under species-independence assumptions. The computation can take a long time, because of combinatorial calculations associated with species-specific and site-specific presence probabilities $(p_{1,1}, p_{1,2}, p_{2,1}, ..., p_{S,1}, p_{S,2})$. To make the formula accessible to as broad researchers as possible, we deviced a fast computable approximation (below).

Box 1: Descriptions of the formal approach and Jaccard dissimilarity

We write $\mathbf{X} \in \Omega^S \otimes \Omega^2$ (with $\otimes$ for a direct product between sets) to indicate that the incidence table $\mathbf{X}$ is a matrix with S rows and 2 columns, each of whose elements is either 0 for absence or 1 for presence. We also write $\mathbf{x}_{\circ,j} = \left(x_{1,j}, x_{2,j}, ..., x_{S,j}\right)^{\top}$ for a column vector within the incidence matrix ($\mathbf{X}$) in a site j, with $\top$ for transpose. Therefore, $\mathbf{x}_{\circ,j} \in \Omega^S$. We write $\left|\mathbf{x}_{\circ,j}\right| := \sum_{i=1}^{S} x_{i,j}$ for alpha-diversity in a site j. In addition, we write $\langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \rangle := \sum_{i=1}^{S} x_{i,1}x_{i,2}$ for the number of species present in sites 1 and 2 both. Note that the number of species present in the landscape (gamma-diversity) is given by $\left|\mathbf{x}_{\circ,1}\right| + \left|\mathbf{x}_{\circ,2}\right| - \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \rangle$, which is integer-valued varying between 0 (no species present) and S (all species present in the landscape).

Jaccard dissimilarity index (Jaccard 1908, 1912) for a given table $\mathbf{X}$, denoted by $\beta^{\mathrm{J}}_{\mathbf{X}}$, is defined by:

$$\beta^{\mathrm{J}}_{\mathbf{X}} = \frac{(\#\text{ unique species})}{(\#\text{ species present in at least one of the sites})} = \frac{\left(\left|\mathbf{x}_{\circ,1}\right| - \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2}\rangle\right) + \left(\left|\mathbf{x}_{\circ,2}\right| - \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2}\rangle\right)}{\left|\mathbf{x}_{\circ,1}\right| + \left|\mathbf{x}_{\circ,2}\right| - \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2}\rangle}, \tag{1}$$

which represents the proportion of the number of unique species (present only one of the two sites) to the number of present species (present at least one of the two sites). Eqn (1) is written as a function of stochastic variables, $x_{i,j}$.

We aim to evaluate the first and second moments (expectation and variance) of Jaccard dissimilarity and assess how the moments vary with parameters of species differences and spatial heterogeneity. If the expectation of Jaccard dissimilarity varies in a certain way with a varying parameter, then the Jaccard dissimilarity is expected to behave accordingly and thereby forms a "beta-diversity pattern"

Box 2: Expectations: exact

We denote the expectation of Jaccard dissimilarity by $\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma > 0\right]$. We found that (Appendix A):

$$\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma > 0\right] = \frac{\sum_{i=1}^{S} \sum_{\mathbf{y}\in\Omega^S} \frac{1}{|\mathbf{y}|} \frac{y_i\left(1-b_{i,\circ}-d_{i,\circ}\right)}{{}_S\mathrm{C}_{|\mathbf{y}|}} \prod_{\ell(\neq i)}^{S} d_{\ell,\circ}^{1-y_\ell}}{1 - \prod_{k=1}^{S} d_{k,\circ}} \tag{2}$$

where ${}_S\mathrm{C}_{|\mathbf{y}|}$ represents the binomial coefficient, which counts the number of ways, disregarding order, that $|\mathbf{y}|$ $(= 0, 1, ..., S)$ species can be chosen from S species (Van Lint & Wilson 2001). Eqn (2) is conditioned on the premise that at least one of the species is present in the landscape (which occurs with probability $1 - \prod_{i=1}^{S} d_{i,\circ}$). Note that $\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma > 0\right]$ has a symmetry in site 1 and 2, by which swapping (permutating) $p_{i,1}$ and $p_{i,2}$ for any species i does not change the result (Baselga 2010).

As is detailed in Appendix A, Eqn (2) recovers the formula for species-equivalence cases (Chase *et al.* 2011; Lu *et al.* 2019). As suggested in previous studies, the species pool size S, which represents the maximum number of species present in the landscape, has no effects on the expected Jaccard dissimilarity conditioned of at least one of the species being present. The invariance is because species being equivalent implies that the incidence of any species occurs equally likely and independently, regardless of how many species can potentially inhabit the landscape.

Approximations for expectation and variance

We here present a "heuristic" approximation for the expectation of Jaccard dissimilarity, which was used in the previous studies (**Kalyuzhny_etal2021**; Ontiveros *et al.* 2021). We write $\beta^{\mathrm{J}}_{\mathrm{heur}}$ for the expectation of the numerator divided by the expectation of the denominator; that is:

$$\begin{aligned}\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma > 0\right] \approx \beta^{\mathrm{J}}_{\mathrm{heur}} &= \frac{\mathbf{E}\left[\#\text{ unique species}\right]}{\mathbf{E}\left[\#\text{ species present at least one of the sites}\right]} \\ &= \frac{\sum_{i=1}^{S}\left(1 - b_{i,\circ} - d_{i,\circ}\right)}{\sum_{i=1}^{S}\left(1 - d_{i,\circ}\right)}\end{aligned} \tag{3}$$

(the derivation and interpretation are provided in Appendix B). In the example below, we will show that Eqn (3) provides a near-identical approximation for the conditional expectation $\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma > 0\right]$.

We also explored for a formula of the variance of Jaccard dissimilarity, but it involves much complication in general, and also our main focus is on expectations. Therefore, we show the results on variance for species-equivalence case and encapsulated the detailed analyses for general cases with species differences in the Appendix D.

Applications

Special case: identical species

We demonstrate a special case where the species are identical in any sense but the landscape can be heterogeneous; that is: $p_{i,j} \equiv p_{\circ,j}$ for each $j = 1, 2$, and consequently we write $b_{\circ,j} \equiv b$ and $d_{\circ,j} \equiv d$ for commonness and double-absence (respectively; note that $p_{\circ,1}$ and $p_{\circ,2}$ may differ). Substituting those into the expected Jaccard dissimilarity index (Eqn (2)) yields:

$$\mathbf{E}\left[\beta^{\mathrm{J}} \middle| \gamma > 0\right]\Bigg|_{\text{species equivalence}} = \frac{1-d-b}{1-d} = \beta^{\mathrm{J}}_{\text{heur}}. \tag{4}$$

That is, the heuristic approximation and the exact conditional expectation completely agree with each other. Eqn (4) also recovers the results of previous studies that assume species' presence probabilities are identical (Chung *et al.* 2019; Lu *et al.* 2019; Lu 2021; Ontiveros *et al.* 2021). In other words, the sufficient condition for the exact and approximated formulae of the expectation to be equal is that species are equal in presence probabilities.

We investigated the dependence of the variance of Jaccard dissimilarity on the pair of presence probabilities $\left(p_{\circ,1}, p_{\circ,2}\right)$ (Figure 2B). We found that: (i) when $p_{\circ,1}$ and $p_{\circ,2}$ are small, the variance tends to take large values; (ii) when only one of $p_{\circ,1}$ and $p_{\circ,2}$ is small but the other is large, the variance is likely to take small values, but the degree of decrease *per se* is very small (<0.05; Figure 2B); (iii) when both values are large, the variance tends to be small; and (iv) when both probabilities are at intermediate values, the dependence of the variance appears to be small.

Effects of species difference and spatial heterogeneity

We examined how species differences and spatial heterogeneity jointly influence the expectation of Jaccard dissimilarity, by means of generating synthetic incidence patterns. For incidence probabilities $\left(p_{1,j}, \dots, p_{S,j}\right)$ with $j = 1, 2$, we define species difference w as the sum, taken between two sites, of the average absolute deviation in the presence probabilities across species, divided by the sum of the species-average presence probabilities in two sites. Also, we define the spatial heterogeneity h as the average, taken among species, of the absolute differences in the presence probabilities between two sites:

$$w := \frac{1}{\mu_1 + \mu_2} \sum_{j=1}^{2} \frac{1}{S} \sum_{i=1}^{S} \left|p_{i,j} - \mu_j\right| \tag{5}$$

$$h := \frac{1}{S} \sum_{i=1}^{S} \left|p_{i,1} - p_{i,2}\right|, \tag{6}$$

with $\mu_j := \sum_{i=1}^{S} p_{i,j}$ for $j = 1$ and 2 the average presence probability in site 1 and 2 respectively (for the definition of species difference, see Hoover index of income inequality; Hoover 1936; Chao & Ricotta 2019 and Appendix C). Note that species differences and spatial heterogeneity are functions of the presence probabilities.

We assess how varying species difference and spatial heterogeneity jointly alter $\mathbf{E}\left[\beta^{\mathrm{J}}\right]$. The first possible case is concerned with $p_{i,1} = p_{i,2}$ for all species, that is, when two sites are homogeneous. In this case, increasing the species difference w reduces beta diversity (Figure 3A-1). This is the case also in the presence of a slight difference between two sites (Figure 3A-2). Hence, when there is at most low spatial heterogeneity, species differences systematically reduce beta-diversity.

As the degree of site heterogeneity h becomes larger, whether species differences result in larger or smaller beta-diversity becomes less consistent. We found both cases where species differences increase and decrease beta-diversity (Figure 3B-1, B-2). The difference between Figure 3B-1 and B-2 is that while Figure 3B-1 assumes that two sites are heterogeneous evenly among species ($\left|p_{i,1} - p_{i,2}\right|$ all equal), Figure 3B-2 uses different values $\left|p_{i,1} - p_{i,2}\right|$ (although both are on average equal). That is, the effects of species differences on beta-diversity depend on the patterns of species differences and spatial heterogeneity.

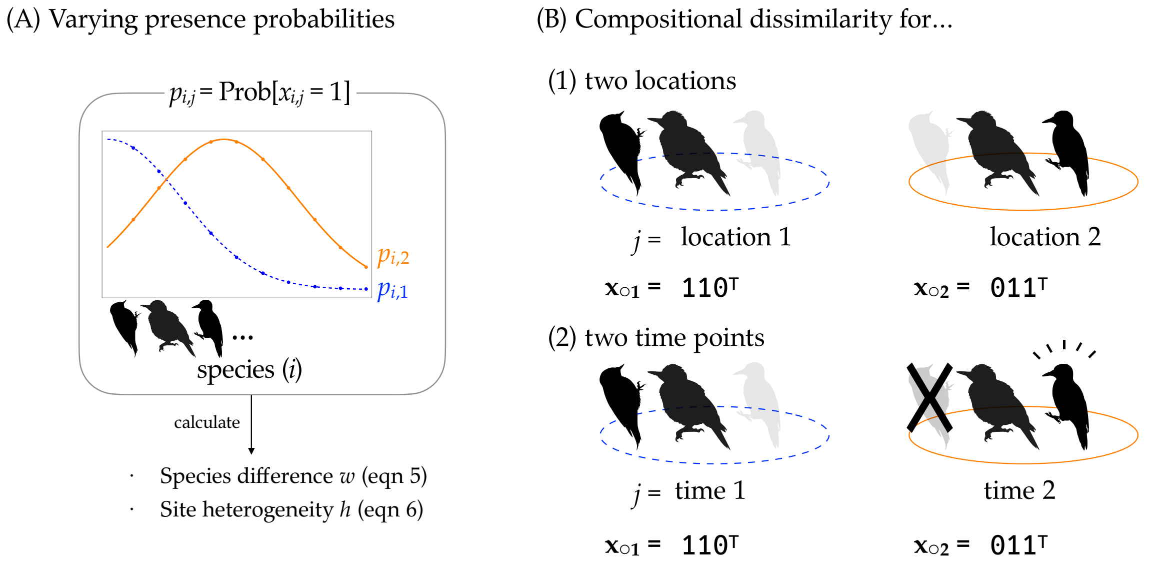
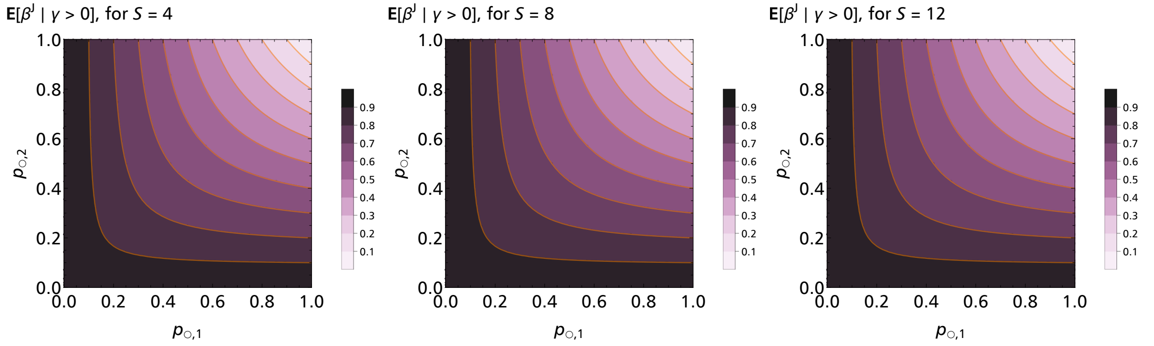

(A) Varying presence probabilities
$p_{i,j} = \mathrm{Prob}[x_{i,j} = 1]$
$p_{i,2}$
$p_{i,1}$
species (i)
calculate
· Species difference w (eqn 5)
· Site heterogeneity h (eqn 6)
(B) Compositional dissimilarity for...
(1) two locations
$j =$ location 1
location 2
$\mathbf{x}_{\circ 1} =$ 110$^\top$
$\mathbf{x}_{\circ 2} =$ 011$^\top$
(2) two time points
$j =$ time 1
time 2
$\mathbf{x}_{\circ 1} =$ 110$^\top$
$\mathbf{x}_{\circ 2} =$ 011$^\top$

Figure 1: Schematic illustration of the framework. (A) The presence probabilities $(p_{1,1}, p_{1,2}, p_{2,1}, \ldots, p_{S,2})$ take different values. Of our interest are the effects of the differences (measured by w) and site heterogeneity (measured by h) on compositional dissimilarity, both of which are computed from given $p_{1,1}, \ldots, p_{S,1}, p_{1,2}, \ldots, p_{S,2}$. (B) The compositional dissimilarity (defined by the Jaccard dissimilarity) can be applied to spatial and temporal changes. When j represents a spatial site, we compute the expected Jaccard dissimilarity from the compositions of the two site (for example 110 and 011). We can carry out the same calculation for time point labels $j = 1, 2$. Note the symbol $\top$ for transpose. The woodpecker pictures are from `http://phylopic.org/` (CC0).

(A) Conditional expectation of Jaccard dissimilarity increases with low presence probabilities



(B) Conditional standard deviation of Jaccard dissimilarity increases with low presence probabilities

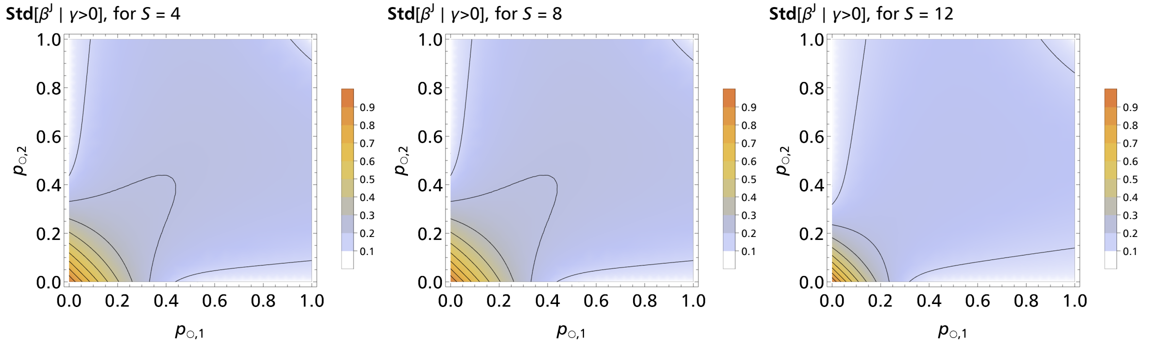
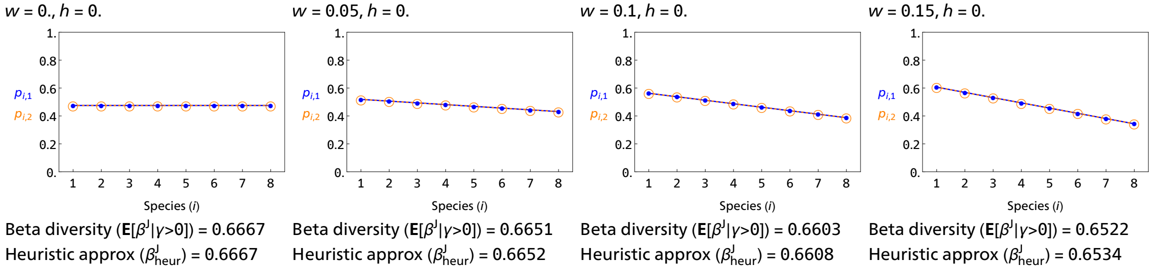
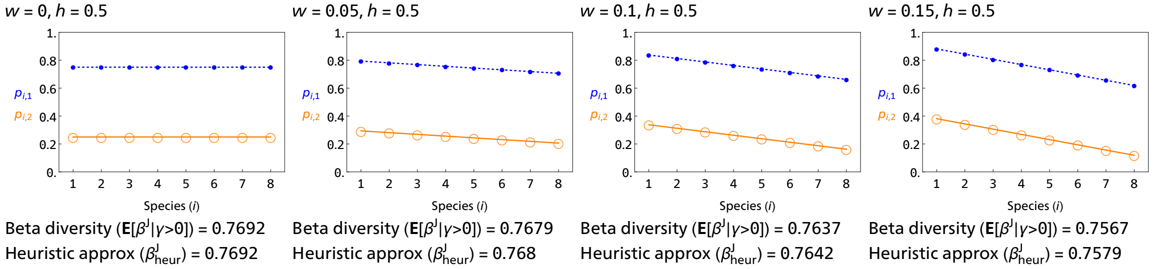


Figure 2: The conditional expectation of Jaccard dissimilarity $\mathbf{E}\left[\beta^J \middle| \gamma > 0\right]$ (panel A) and the conditional standard deviation of Jaccard dissimilarity $\mathbf{Std}\left[\beta^J \middle| \gamma > 0\right] := \sqrt{\mathbf{V}\left[\beta^J \middle| \gamma > 0\right]}$ (B), plotted on $p_{\circ,1}$ (horizontal axis), $p_{\circ,2}$ (vertical axis) for varying species pool sizes, S. (A) The expected Jaccard dissimilarity is lower when the probability of commonness is larger (right top zones), and higher when the probability of being present in only one of the sites is higher (left top and right bottom zones). The effect of S on the expectation is negligibly small, and thus increasing S (from left to right panels) does not appear to change the overall trend. (B) The standard deviation of Jaccard dissimilarity is lower when the presence probabilities take extreme values (four corners). As S increases, the standard deviation tends to be less dependent on intermediate values of presence probabilities, as seen from the observation that over the wide region of the rectangle, the standard deviation is relatively unchanged with presence probabilities.

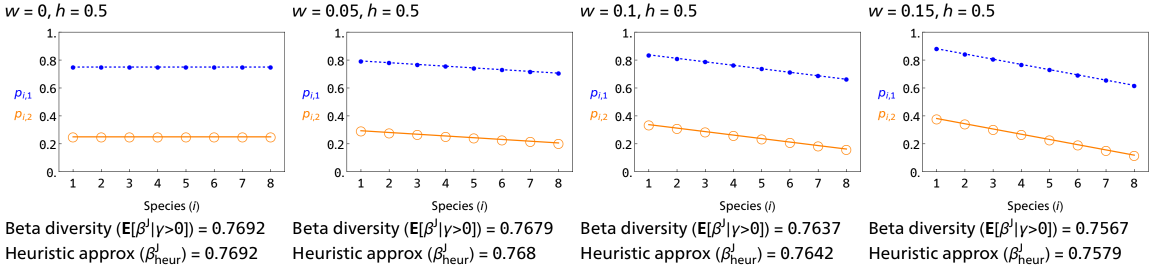
(A-1) Sites are homogeneous ⇒ beta decreases with species difference w



(A-2) Sites are weakly heterogeneous ⇒ beta decreases with w



(B-1) Sites are strongly heterogeneous ⇒ beta *may decrease* with w



(B-2) Sites are strongly heterogeneous ⇒ beta *may increase* with w

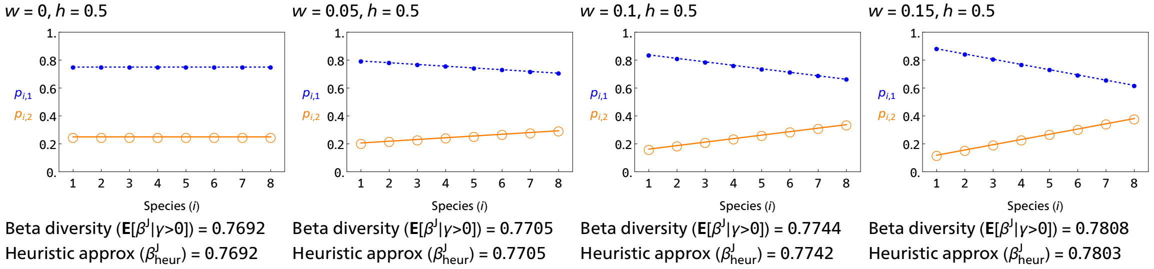


Figure 3: The beta-diversity patterns under species differences and spatial heterogeneity. Setup: blue dashed curves plot $p_{i,1}$ and orange $p_{i,2}$ respectively (referred to as incidence gradient curves). In all panels, the average presence probability per species per site is set one half: $\left(\mu_1 + \mu_2\right) / 2 = 0.5$. Species differences and site-heterogeneity are calculated from the incidence curves, each indicated. (A-1) When sites are completely homogeneous (i.e., when $p_{i,1} = p_{i,2}$ for all species $i = 1, ..., S$), increasing species difference results in lower beta-diversity. (A-2) When sites are weakly heterogeneous $h = 0.05$, species difference reduces beta-diversity, as in (A). (B) When two sites are strongly heterogeneous with $h = 0.40$ in this example, increasing species differences can either decrease (B-1) or increase (B-2) beta-diversity.

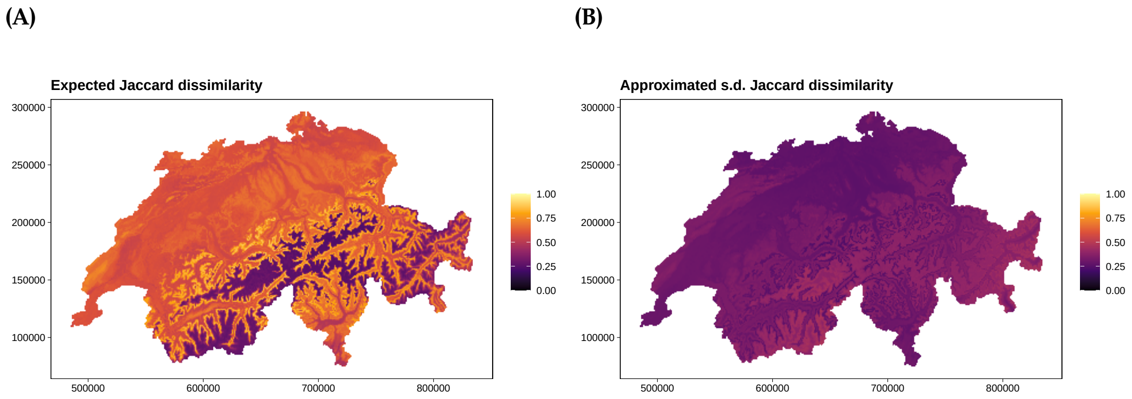


Figure 4: Outcomes of species distribution models, using a climatic variable. We quantified the expected, compositional dissimilarity of five woodpecker species at two time points, current and future, over the region of Switzerland. That is, we assessed the expectation and standard deviation of temporal Jaccard dissimilarity (Legendre 2019; Magurran *et al.* 2019). We used occupancy estimations for current and future climatic conditions over Switzerland. (A) Expectation. Compositional changes are expected to be high in the upper limit of the current distribution and lowlands. (B) Standard deviation (approximated). The standard deviation tends to be small, which is consistent with the analytical prediction in Figure 2.

Temporal Jaccard dissimilarity with Species Distribution Models

We provide a case application of our method using the Species Distribution Models (SDMs; Elith & Leathwick 2009; Guisan *et al.* 2017; Zurell *et al.* 2020). SDMs seek to estimate the probability that each species is present at a given site given information on the environment found at that site. Our approach enables quantification of the changes in species distribution on a location-by-location basis, based on temporal Jaccard dissimilarity (the proportion of the number of species that are present only one of two time points relative to the number of species that are present in at least one of the two time points; Legendre 2019; Magurran *et al.* 2019, Figure 1). The details are described in Appendix E.

Note that the temporal data systematically comes with temporal autocorrelation (i.e., correltion between compositions at two time points), but our analysis takes advantage of the site-permutation symmetry in Jaccard dissimilarity so that we can omit the correlations between two compositions.

Based on Schmid *et al.* (1998, 2018) and Zurell *et al.* (2019b, 2020), we used data of five woodpecker species, *Picus viridis*, *P. canus*, *Dendrocopos major*, *D. minor*, and *Dryocopus martius* in Switzerland. These species have common evolutionary history but use different habitats (Benz *et al.* 2006; Pasinelli 2007; Pons *et al.* 2010). For example, *P. canus* and *D. minor* occur at lowlands, while *P. viridis* is more widely found across Switzerland (Appendix E). The variation in geographic habitat use arguably reflects species differences, making the system ideal for an application of temporal Jaccard dissimilarity. Note also that Incidence-based SDMs rely on the species independence assumption and therefore are reconciled with the species-independence assumption.

We examined how dissimilar woodpeckers metacommunity is going to be under future climatic conditions across Switzerland (i.e., time-decay of similarity), compared with the sub-community under the current climate. Our methods predicted that the expectations of temporal Jaccard dissimilarity were unexpectedly high (Figure 4), indicating significant future changes in the distribution of the woodpeckers. This result was primarily explained by species dynamics in lowland sites where some species thrived and others failed: *P. canus*, that would decrease its occupancy rate near the rivers and would increase in surrounding areas (SI Figure 2), and *D. minor*, whose occupancy is expected to increase in lowlands and valleys (SI Figure 2). Dissimilarity in hillsides is expected to be moderate due to a general increment in richness (SI Figure 2). These results are consistent with a general trend of Switzerland forest birds moving to higher grounds as a response to environmental change (Maggini *et al.* 2014). Our results demonstrate the dramatic effects of climate change on temporal beta-diversity of a bird metacommunity.

3 Discussion

¶1 We derived the formulae of the expectation and variance of Jaccard dissimilarity index as a measure of compositional dissimilarity between two sites for nonidentical species in heterogeneous landscapes. We showed that species differences may have nonmonotonic effects on Jaccard dissimilarity (Figure 3). When sites are relatively similar in their species presence probabilities, species difference reduces Jaccard dissimilarity. When sites are different in their presence probabilities, species differences can cause varying effects on Jaccard dissimilarity, making robust prediction challenging. Our work allows us to implement empirical incidence data of presence probabilities and assess the long term estimations for Jaccard dissimilarity. In addition, our method enables quantitative comparison of predicted outcomes with observed species distributions. Hence, the present work offers a powerful framework to theoretically and quantitatively investigate spatial variations in species compsition.

¶2 Our results suggest that knowledge of the species pool will be important for some properties but not others. When all species are identical, we found that the conditional expectation of Jaccard dissimilarity (given that at least one species is likely present) is independent of species pool size, as was revealed in the literature (Lu *et al.* 2019; Lu 2021). In contrast, we found that the variance more substantially depends on the species pool size even when the species presence probabilities are equal in each pair of sites (Figure 2B). Specifically, the variance becomes smaller when the species pool size is larger, suggesting that large species pool sizes may reduce the uncertainty in statistical inferences for Jaccard dissimilarity. However, the uncertainty may become larger when the baseline presence probabilities are smaller (Figure 2B), suggesting that the effect of opportunistic species (species with the presence probabilities are very low but not zero) may influence the robust inference of Jaccard dissimilarity, and also that rarely observed species can increase the uncertainty of estimations for Jaccard dissimilarity (Wolda 1981; Colwell & Coddington 1994; Plotkin & Muller-Landau 2002; Chao *et al.* 2004). These predictions are consistent with the concept of dark diversity (formally defined as the number of species that are absent from an ecosystem but which belong to its species pool; Pärtel *et al.* 2011; Carmona & Pärtel 2020), and our work suggests that dealing with the uncertainty is the key to predict the dark diversity.

¶3 Comparison with simulated data showed that the approximation of the Jaccard dissimilarity agrees well with the exact formula. The accuracy increases with the increasing species pool size (Appendix B). Considering that calculating the exact expectation of Jaccard dissimilarity in a brute force approach incurs large computational expenses, we suggest that, when the species pool size is relatively small, exact evaluation of the Jaccard dissimilarity is suitable and should be compared with the approximated value which can be computed faster. When the species pool size is large, we suggest using our approximations and exact values both, e.g., by applying Gauß' fast Fourier Transforms (Cooley & Tukey 1965; Heideman *et al.* 1984). When the species pool size is unknown, the Eqn (3) is beneficial, because the interpretation is clear and the expectations in the numerator and denominator can be assessed separately.

¶4 Using data on woodpecker distributions, we projected shifts in habitat use. This approach required only information on the expectation and variance. Likely mechanisms of the species differences and temporal heterogeneity in this system include colonization abilities, habitat selection, and species-specific tolerance to environmental challenges. By specifying possible mechanisms, we can obtain deeper insight into the processes by which biotic homogeneization occurs. For instance, partitioning beta-diversity into colonization and extinction components can be of great use to better understand dynamic processes of beta-diversity (Tatsumi *et al.* 2021). Our study can be used in conjunction with data such as SDMs for better understanding and management of the spatio-temporal dynamics of biodiversity.

¶5 Our key finding is that beta-diversity increases or decreases with species differences depending on the degree of spatial heterogeneity, other things being equal (Figure 3). To facilitate biological interpretations for the prediction, we first list species in the descending order of presence probability in site 1 (i.e., $p_{1,1} \geq p_{2,1} \geq \cdots \geq p_{S,1}$). We plotted the curve of $p_{i,j}$ against $i = 1, 2, ..., S$ and termed this curve as a "species incidence gradient" for site j. Under no spatial heterogeneity (i.e., $p_{i,1} = p_{i,2}$ for all species

$i = 1, \dots, S$, implying $h = 0$), the two species incidence gradients synchronize perfectly. In this case, increasing species differences increases the probability of some species being either (i) double-absent or (ii) present in both sites. In either case (i) or (ii), fewer species are present in only one of the sites than the case with no species difference, thus leading to lower beta-diversity. Biologically, spatial homogeneity tends to constrain patterns of species incidence gradients in a way that species differences cause species to be either present in both sites or absent from both sites. In other words, species differences under homogeneity do not foster uniquely present species. This prediction is robust against small increases in spatial heterogeneity (Figure 3A-2). In a weakly heterogeneous metacommunity, beta-diversity is predicted to decrease with species differences due to the synchronization of species incidence gradients between two sites. Maintaining large beta-diversity under biotic homogeneization requires all species to be equally likely to be present in both sites.

With increasing spatial heterogeneity, however, the effect of species difference on beta-diversity becomes less clear (Figure 3B). This is because in the presence of spatial heterogeneity, species incidence gradients tend to become asyncrhonous. When the degree of spatial heterogeneity is very large, two species incidence gradients are necessarily desynchronized (or form a "nestedness" or "turnover pattern", *sensu* Harrison *et al.* 1992, by which species incidences tend to be spatially segregated). The difference between Figure 3B-1 (in which beta-diversity decreases with species difference) and B-2 (otherwise increases) predicts that the variation in species-wise spatial heterogeneity $|p_{i,1} - p_{i,2}|$ matter, though we were unable to corroborate this speculation. Hence, heterogeneous environments make the prediction and management of beta-diversity even more challenging.

¶6 Our study has significant implications for conservation. Generally, beta-diversity is a key factor for ecosystem functioning from local to global scales (Socolar *et al.* 2016; Mori *et al.* 2018). Local ecosystem functioning may be driven by species' functional dissimilarity, e.g., in niches (Godoy *et al.* 2020). For example, Loiseau *et al.* (2016) pointed out that conservation policy designed to protect taxonomic diversity cannot be fully reconciled with functional diversity management. Indeed our finding predicts that with the environmental homogeneity, increasing local species differences lead to lower beta-diversity (Figure 3A-1, B). Given that the species difference is associated with functional diversity (or transformed quantity thereof; Palacio *et al.* 2022), the present prediction suggests that a conservation policy aiming to maintain high beta-diversity be traded-off against the local, functional diversity. This dilemma becomes more complicated when the spatial heterogeneity is strong, by which beta-diversity may or may not decrease with species-differences, thereby making the prediction of beta-diversity rather difficult. One promising approach is thus to identify species traits and environmental factors for maintaining a balance (evenness) in the likeliness of species presence and absence, which may produce new interesting questions. Moving forward, open questions include: how does incidence-based beta-diversity respond to changes in functional diversity in colonization ability and extinction tolerance? How does functional diversity, in turn, respond against the reduction in compositional dissimilarity (biotic homogeneization)?

¶7 To conclude, we have derived the analytic formula of the expectation and variance of Jaccard dissimilarity index incorporating different species in a pair of heterogeneous sites, and revealed the opposing effects of species differences on beta-diversity. Assuming that species incidences are uncorrelated with each other, we found that species differences in presence probabilities may lead to a complex pattern (Figure 3). This work will help researchers better understand the probabilistic, or stochastic, nature of Jaccard dissimilarity (Real & Vargas 1996). Future studies may explore the effects of species associations on the probabilistic properties of Jaccard dissimilarity, and also carry out occupancy dynamics analyses, beyond pairwise dissimilarity analyses (MacKenzie *et al.* 2018). One of the most promising approaches is a process-based approach (Pilowsky *et al.* 2022), by which we can incorporate further complications that influence beta-diversity. Our method can incorporate additional realities to track and manage the changes in species distributions under global changes.

4 Acknowledgement

The authors thank Ryosuke Nakadai, Naoto Shinohara and Akira Terui, for helpful comments. We thank JSPS-KAKENHI (grant numbers 19K22457, 19K23768, and 20K15882 to RI) for funding. The authors also thank Margarita Salas grant funded by the Spanish Ministry of Universities and the "European Union - Next GenerationEU" to VJO, CRISIS (PGC2018-096577-B-I00) to DA and JAC, UNIQUE (PID2021-127202NB-C21) to DA, and PRIORITY (PID2021-127202NB-C22) to JAC, all funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe". RI was inspired by some of the questions and answers on Cross Validated (Stack Exchange), in calculating the expectation of reciprocals (c.f., Cressie *et al.* 1981; Lange 2010).

References

Albrecht, J., Peters, M. K., Becker, J. N., Behler, C., Classen, A., Ensslin, A., Ferger, S. W., Gebert, F., Gerschlauer, F., Helbig-Bonitz, M., Kindeketa, W. J., Kühnel, A., Mayr, A. V., Njovu, H. K., Pabst, H., Pommer, U., Röder, J., Rutten, G., Costa, D. S., Sierra-Cornejo, N., Vogeler, A., Vollstädt, M. G. R., Dulle, H. I., Eardley, C. D., Howell, K. M., Keller, A., Peters, R. S., Kakengi, V., Hemp, C., Zhang, J., Manning, P., Mueller, T., Bogner, C., Böhning-Gaese, K., Brandl, R., Hertel, D., Huwe, B., Kiese, R., Kleyer, M., Leuschner, C., Kuzyakov, Y., Nauss, T., Tschapka, M., Fischer, M., Hemp, A., Steffan-Dewenter, I., & Schleuning, M. (2021). Species richness is more important for ecosystem functioning than species turnover along an elevational gradient. *Nature Ecology & Evolution*, **5**.12, pp. 1582–1593. DOI: `10.1038/s41559-021-01550-9` (cit. on p. 2).

Allouche, O., Kalyuzhny, M., Moreno-Rueda, G., Pizarro, M., & Kadmon, R. (2012). Area–heterogeneity tradeoff and the diversity of ecological communities. *Proceedings of the National Academy of Sciences*, **109**.43, pp. 17495–17500. DOI: `10.1073/pnas.1208652109` (cit. on p. 3).

Alonso, D., Pinyol-Gallemí, A., Alcoverro, T., & Arthur, R. (2015). Fish community reassembly after a coral mass mortality: higher trophic groups are subject to increased rates of extinction. *Ecology Letters*, **18**.5, pp. 451–461. DOI: `10.1111/ele.12426` (cit. on p. 3).

Altermatt, F., Schreiber, S., & Holyoak, M. (2011). Interactive effects of disturbance and dispersal directionality on species richness and composition in metacommunities. *Ecology*, **92**.4, pp. 859–870. DOI: `10.1890/10-1095.1` (cit. on p. 2).

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C., & Swenson, N. G. (2010). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**.1, pp. 19–28. DOI: `10.1111/j.1461-0248.2010.01552.x` (cit. on p. 2).

Arita, H. T. (2017). Multisite and multispecies measures of overlap, co-occurrence, and co-diversity. *Ecography*, **40**.6, pp. 709–718. DOI: `10.1111/ecog.01942` (cit. on p. 3).

Arnold, B. C. (2012). *Majorization and the Lorenz order: A brief introduction*. **43**. Springer Science & Business Media (cit. on p. 27).

Astorga, A., Death, R., Death, F., Paavola, R., Chakraborty, M., & Muotka, T. (2014). Habitat heterogeneity drives the geographical distribution of beta diversity: the case of N ew Z ealand stream invertebrates. *Ecology and Evolution*, **4**.13, pp. 2693–2702. DOI: `10.1002/ece3.1124` (cit. on p. 3).

Báldi, A. (2008). Habitat heterogeneity overrides the species–area relationship. *Journal of Biogeography*, **35**.4, pp. 675–681. DOI: `10.1111/j.1365-2699.2007.01825.x` (cit. on p. 3).

Bar-Massada, A. (2015a). Complex relationships between species niches and environmental heterogeneity affect species co-occurrence patterns in modelled and real communities. *Proceedings of the Royal Society B: Biological Sciences*, **282**.1813, p. 20150927. DOI: `10.1098/rspb.2015.0927` (cit. on p. 3).

Bar-Massada, A. (2015b). Immigration rates and species niche characteristics affect the relationship between species richness and habitat heterogeneity in modeled meta-communities. *PeerJ*, **3**, e832. DOI: `10.7717/peerj.832` (cit. on p. 3).

Bar-Massada, A. & Belmaker, J. (2017). Non-stationarity in the co-occurrence patterns of species across environmental gradients. *Journal of Ecology*, **105**.2, pp. 391–399. DOI: `10.1111/1365-2745.12713` (cit. on p. 3).

Bar-Massada, A. & Wood, E. M. (2013). The richness-heterogeneity relationship differs between heterogeneity measures within and among habitats. *Ecography*, **37**.6, pp. 528–535. DOI: `10.1111/j.1600-0587.2013.00590.x` (cit. on p. 3).

Baroni-Urbani, C. (1980). A statistical table for the degree of coexistence between two species. **44**.3, pp. 287–289. DOI: `10.1007/bf00545229` (cit. on p. 3).

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**.1, pp. 134–143. DOI: `10.1111/j.1466-8238.2009.00490.x` (cit. on p. 6).

Ben-Hur, E. & Kadmon, R. (2020). Heterogeneity–diversity relationships in sessile organisms: a unified framework. *Ecology Letters*, **23**.1, pp. 193–207. DOI: `10.1111/ele.13418` (cit. on p. 3).

Benz, B. W., Robbins, M. B., & Peterson, A. T. (2006). Evolutionary history of woodpeckers and allies (Aves: Picidae): Placing key taxa on the phylogenetic tree. *Molecular Phylogenetics and Evolution*, **40**.2, pp. 389–399. DOI: `10.1016/j.ympev.2006.02.021` (cit. on p. 10).

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2013). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, **23**.1, pp. 99–112. DOI: `10.1111/geb.12102` (cit. on p. 3).

Cao, K., Condit, R., Mi, X., Chen, L., Ren, H., Xu, W., Burslem, D. F. R. P., Cai, C., Cao, M., Chang, L.-W., Chu, C., Cui, F., Du, H., Ediriweera, S., Gunatilleke, C. S. V., Gunatilleke, I. U. A. N., Hao, Z., Jin, G., Li, J., Li, B., Li, Y., Liu, Y., Ni, H., O'Brien, M. J., Qiao, X., Shen, G., Tian, S., Wang, X., Xu, H., Xu, Y., Yang, L., Yap, S. L., Lian, J., Ye, W., Yu, M., Su, S.-H., Chang-Yang, C.-H., Guo, Y., Li, X., Zeng, F., Zhu, D., Zhu, L., Sun, I.-F., Ma, K., & Svenning, J.-C. (2021). Species packing and the latitudinal gradient in beta-diversity. *Proceedings of the Royal Society B: Biological Sciences*, **288**.1948. DOI: `10.1098/rspb.2020.3045` (cit. on p. 3).

Carmona, C. P. & Pärtel, M. (2020). Estimating probabilistic site-specific species pools and dark diversity from co-occurrence data. *Global Ecology and Biogeography*, **30**.1, pp. 316–326. DOI: `10.1111/geb.13203` (cit. on pp. 5, 11).

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T.-J. (2004). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**.2, pp. 148–159. DOI: `10.1111/j.1461-0248.2004.00707.x` (cit. on p. 11).

Chao, A. & Ricotta, C. (2019). Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, **100**.12. DOI: `10.1002/ecy.2852` (cit. on pp. 7, 28).

Chase, J. M., Kraft, N. J. B., Smith, K. G., Vellend, M., & Inouye, B. D. (2011). Using null models to disentangle variation in community dissimilarity from variation in α-diversity. **2**.2, art24. DOI: `10.1890/es10-00117.1` (cit. on pp. 5, 6).

Chase, J. M., McGill, B. J., Thompson, P. L., Antão, L. H., Bates, A. E., Blowes, S. A., Dornelas, M., Gonzalez, A., Magurran, A. E., Supp, S. R., Winter, M., Bjorkman, A. D., Bruelheide, H., Byrnes, J. E. K., Cabral, J. S., Elahi, R., Gomez, C., Guzman, H. M., Isbell, F., Myers-Smith, I. H., Jones, H. P., Hines, J., Vellend, M., Waldock, C., & O'Connor, M. (2019). Species richness change across spatial scales. *Oikos*, **128**.8, pp. 1079–1091. DOI: `10.1111/oik.05968` (cit. on p. 2).

Chung, N. C., Miasojedow, B., Startek, M., & Gambin, A. (2019). Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. **20**.S15. DOI: `10.1186/s12859-019-3118-5` (cit. on pp. 3, 7).

Colwell, R. K. & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **345**.1311, pp. 101–118 (cit. on p. 11).

Cooley, J. W. & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. **19**.90, pp. 297–301. DOI: `10.1090/s0025-5718-1965-0178586-1` (cit. on p. 11).

Cressie, N., Davis, A. S., & Folks, J. L. (1981). The moment-generating function and negative integer moments. *The American Statistician*, **35**.3, pp. 148–150. DOI: `10.1080/00031305.1981.10479334` (cit. on p. 13).

Crowther, T. W., Glick, H. B., Covey, K. R., Bettigole, C., Maynard, D. S., Thomas, S. M., Smith, J. R., Hintler, G., Duguid, M. C., Amatulli, G., Tuanmu, M.-N., Jetz, W., Salas, C., Stam, C., Piotto, D., Tavani, R., Green, S., Bruce, G., Williams, S. J., Wiser, S. K., Huber, M. O., Hengeveld, G. M., Nabuurs, G.-J., Tikhonova, E., Borchardt, P., Li, C.-F., Powrie, L. W., Fischer, M., Hemp, A., Homeier, J., Cho, P., Vibrans, A. C., Umunay, P. M., Piao, S. L., Rowe, C. W., Ashton, M. S., Crane, P. R., & Bradford, M. A. (2015). Mapping tree density at a global scale. *Nature*, **525**.7568, pp. 201–205. DOI: `10.1038/nature14967` (cit. on p. 2).

De Bie, T., De Meester, L., Brendonck, L., Martens, K., Goddeeris, B., Ercken, D., Hampel, H., Denys, L., Vanhecke, L., Gucht, K. Van der, *et al.* (2012). Body size and dispersal mode as key traits determining metacommunity structure of aquatic organisms. *Ecology Letters*, **15**.7, pp. 740–747 (cit. on p. 3).

Elith, J. & Leathwick, J. R. (2009). Species Distribution Models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**.1, pp. 677–697. DOI: `10.1146/annurev.ecolsys.110308.120159` (cit. on pp. 3, 10).

Godoy, O., Gómez-Aparicio, L., Matías, L., Pérez-Ramos, I. M., & Allan, E. (2020). An excess of niche differences maximizes ecosystem functioning. *Nature Communications*, **11**.1. DOI: `10.1038/s41467-020-17960-5` (cit. on p. 12).

Gotelli, N. J. & Ulrich, W. (2011). Statistical challenges in null model analysis. *Oikos*, **121**.2, pp. 171–180. DOI: `10.1111/j.1600-0706.2011.20301.x` (cit. on p. 3).

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press (cit. on pp. 3, 10).

Harrison, S., Ross, S. J., & Lawton, J. H. (1992). Beta diversity on geographic gradients in britain. *Journal of Animal Ecology*, **61**.1, p. 151. DOI: `10.2307/5518` (cit. on p. 12).

Hautier, Y., Isbell, F., Borer, E. T., Seabloom, E. W., Harpole, W. S., Lind, E. M., MacDougall, A. S., Stevens, C. J., Adler, P. B., Alberti, J., Bakker, J. D., Brudvig, L. A., Buckley, Y. M., Cadotte, M., Caldeira, M. C., Chaneton, E. J., Chu, C., Daleo, P., Dickman, C. R., Dwyer, J. M., Eskelinen, A., Fay, P. A., Firn, J., Hagenah, N., Hillebrand, H., Iribarne, O., Kirkman, K. P., Knops, J. M. H., Pierre, K. J. L., McCulley, R. L., Morgan, J. W., Pärtel, M., Pascual, J., Price, J. N., Prober, S. M., Risch, A. C., Sankaran, M., Schuetz, M., Standish, R. J., Virtanen, R., Wardle, G. M., Yahdjian, L., & Hector, A. (2017). Local loss and spatial homogenization of plant diversity reduce ecosystem multifunctionality. *Nature Ecology & Evolution*, **2**.1, pp. 50–56. DOI: `10.1038/s41559-017-0395-0` (cit. on p. 2).

Heideman, M., Johnson, D., & Burrus, C. (1984). Gauss and the history of the fast Fourier transform. **1**.4, pp. 14–21. DOI: `10.1109/massp.1984.1162257` (cit. on p. 11).

Heino, J., Grönroos, M., Ilmonen, J., Karhu, T., Niva, M., & Paasivirta, L. (2013). Environmental heterogeneity and β diversity of stream macroinvertebrate communities at intermediate spatial scales. *Freshwater Science*, **32**.1, pp. 142–154. DOI: `10.1899/12-083.1` (cit. on p. 3).

Heino, J., Melo, A. S., & Bini, L. M. (2014). Reconceptualising the beta diversity-environmental heterogeneity relationship in running water systems. *Freshwater Biology*, **60**.2, pp. 223–235. DOI: `10.1111/fwb.12502` (cit. on p. 3).

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**.15, pp. 1965–1978. DOI: `10.1002/joc.1276` (cit. on p. 32).

Hoover, E. M. (1936). The Measurement of Industrial Localization. *The Review of Economics and Statistics*, **18**.4, p. 162. DOI: `10.2307/1927875` (cit. on p. 7).

Hubbard, J. (1959). Calculation of partition functions. *Physical Review Letters*, **3**.2, pp. 77–78. DOI: `10.1103/physrevlett.3.77` (cit. on p. 29).

Hui, C. & McGeoch, M. A. (2014). Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. *The American Naturalist*, **184**.5, pp. 684–694. DOI: `10.1086/678125` (cit. on p. 3).

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, **44**, pp. 223–270 (cit. on pp. 3, 5).

— (1912). The Distribution of the flora in the alpine zone. *New Phytologist*, **11**.2, pp. 37–50. DOI: `10.1111/j.1469-8137.1912.tb05611.x` (cit. on pp. 3, 5).

Jankowski, J. E., Ciecka, A. L., Meyer, N. Y., & Rabenold, K. N. (2009). Beta diversity along environmental gradients: implications of habitat specialization in tropical montane landscapes. *Journal of Animal Ecology*, **78**.2, pp. 315–327. DOI: `10.1111/j.1365-2656.2008.01487.x` (cit. on p. 3).

Jost, L. (2006). Entropy and diversity. *Oikos*, **113**.2, pp. 363–375. DOI: `10.1111/j.2006.0030-1299.14714.x` (cit. on p. 28).

— (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, **88**.10, pp. 2427–2439. DOI: `10.1890/06-1736.1` (cit. on p. 28).

Karp, D. S., Frishkoff, L. O., Echeverri, A., Zook, J., Juárez, P., & Chan, K. M. A. (2017). Agriculture erases climate-driven β-diversity in Neotropical bird communities. *Global Change Biology*, **24**.1, pp. 338–349. DOI: `10.1111/gcb.13821` (cit. on p. 2).

Keil, P., Wiegand, T., Tóth, A. B., McGlinn, D. J., & Chase, J. M. (2021). Measurement and analysis of interspecific spatial associations as a facet of biodiversity. *Ecological Monographs*, **91**.3. DOI: `10.1002/ecm.1452` (cit. on p. 3).

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence-absence data. *Journal of Animal Ecology*, **72**.3, pp. 367–382. DOI: `10.1046/j.1365-2656.2003.00710.x` (cit. on p. 2).

Lange, K. (2010). *Applied Probability*. 2nd ed. Springer Texts in Statistics. Springer (cit. on p. 13).

Legendre, P. (2019). A temporal beta-diversity index to identify sites that have changed in exceptional ways in space–time surveys. *Ecology and Evolution*, **9**.6, pp. 3500–3514. DOI: `10.1002/ece3.4984` (cit. on pp. 2, 3, 10).

Leinster, T. (2021). *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press (cit. on p. 28).

Loiseau, N., Legras, G., Kulbicki, M., Mérigot, B., Harmelin-Vivien, M., Mazouni, N., Galzin, R., & Gaertner, J. (2016). Multi-component β-diversity approach reveals conservation dilemma between species and functions of coral reef fishes. *Journal of Biogeography*, **44**.3, pp. 537–547. DOI: `10.1111/jbi.12844` (cit. on p. 12).

Loreau, M. (2000). Are communities saturated? On the relationship between alpha, beta and gamma diversity. *Ecology Letters*, **3**.2, pp. 73–76. DOI: `10.1046/j.1461-0248.2000.00127.x` (cit. on p. 2).

Lu, M. (2021). Complex relationships between beta diversity and dispersal in meta-community models. *Ecography*, **44**.12, pp. 1769–1780. DOI: `10.1111/ecog.05937` (cit. on pp. 2, 7, 11).

Lu, M., Vasseur, D., & Jetz, W. (2019). Beta diversity patterns derived from island biogeography theory. *The American Naturalist*, **194**.3, E52–E65. DOI: `10.1086/704181` (cit. on pp. 2, 6, 7, 11, 22).

MacArthur, R. H. & Wilson, E. O. (1963). An equilibrium theory of insular zoogeography. *Evolution*, **17**.4, p. 373. DOI: `10.2307/2407089` (cit. on pp. 3, 5).

MacArthur, R. H. & Wilson, E. O. (1967). *The theory of island biogeography*. **1**. Princeton university press (cit. on p. 3).

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2018). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. 2nd ed. Elsevier (cit. on p. 12).

Maggini, R., Lehmann, A., Zbinden, N., Zimmermann, N. E., Bolliger, J., Schröder, B., Foppen, R., Schmid, H., Beniston, M., & Jenni, L. (2014). Assessing species vulnerability to climate and land use change: the case of the Swiss breeding birds. *Diversity and Distributions*, **20**.6, pp. 708–719. DOI: `10.1111/ddi.12207` (cit. on p. 10).

Magurran, A. E., Dornelas, M., Moyes, F., & Henderson, P. A. (2019). Temporal β diversity–A macroecological perspective. *Global Ecology and Biogeography*, **28**.12, pp. 1949–1960. DOI: `10.1111/geb.13026` (cit. on pp. 3, 10).

Marshall, A. W., Olkin, I., & Arnold, B. C. (1979). *Inequalities: theory of majorization and its applications*. **143**. Springer (cit. on p. 28).

Maynard, D. S., Crowther, T. W., & Bradford, M. A. (2017). Competitive network determines the direction of the diversity–function relationship. *Proceedings of the National Academy of Sciences*, **114**.43, pp. 11464–11469. DOI: `10.1073/pnas.1712211114` (cit. on p. 2).

McKinney, M. L. (2006). Urbanization as a major cause of biotic homogenization. *Biological Conservation*, **127**.3, pp. 247–260. DOI: `10.1016/j.biocon.2005.09.005` (cit. on p. 2).

McVinish, R. & Lester, R. J. G. (2020). Measuring aggregation in parasite populations. *Journal of The Royal Society Interface*, **17**.165, p. 20190886. DOI: `10.1098/rsif.2019.0886` (cit. on p. 27).

Mori, A. S., Isbell, F., & Seidl, R. (2018). β-diversity, community assembly, and ecosystem functioning. *Trends in Ecology & Evolution*, **33**.7, pp. 549–564. DOI: `10.1016/j.tree.2018.04.012` (cit. on pp. 2, 12).

Mouquet, N. & Loreau, M. (2003). Community Patterns in Source-Sink Metacommunities. *The American Naturalist*, **162**.5, pp. 544–557. DOI: `10.1086/378857` (cit. on p. 2).

Myers, J. A., Chase, J. M., Crandall, R. M., & Jiménez, I. (2015). Disturbance alters beta-diversity but not the relative importance of community assembly mechanisms. *Journal of Ecology*, **103**.5, pp. 1291–1299. DOI: `10.1111/1365-2745.12436` (cit. on p. 2).

Ojima, M. N. & Jiang, L. (2016). Interactive effects of disturbance and dispersal on community assembly. *Oikos*, **126**.5, pp. 682–691. DOI: `10.1111/oik.03265` (cit. on p. 2).

Olden, J. D., Comte, L., & Giam, X. (2018). The Homogocene: a research prospectus for the study of biotic homogenisation. *NeoBiota*, **37**, pp. 23–36. DOI: `10.3897/neobiota.37.22552` (cit. on p. 2).

Olden, J. D. & Poff, N. L. (2003). Toward a Mechanistic Understanding and Prediction of Biotic Homogenization. *The American Naturalist*, **162**.4, pp. 442–460. DOI: `10.1086/378212` (cit. on p. 2).

Olden, J. D. & Rooney, T. P. (2006). On defining and quantifying biotic homogenization. *Global Ecology and Biogeography*, **15**.2, pp. 113–120. DOI: `10.1111/j.1466-822x.2006.00214.x` (cit. on p. 2).

Ontiveros, V. J., Capitán, J. A., Casamayor, E. O., & Alonso, D. (2021). The characteristic time of ecological communities. *Ecology*, **102**.2. DOI: `10.1002/ecy.3247` (cit. on pp. 3, 6, 7, 25).

Palacio, F. X., Callaghan, C. T., Cardoso, P., Hudgins, E. J., Jarzyna, M. A., Ottaviani, G., Riva, F., Graco-Roza, C., Shirey, V., & Mammola, S. (2022). A protocol for reproducible functional diversity analyses. *Ecography*, **2022**.11. DOI: `10.1111/ecog.06287` (cit. on p. 12).

Pärtel, M., Szava-Kovats, R., & Zobel, M. (2011). Dark diversity: shedding light on absent species. *Trends in Ecology & Evolution*, **26**.3, pp. 124–128. DOI: `10.1016/j.tree.2010.12.004` (cit. on p. 11).

Pasinelli, G. (2007). "Nest site selection in middle and great spotted woodpeckers *Dendrocopos medius* and *D. major*: implications for forest management and conservation". *Topics in Biodiversity and Conservation*. Springer Netherlands, pp. 457–472. DOI: `10.1007/978-1-4020-6320-6_30` (cit. on p. 10).

Peres-Neto, P. R., Olden, J. D., & Jackson, D. A. (2001). Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**.1, pp. 110–120. DOI: `10.1034/j.1600-0706.2001.930112.x` (cit. on p. 3).

Pilowsky, J. A., Colwell, R. K., Rahbek, C., & Fordham, D. A. (2022). Process-explicit models reveal the structure and dynamics of biodiversity patterns. *Science Advances*, **8**.31. DOI: `10.1126/sciadv.abj2271` (cit. on p. 12).

Plotkin, J. B. & Muller-Landau, H. C. (2002). Sampling the species composition of a landscape. *Ecology*, **83**.12, pp. 3344–3356 (cit. on p. 11).

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*, **36**.5, pp. 391–401. DOI: `10.1016/j.tree.2021.01.002` (cit. on p. 2).

Pomara, L. Y., Ruokolainen, K., & Young, K. R. (2013). Avian species composition across the Amazon River: the roles of dispersal limitation and environmental heterogeneity. *Journal of Biogeography*, **41**.4, pp. 784–796. DOI: `10.1111/jbi.12247` (cit. on p. 3).

Pons, J.-M., Olioso, G., Cruaud, C., & Fuchs, J. (2010). Phylogeography of the Eurasian green woodpecker (*Picus viridis*). *Journal of Biogeography*, **38**.2, pp. 311–325. DOI: `10.1111/j.1365-2699.2010.02401.x` (cit. on p. 10).

Powell, K. I., Chase, J. M., & Knight, T. M. (2013). Invasive plants have scale-dependent effects on diversity by altering species-area relationships. *Science*, **339**.6117, pp. 316–318. DOI: `10.1126/science.1226817` (cit. on p. 2).

Real, R., Barbosa, A. M., & Bull, J. W. (2016). Species distributions, quantum theory, and the enhancement of biodiversity measures. *Systematic Biology*, syw072. DOI: `10.1093/sysbio/syw072` (cit. on p. 5).

Real, R. & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, **45**.3, pp. 380–385. DOI: `10.1093/sysbio/45.3.380` (cit. on pp. 3, 12).

Routledge, R. D. (1983). Evenness indices: are any admissible? *Oikos*, **40**.1, p. 149. DOI: `10.2307/3544211` (cit. on p. 28).

Royan, A., Reynolds, S. J., Hannah, D. M., Prudhomme, C., Noble, D. G., & Sadler, J. P. (2015). Shared environmental responses drive co-occurrence patterns in river bird communities. *Ecography*, **39**.8, pp. 733–742. DOI: `10.1111/ecog.01703` (cit. on p. 3).

Schmid, H., M. Kestenholz, Knaus, P., L. Rey, & Sattler, T. (2018). The state of birds in Switzerland: special issue on the breeding bird atlas 2013–2016. en. DOI: `10.13140/RG.2.2.17872.15366` (cit. on pp. 10, 32).

Schmid, H., Posse, B., & Vogelwarte, S. S. (1998). *Schweizer Brutvogelatlas: Verbreitung der Brutvögel in der Schweiz und im Fürstentum Liechtenstein 1993-1996*. Schweizerische Vogelwarte (cit. on pp. 10, 32).

Shinohara, N., Hongo, Y., Ichinokawa, M., Nishijima, S., Sawayama, S., Kurogi, H., Uto, Y., Mita, H., Ishii, M., Kusano, A., & Akimoto, S. (2022). Similar fish species composition despite larger environmental heterogeneity during severe hypoxia in a coastal ecosystem. *Ecology and Evolution*, **12**.5. DOI: `10.1002/ece3.8884` (cit. on p. 3).

Socolar, J. B., Gilroy, J. J., Kunin, W. E., & Edwards, D. P. (2016). How should beta-diversity inform biodiversity conservation? *Trends in Ecology & Evolution*, **31**.1, pp. 67–80. DOI: `10.1016/j.tree.2015.11.005` (cit. on pp. 2, 12).

Soininen, J., Lennon, J. J., & Hillebrand, H. (2007). A multivariate analysis of beta diversity across organisms and environments. *Ecology*, **88**.11, pp. 2830–2838. DOI: `10.1890/06-1730.1` (cit. on p. 3).

Tatsumi, S., Iritani, R., & Cadotte, M. W. (2021). Temporal changes in spatial variation: partitioning the extinction and colonisation components of beta diversity. *Ecology Letters*, **24**.5, pp. 1063–1072. DOI: `10.1111/ele.13720` (cit. on p. 11).

Thompson, P. L., Guzman, L. M., Meester, L. D., Horváth, Z., Ptacnik, R., Vanschoenwinkel, B., Viana, D. S., & Chase, J. M. (2020). A process-based metacommunity framework linking local and regional scale community ecology. **23**.9, pp. 1314–1329. DOI: `10.1111/ele.13568` (cit. on p. 2).

Van Lint, J. H. & Wilson, R. M. (2001). *A course in combinatorics*. Cambridge University Press (cit. on pp. 6, 23).

Vannette, R. L. & Fukami, T. (2017). Dispersal enhances beta diversity in nectar microbes. *Ecology Letters*, **20**.7, pp. 901–910. DOI: `10.1111/ele.12787` (cit. on p. 2).

Veech, J. A. (2012). A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, **22**.2, pp. 252–260. DOI: `10.1111/j.1466-8238.2012.00789.x` (cit. on p. 3).

Veech, J. A. & Crist, T. O. (2007). Habitat and climate heterogeneity maintain beta-diversity of birds among landscapes within ecoregions. *Global Ecology and Biogeography*, **16**.5, pp. 650–656. DOI: `10.1111/j.1466-8238.2007.00315.x` (cit. on pp. 2, 3).

Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, **85**.2, pp. 183–206. DOI: `10.1086/652373` (cit. on p. 2).

Vellend, M., Verheyen, K., Flinn, K. M., Jacquemyn, H., Kolb, A., Calster, H. van, Peterken, G., Graae, B. J., Bellemare, J., Honnay, O., Brunet, J., Wulf, M., Gerhardt, F., & Hermy, M. (2007). Homogenization of forest plant communities and weakening of species? environment relationships via agricultural land use. *Journal of Ecology*, **95**.3, pp. 565–573. DOI: `10.1111/j.1365-2745.2007.01233.x` (cit. on p. 2).

Wang, S., Loreau, M., Mazancourt, C., Isbell, F., Beierkuhnlein, C., Connolly, J., Deutschman, D. H., Doležal, J., Eisenhauer, N., Hector, A., Jentsch, A., Kreyling, J., Lanta, V., Lepš, J., Polley, H. W., Reich, P. B., Ruijven, J., Schmid, B., Tilman, D., Wilsey, B., & Craven, D. (2021). Biotic homogenization destabilizes ecosystem functioning by decreasing spatial asynchrony. *Ecology*, **102**.6. DOI: 10.1002/ecy.3332 (cit. on p. 2).

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, **21**.2-3, pp. 213–251. DOI: 10.2307/1218190 (cit. on p. 2).

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, **50**.3, pp. 296–302 (cit. on p. 11).

Zorzal-Almeida, S., Bini, L. M., & Bicudo, D. C. (2017). Beta diversity of diatoms is driven by environmental heterogeneity, spatial extent and productivity. *Hydrobiologia*, **800**.1, pp. 7–16. DOI: 10.1007/s10750-017-3117-3 (cit. on p. 3).

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., Yates, K. L., Zimmermann, N. E., & Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, **43**.9, pp. 1261–1277. DOI: 10.1111/ecog.04960 (cit. on pp. 3, 10).

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2019a). *Data from: Testing species assemblage predictions from stacked and joint species distribution models*. en. DOI: 10.5061/DRYAD.K88V330 (cit. on p. 32).

— (2019b). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, **47**.1, pp. 101–113. DOI: 10.1111/jbi.13608 (cit. on pp. 10, 32).

Notation

- $\Omega := \{ 0, 1 \}$
- S, the species pool size, defined as the number of elements $\{ i \}$ such that $\{ \max_{j=1,2} p_{i,j} > 0 \}$
- $x_{i,j} \in \Omega$: incidence
- $\overline{x_{i,j}} \in \Omega$: logical negation, i.e., $\overline{x_{i,j}} = 1 - x_{i,j}$
- $p_{i,j}$: probability that $x_{i,j} = 1$
- $a_{i,j}$: probability that $x_{i,j} = 0$
- $\mathbf{X} \in \Omega^S \otimes \Omega^2$: Incidence table of size S rows and N columns
- $\mathbf{x}_{\circ,j} := \left(x_{1,j}, ..., x_{S,j} \right)^\top$ $\left(\in \Omega^S \right)$, referred to as "local compositional profile"
- $\mathbf{x}_{i,\circ} := \left(x_{i,1}, ..., x_{i,N} \right)$ $\left(\in \Omega^N \right)$, referred to as "species occurrence profile"
- $P_{\mathbf{X}}$: Probability that the incidence table $\mathbf{X}$ realizes
- $\langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \rangle := \sum_{i=1}^{S} x_{i,1} x_{i,2}$, or the inner product between local compositional profiles to count the number of common species
- $\left| \mathbf{x}_{\circ,j} \right| := \langle \mathbf{x}_{\circ,j}, \mathbf{x}_{\circ,j} \rangle$: the total number of species present in a site j
- $\mathbf{X} = \left(\mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \right)$ as we consider only two sites.

Appendix A Expectation of Jaccard dissimilarity

Definition of Jaccard dissimilarity

We write $\beta^{\mathrm{J}}_{\mathbf{X}}$ for the Jaccard dissimilarity measure for a table $\mathbf{X}$, defined by:

$$\beta^{\mathrm{J}}_{\mathbf{X}} := \frac{\left| \mathbf{x}_{\circ,1} \right| + \left| \mathbf{x}_{\circ,2} \right| - 2 \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \rangle}{\left| \mathbf{x}_{\circ,1} \right| + \left| \mathbf{x}_{\circ,2} \right| - \langle \mathbf{x}_{\circ,1}, \mathbf{x}_{\circ,2} \rangle}. \tag{A7}$$

For $\mathbf{X} = \mathbf{O}$ (zero-matrix), we define $\beta^{\mathrm{J}}_{\mathbf{O}} := 0$, which follows from two facts: (i) two all-zero vectors are (or axiomatically should be) completely similar, and (ii) the nullification of the denominator (which is always larger or equal to the numerator) should imply the nullification of the numerator (which is smaller or at most equal) as well. To avoid confusion, we suppose that numerator being zero implies the Jaccard dissimilarity be zero (otherwise resulting in erroneous calculations). It makes sense to exclude the zero-matrix, because zero-matrix indicates that there is no species in the landscape. Therefore we will focus on the conditional expectation.

Step 1: express the Jaccard dissimilarity as an integral

We note that any fraction of ν_1 to ν_2 (≥ 0) has a form of integration:

$$\frac{\nu_1}{\nu_2} \equiv \int_0^{+\infty} \frac{\partial}{\partial \xi} \exp(\nu_1 \xi - \nu_2 \theta) \, \mathrm{d}\theta \bigg|_{\xi=0}, \tag{A8}$$

which yields:

$$\beta^{\mathrm{J}}_{\mathbf{X}} \equiv \int_0^{\infty} \frac{\partial}{\partial \xi} \exp\left(\xi \sum_{i=1}^{S} \left(x_{i,1} + x_{i,2} - 2 x_{i,1} x_{i,2} \right) \right) \exp\left(-\theta \sum_{i=1}^{S} \left(x_{i,1} + x_{i,2} - x_{i,1} x_{i,2} \right) \right) \mathrm{d}\theta \bigg|_{\xi=0}, \tag{A9}$$

where we assign that we do not interchange the integral with the derivative unless otherwise stated, in order to remind that the integral should be defined as zero whenever the numerator is zero. We compute the expectation of $\beta^{\mathrm{J}}_{\mathbf{X}}$ (which is a stochastic variable) over the distribution $P_{\mathbf{X}}$.

Step 2: Independence yields product

Assuming the species independence, the probability that a given incidence table $\mathbf{X}$ is observed is given by:

$$P_{\mathbf{X}} = \prod_{i=1}^{S} \prod_{j=1}^{2} p_{i,j}^{x_{i,j}} a_{i,j}^{\overline{x_{i,j}}}, \tag{A10}$$

as the incidence probabilities are independent across species. We can then see the following facts:

$$\begin{aligned} \sum_{\mathbf{X}} P_{\mathbf{X}} \prod_{i=1}^{S} (\bullet) &\equiv \prod_{i=1}^{S} \sum_{\mathbf{x}_{i,\circ} \in \Omega^2} p_{i,1}^{x_{i,1}} p_{i,2}^{x_{i,2}} a_{i,1}^{1-x_{i,1}} a_{i,2}^{1-x_{i,2}} (\bullet) \\ \exp\left(\sum_{i=1}^{S} (\bullet) \right) &\equiv \prod_{i=1}^{S} \exp(\bullet) \end{aligned} \tag{A11}$$

Then we get the (unconditional) expectation as:

$$\begin{aligned} \mathbf{E}\left[\beta^{\mathrm{J}}\right] &:= \sum_{\mathbf{X}} P_{\mathbf{X}} \beta_{\mathbf{X}}^{\mathrm{J}} \\ &= \int_0^\infty \frac{\partial}{\partial \xi} \prod_{i=1}^{S} \sum_{\mathbf{x}_{i,\circ} \in \Omega^2} p_{i,1}^{x_{i,1}} p_{i,2}^{x_{i,2}} a_{i,1}^{1-x_{i,1}} a_{i,2}^{1-x_{i,2}} \exp\Big(\xi \left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2} \right) - \theta \left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2} \right) \Big) \, \mathrm{d}\theta \Big|_{\xi=0} . \end{aligned} \tag{A12}$$

Step 3: Boolean thinking

Let us evaluate the Boolean variable in the argument of exponential:

$$\xi \left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2} \right) - \theta \left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2} \right) = \begin{cases} 0, & \text{if } x_{i,1} = x_{i,2} = 0 \text{ (double-absence)} \\ -\theta, & \text{if } x_{i,1} = x_{i,2} = 1 \text{ (double-presence)} \\ \xi - \theta, & \text{otherwise (uniqueness)} \end{cases} \tag{A13}$$

Using this can allow us to expand the summation $\sum_{\mathbf{x}_{i,\circ} \in \Omega^2}$; that is:

$$\begin{aligned} &\sum_{\mathbf{x}_{i,\circ} \in \Omega^2} p_{i,1}^{x_{i,1}} p_{i,2}^{x_{i,2}} a_{i,1}^{1-x_{i,1}} a_{i,2}^{1-x_{i,2}} \exp\Big(\xi \left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2} \right) - \theta \left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2} \right) \Big) \\ &= d_{i,\circ} + \mathrm{e}^{-\theta} b_{i,\circ} + \mathrm{e}^{\xi-\theta} \left(1 - b_{i,\circ} - d_{i,\circ} \right) \end{aligned} \tag{A14}$$

for all $i \in \{1, ..., S\}$. Therefore, substituting this into Eqn (A12) results in:

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right] = \int_0^\infty \frac{\partial}{\partial \xi} \prod_{i=1}^{S} \Big(d_{i,\circ} + \mathrm{e}^{-\theta} b_{i,\circ} + \mathrm{e}^{\xi-\theta} \left(1 - b_{i,\circ} - d_{i,\circ} \right) \Big) \, \mathrm{d}\theta \Big|_{\xi=0} . \tag{A15}$$

Step 4: apply Leibniz rule

By using Leibniz rule of the derivative of a product, we can get:

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right] = \int_0^\infty \mathrm{e}^{-\theta} \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ} \right) \prod_{\ell=1, \ell \neq i}^{S} \Big(d_{\ell,\circ} + \mathrm{e}^{-\theta} \left(1 - d_{\ell,\circ} \right) \Big) \, \mathrm{d}\theta. \tag{A16}$$

By transforming the variable $z = 1 - \mathrm{e}^{-\theta}$ with $\mathrm{d}\theta = (1 - z)\,\mathrm{d}z$, we can rewrite Eqn (A16) as:

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right] = \int_0^1 \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ} \right) \prod_{\ell=1, \ell \neq i}^{S} \left(1 - z + z d_{\ell,\circ} \right) \mathrm{d}z. \tag{A17}$$

Eqn (A17) represents the general expression for the expectation of Jaccard dissimilarity provided that species incidences are uncorrelated.

Break to check: experiments

Experiment 1 | When $S = 1$, we immediately get $\mathbf{E}\left[\beta^{\mathrm{J}}\right]\Big|_{S=1} = 1 - b_{1,\circ} - d_{1,\circ}$. Thus the conditional expectation is $\left(1 - b_{1,\circ} - d_{1,\circ}\right) / \left(1 - d_{1,\circ}\right)$.

Experiment 2 | When $S = 2$,

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right]\Bigg|_{S=2} = \int_0^1 \left(1 - b_{1,\circ} - d_{1,\circ}\right)\left(1 - z + zd_{2,\circ}\right) + \left(1 - b_{2,\circ} - d_{2,\circ}\right)\left(1 - z + zd_{1,\circ}\right) \mathrm{d}z, \tag{A18}$$

which is $\left(1 - b_{1,\circ} - d_{1,\circ}\right)\left(1 - 1 / 2 + d_{2,\circ} / 2\right) + \left(1 - b_{2,\circ} - d_{2,\circ}\right)\left(1 - 1 / 2 + d_{1,\circ} / 2\right)$. Thus the conditional expectation is

$$\mathbf{E}\left[\beta^{\mathrm{J}}\Big|\gamma > 0\right]\Bigg|_{S=2} = \frac{1}{2\left(1 - d_{1,\circ}d_{2,\circ}\right)}\left(\left(1 - b_{1,\circ} - d_{1,\circ}\right)\left(1 + d_{2,\circ}\right) + \left(1 - b_{2,\circ} - d_{2,\circ}\right)\left(1 + d_{1,\circ}\right)\right) \tag{A19}$$

Experiment 3 | When all species are equal, that is when $\left(p_{i,1}, p_{i,2}\right) \equiv \left(p_{\circ,1}, p_{\circ,2}\right)$ with $p_{i,1}p_{i,2} = b$ and $a_{i,1}a_{i,2} = d$,

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right] = \int_0^1 S\left(1 - b - d\right)\left(1 - z + zd\right)^{S-1} \mathrm{d}z = \frac{1 - b - d}{1 - d} \cdot \left(1 - d^S\right), \tag{A20}$$

thus recovering Lu *et al.*'s (2019) results by dividing the RHS by $1 - d^S$ the probability that some species is present ($\gamma > 0$).

Rationale

First, notice that $1 - b_{i,\circ} - d_{i,\circ}$ represents the probability that species i is unique to one of the sites. Given that species i is unique, we aim to identify which species are present, regardless of being unique or common. For instance, given that species $i = 1$ is unique, the other species $i = 2, 3, \ldots, S$, each of which is either unique, common or double-absent, we can count the number of present species and put it in the denominator by calculating an integral of the product:

$$M_1 := \int_0^1 \prod_{\ell \geq 2}^{S} \left(1 - z + zd_{\ell,\circ}\right) \mathrm{d}z. \tag{A21}$$

Indeed, when $S = 2$, the integral results in $M_1 = \left(1 + d_{2,\circ}\right) / 2$, because with probability $d_{2,\circ}$, species 2 is absent from both sites, in which the contribution of species 1 to Jaccard dissimilarity is 1, while with probability $1 - d_{2,\circ}$, species 2 is present, in which case the contribution of species 1 to Jaccard dissimilarity is 1/2 (with species 2's contribution not counted here), thus giving the expectation of $\left(1 + d_{2,\circ}\right) / 2$. When $S = 3$, given that species i is unique, writing 00 for double-absence of species 2 and 3 and $\overline{00}$ for non double-absence of species 2 and 3,

sp 2	sp 3	probability	# present sp (incl 1)	species 1's contribution to Jaccard
00	00	$d_{2,\circ}d_{3,\circ}$	1	1
00	$\overline{00}$	$d_{2,\circ}\left(1 - d_{3,\circ}\right)$	2	1 / 2
$\overline{00}$	00	$\left(1 - d_{2,\circ}\right)d_{3,\circ}$	2	1 / 2
$\overline{00}$	$\overline{00}$	$\left(1 - d_{2,\circ}\right)\left(1 - d_{3,\circ}\right)$	3	1 / 3

(A22)

The expected contribution of species 1 to Jaccard dissimilarity, conditioned on species 1 being unique, is thus given by

$$\begin{aligned} d_{2,\circ}d_{3,\circ}\cdot 1 + d_{2,\circ}\left(1-d_{3,\circ}\right)\cdot\frac{1}{2} + \left(1-d_{2,\circ}\right)d_{3,\circ}\cdot\frac{1}{2} + \left(1-d_{2,\circ}\right)\left(1-d_{3,\circ}\right)\cdot\frac{1}{3} &= \frac{2+d_{2,\circ}+d_{3,\circ}+2d_{2,\circ}d_{3,\circ}}{6} \\ &= M_1 \end{aligned} \tag{A23}$$

where the second line results from calclation of M_1 for $S = 3$. From this reasoning, we can interpret Eqn (A17) as the sum of the conditional expectations of species' contribution to Jaccard dissimilarity.

Step 5: reach Beta function

Expanding the product in Eqn (A17) in terms of $1-z$ and z, we get:

$$\begin{aligned} \mathbf{E}\left[\beta^{\mathrm{J}}\right] &= \sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)\int_0^1\left(\prod_{\ell=1;(\ell\neq i)}^{S}\left(1-z+zd_{\ell,\circ}\right)\right)\mathrm{d}z \\ &= \sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)\int_0^1\sum_{k=1}^{S}\sum_{\mathbf{y}\in\Omega^S,\left|\mathbf{y}\right|=k,y_i=1}(1-z)^{k-1}z^{S-k}\prod_{\ell=1,\ell\neq i}^{S}d_{i,\circ}^{1-y_\ell}\,\mathrm{d}z \end{aligned} \tag{A24}$$

Using the Beta function $\mathrm{B}(k, S-k+1) := \int_0^1 (1-z)^{k-1}z^{S-k}\,\mathrm{d}z = (k-1)!(S-k)!\,/\,S!$, we can rewrite $\mathbf{E}\left[\beta^{\mathrm{J}}\right]$ as:

$$\begin{aligned} \mathbf{E}\left[\beta^{\mathrm{J}}\right] &= \sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)\int_0^1\sum_{k=1}^{S}\sum_{\mathbf{y}\in\Omega^S,\left|\mathbf{y}\right|=k,y_i=1}(1-z)^{k-1}z^{S-k}\prod_{\ell=1,\ell\neq i}^{S}d_{i,\circ}^{1-y_\ell}\,\mathrm{d}z \\ &= \sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)\sum_{k=1}^{S}\sum_{\mathbf{y}\in\Omega^S,\left|\mathbf{y}\right|=k,y_i=1}\frac{(k-1)!\,(S-k)!}{S!}\prod_{\ell=1,\ell\neq i}^{S}d_{i,\circ}^{1-y_\ell} \\ &= \frac{1}{S}\sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)\sum_{k=1}^{S}\sum_{\mathbf{y}\in\Omega^S,\left|\mathbf{y}\right|=k,y_i=1}\frac{1}{{}_{S-1}\mathrm{C}_{k-1}}\prod_{\ell=1,\ell\neq i}^{S}d_{i,\circ}^{1-y_\ell} \end{aligned} \tag{A25}$$

where ${}_{S-1}\mathrm{C}_{k-1}$ (with $k = \left|\mathbf{y}\right|$) represents the binomial coefficient, which counts the number of ways, disregarding order, that $k-1$ species can be chosen from among $S-1$ species (Van Lint & Wilson 2001). This is the exact expression of the expectation of Jaccard dissimilarity. When we consider the conditional expectation, we divide the unconditional expectation by $1-\prod_{i=1}^{S} d_{i,\circ}$. An alternative expression (displayed in the main text) is given by:

$$\mathbf{E}\left[\beta^{\mathrm{J}}\middle|\,\gamma>0\right] = \frac{\sum_{i=1}^{S}\sum_{\mathbf{y}\in\Omega^S}\frac{1}{\left|\mathbf{y}\right|}\frac{y_i\left(1-b_{i,\circ}-d_{i,\circ}\right)}{{}_S\mathrm{C}_{\left|\mathbf{y}\right|}}\prod_{\ell(\neq i)}^{S}d_{\ell,\circ}^{1-y_\ell}}{1-\prod_{k=1}^{S}d_{k,\circ}} \tag{A26}$$

which follows by rearranging the binomial coefficients.

Shortcut method using the generating function

By noticing that integration is the key, one can take a shortcut approach. Let $\tau(z_{\mathsf{b}}, z_{\mathsf{u}}, z_{\mathsf{d}}) := \prod_{i=1}^{S}\left(u_{i,\circ}z_{\mathsf{u}} + b_{i,\circ}z_{\mathsf{b}} + d_{i,\circ}z_{\mathsf{d}}\right)$ be the joint generating function of the ternary distributions for species i to be common, unique, or double-absent ($i = 1, 2, ..., S$). By expanding the polynomial we get the identity:

$$\tau(z_{\mathsf{b}}, z_{\mathsf{u}}, z_{\mathsf{d}}) \equiv \sum_{\{\,i_{\mathsf{u}}+i_{\mathsf{b}}+i_{\mathsf{d}}=S\,\}}\tau_{(i_{\mathsf{u}},i_{\mathsf{b}},i_{\mathsf{d}})}z_{\mathsf{u}}^{i_{\mathsf{u}}}z_{\mathsf{b}}^{i_{\mathsf{b}}}z_{\mathsf{d}}^{i_{\mathsf{d}}}, \tag{A27}$$

where the coefficient $\tau_{(i_\mathsf{u}, i_\mathsf{b}, i_\mathsf{d})}$ represents the probability that (i) i_u species are unique, (ii) i_b species are common, and (iii) $i_\mathsf{d} = S - i_\mathsf{u} - i_\mathsf{b}$ species are double-absent.[1] A vector $\mathbf{i} := (i_\mathsf{u}, i_\mathsf{b}, i_\mathsf{d})$ (with $i_\mathsf{u} + i_\mathsf{b} + i_\mathsf{d} = S$) therefore represents the state, or species-implicit incidence-table, of the community, with each species categorized as either unique, common, or double-absent. What we ought to compute is then:

$$\mathbf{E}\left[\beta^\mathrm{J}\right] = \sum_{\mathbf{i}} \frac{i_\mathsf{u}}{i_\mathsf{u} + i_\mathsf{b}} \tau_{\mathbf{i}}, \tag{A28}$$

which equals:

$$\sum_{\mathbf{i}} \frac{i_\mathsf{u}}{i_\mathsf{u} + i_\mathsf{b}} \tau_{\mathbf{i}} z_\mathsf{u}^{i_\mathsf{u}} z_\mathsf{b}^{i_\mathsf{b}} z_\mathsf{d}^{i_\mathsf{d}} \Bigg|_{z_\mathsf{u} = z_\mathsf{b} = z_\mathsf{d} = 1} . \tag{A29}$$

We wish to algebraically extract i_u (the numerator of Jaccard dissimilarity) and the reciprocal of $(i_\mathsf{b} + i_\mathsf{u})$ (the denominator of Jaccard dissimilarity). With the inspiration of integration, we can come up with:

$$\mathbf{E}\left[\beta^\mathrm{J}\right] = \int_0^1 \frac{\partial}{\partial z_\mathsf{u}} \sum_{\mathbf{i}} \tau_{(i_\mathsf{u}, i_\mathsf{b}, i_\mathsf{d})} z_\mathsf{u}^{i_\mathsf{u}} z_\mathsf{b}^{i_\mathsf{b}} 1^{i_\mathsf{d}} \Bigg|_{z_\mathsf{u} = z_\mathsf{b}} \mathrm{d}z_\mathsf{b} \equiv \int_0^1 \frac{\partial}{\partial z_\mathsf{u}} \tau(z_\mathsf{u}, z_\mathsf{b}, 1) \Bigg|_{z_\mathsf{u} = z_\mathsf{b}} \mathrm{d}z_\mathsf{b}. \tag{A30}$$

Using the original definition of the generating function, its derivative is given by:

$$\begin{aligned} \frac{\partial}{\partial z_\mathsf{u}} \tau(z_\mathsf{b}, z_\mathsf{u}, z_\mathsf{d}) \Bigg|_{z_\mathsf{u} = z_\mathsf{b}} &= \frac{\partial}{\partial z_\mathsf{u}} \prod_{i=1}^{S} \left(u_{i,\circ} z_\mathsf{u} + b_{i,\circ} z_\mathsf{b} + d_{i,\circ} z_\mathsf{d}\right) \Bigg|_{z_\mathsf{u} = z_\mathsf{b}, z_\mathsf{d} = 1} \\ &= \sum_{i=1}^{S} u_{i,\circ} \prod_{\ell(\neq i)} \left(u_\ell z_\mathsf{b} + b_\ell z_\mathsf{b} + d_\ell\right) \\ &= \sum_{i=1}^{S} u_{i,\circ} \prod_{\ell(\neq i)} \left(\left(1 - d_{\ell,\circ}\right) z_\mathsf{b} + d_{\ell,\circ}\right), \end{aligned} \tag{A31}$$

which thus returns us back to Eqn (A17) by integrating the last line.

Appendix B Approximations

Upper bound

Although the exact calculation of the expectation is correct, the computational speed may be too slow to be practical, especially when S is large. Therefore we consider approximating it. We will make use of the property of the bivariate function:

$$J(b, d) := \frac{1 - b - d}{1 - d}, \tag{B32}$$

which is "almost" linear, on the feasible domain $\{\, (b, d) \in [0, 1]^2 \mid \sqrt{b} + \sqrt{d} \leq 1 \,\}$.

We rewrite Eqn (A17) as:

$$\int_0^1 \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ}\right) \exp\left(\sum_{\ell=1, \ell \neq i}^{S} \log\left(1 - z + z d_{\ell,\circ}\right)\right) \mathrm{d}z. \tag{B33}$$

Using Jensen's inequality,

$$\sum_{\ell=1, \ell \neq i}^{S} \log\left(1 - z + z d_{\ell,\circ}\right) \leq (S - 1) \log\left(1 - z + z \frac{1}{S - 1} \sum_{\ell=1, \ell \neq i}^{S} d_{\ell,\circ}\right), \tag{B34}$$

---

[1] z_d really is unneeded but is incorporated for symmetry.

where the equality achieves when $d_{\ell,\circ}$s are all identical across species $i = 1, 2, \dots, S$. Now put:

$$h_{\setminus i} := \frac{1}{S-1} \sum_{\ell=1,\ell\neq i}^{S} d_{\ell,\circ}. \tag{B35}$$

Then we get:

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right] \leq \int_0^1 \sum_{i=1}^{S} \left(1 - b_{\ell,\circ} - d_{\ell,\circ}\right) \left(1 - z + z h_{\setminus i}\right)^{S-1} \mathrm{d}z = \frac{1}{S} \sum_{i=1}^{S} \frac{1 - b_{i,\circ} - d_{i,\circ}}{1 - h_{\setminus i}} \left(1 - h_{\setminus i}^{S}\right) = \beta^{\mathrm{J}}_{+}. \tag{B36}$$

RHS gives a very good approximation, because $J(b, d)$ is almost linear and only very moderately concave.

We divide both sides by $1 - \prod_{i=1}^{S} d_{i,\circ}$ to get the approximation of the conditional expectation of Jaccard dissimilarity.

Lower bound

The lower-bound approximation of the expected Jaccard dissimilarity is given by:

$$\begin{aligned}
\mathbf{E}\left[\beta^{\mathrm{J}}\right] &= \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ}\right) \int_0^1 \left(\prod_{\ell=1;\ell\neq i}^{S} \left(1 - z + z d_{\ell,\circ}\right) \right) \mathrm{d}z \\
&\geq \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ}\right) \int_0^1 \left(1 - z + z \prod_{\ell=1;\ell\neq i}^{S} d_{\ell,\circ}^{\frac{1}{S-1}} \right)^{S-1} \mathrm{d}z \\
&= \frac{1}{S} \sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ}\right) \frac{1 - \prod_{\ell=1;\ell\neq i}^{S} d_{\ell,\circ}^{\frac{S}{S-1}}}{1 - \prod_{\ell=1;\ell\neq i}^{S} d_{\ell,\circ}^{\frac{1}{S-1}}} = \beta^{\mathrm{J}}_{-},
\end{aligned} \tag{B37}$$

where the second line follows by applying the induction, and the equality holds when $d_{i,\circ}$s are all identical. [2] Dividing both sides by $1 - \prod_{i=1}^{S} d_{i,\circ}$ yields the approximation for the conditional expectation.

Heuristic approximation

The other approximation for the conditional expectation can be obtained heuristically (Ontiveros *et al.* 2021):

$$\beta^{\mathrm{J}}_{\mathrm{heur}} = \frac{\sum_{i=1}^{S} \left(1 - b_{i,\circ} - d_{i,\circ}\right)}{\sum_{i=1}^{S} \left(1 - d_{i,\circ}\right)} = J\left(\frac{1}{S} \sum_{i=1}^{S} b_{i,\circ}, \frac{1}{S} \sum_{i=1}^{S} d_{i,\circ} \right) \tag{B38}$$

which represents the expected number of unique species divided by the expected number of present species. Deriving this formula requires quite a bit of calculations, but if we notice:

$$1 - \prod_{i=1}^{S} d_{i,\circ} = - \int_0^1 \frac{\mathrm{d}}{\mathrm{d}z} \prod_{i=1}^{S} \left(1 - z + z d_{i,\circ}\right) \mathrm{d}z = \int_0^1 \sum_{i=1}^{S} \left(1 - d_{i,\circ}\right) \prod_{\ell=1;\ell\neq i}^{S} \left(1 - z + z d_{\ell,\circ}\right) \mathrm{d}z, \tag{B39}$$

---

[2] First, when $S = 2$, some convexity arguments of arithmetic and geometric means work. Second, hypothesizing that the inequality is valid for a certain $S = 2^{\ell}$, it is easy to prove that the inequality holds for $S = 2^{\ell+1}$. Finally, hypothesizing that the inequality is valid for a certain S, we prove the inequality is the case for $S - 1$, by using the binomial expansion of the product and applying the arithmetic-geometric means relation repeatedly. The equality achieves when all $d_{i,\circ}$ are equal. This completes the proof.

then we get:

$$\begin{aligned}
\mathbf{E}\left[\beta^{\mathrm{J}} \,\middle|\, \gamma > 0\right] &= \frac{\mathbf{E}\left[\beta^{\mathrm{J}}\right]}{1 - \prod_{i=1}^{S} d_{i,\circ}} \\
&= \frac{\int_0^1 \sum_{i=1}^{S} \left(1 - d_{i,\circ} - b_{i,\circ}\right) \prod_{\ell=1;\ell\neq i}^{S} \left(1 - z + z d_{\ell,\circ}\right) \mathrm{d}z}{\int_0^1 \sum_{i=1}^{S} \left(1 - d_{i,\circ}\right) \prod_{\ell=1;\ell\neq i}^{S} \left(1 - z + z d_{\ell,\circ}\right) \mathrm{d}z} \\
&= \frac{\sum_{i=1}^{S} \left(1 - d_{i,\circ} - b_{i,\circ}\right) M_i}{\sum_{i=1}^{S} \left(1 - d_{i,\circ}\right) M_i},
\end{aligned} \tag{B40}$$

where we have put:

$$M_i := \int_0^1 \prod_{\ell=1;\ell\neq i}^{S} \left(1 - z + z d_{i,\circ}\right) \mathrm{d}z \tag{B41}$$

for $i = 1, 2, \ldots, S$. If we replace the integral M_i, which shows up in both the denominator and numerator (but is multiplied by different coefficients), with (say) its average over $i = 1, 2, \ldots, S$, written as $M_* := \sum_{i=1}^{S} M_i \,/\, S$, then:

$$\mathbf{E}\left[\beta^{\mathrm{J}} \,\middle|\, \gamma > 0\right] \approx \frac{\sum_{i=1}^{S} \left(1 - d_{i,\circ} - b_{i,\circ}\right) M_*}{\sum_{i=1}^{S} \left(1 - d_{i,\circ}\right) M_*} = \frac{\sum_{i=1}^{S} \left(1 - d_{i,\circ} - b_{i,\circ}\right)}{\sum_{i=1}^{S} \left(1 - d_{i,\circ}\right)} = \beta^{\mathrm{J}}_{\mathrm{heur}}, \tag{B42}$$

which thus gives the heuristic approximation. The approximation is exact when $M_1 = \cdots = M_S$ (which is not a necessary condition). This explains why the heuristic approximation works for the conditional expectation, while not for the unconditional, and also explains why the heuristic approximation may be larger or smaller than the exact expectation depending on the variance in M_is. We may observe that M_i becomes increasingly small with S larger ($M_* = \mathcal{O}\left(\left(\sum_{i=1}^{S} d_{i,\circ}\right)^{-1}\right)$), and thus, as S increases, the contribution of the replacement (M_i with M_*) to the difference between the exact and approximated expectation becomes smaller.

The heuristic approximation may, however, be either larger or smaller than the exact conditional expectation, and thus the upper and lower bounds, (β^{J}_{+} and β^{J}_{-}), both may be also recommended.

Altogether, we obtained the three approximations:

- Approximation from below: β^{J}_{-}
- Approximation from above: β^{J}_{+}
- Heuristic approximation: $\beta^{\mathrm{J}}_{\mathrm{heur}}$

Performances

We numerically compared the accuracy and precision of the approximations, in the following procedure.

(1) Generate two vectors (with $S = 100$):

$$\mathbf{p}_{\circ,1} = \begin{pmatrix} p_{1,1} \\ p_{2,1} \\ \vdots \\ p_{S,1} \end{pmatrix}, \mathbf{p}_{\circ,2} = \begin{pmatrix} p_{1,2} \\ p_{2,2} \\ \vdots \\ p_{S,2} \end{pmatrix}, \tag{B43}$$

in which each value of $p_{i,1}$ drawn from the Beta distribution with parameters 1.2 and 1.5, and $p_{i,2}$ drawn from the Beta distribution with parameters 1.6 and 0.8.

(2) Calculate double-presence and -absence vectors:

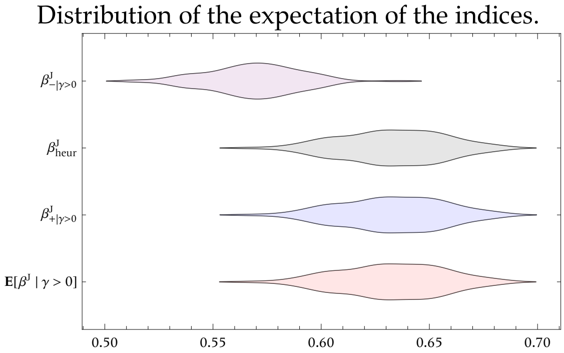
$$\mathbf{b}_{\circ,\circ} = \begin{pmatrix} b_{1,\circ} \\ b_{2,\circ} \\ \vdots \\ b_{S,\circ} \end{pmatrix} = \begin{pmatrix} p_{1,1}p_{1,2} \\ p_{2,1}p_{2,2} \\ \vdots \\ p_{S,1}p_{S,2} \end{pmatrix}, \mathbf{d}_{\circ,\circ} = \begin{pmatrix} d_{1,\circ} \\ d_{2,\circ} \\ \vdots \\ d_{S,\circ} \end{pmatrix} = \begin{pmatrix} a_{1,1}a_{1,2} \\ a_{2,1}a_{2,2} \\ \vdots \\ a_{S,1}a_{S,2} \end{pmatrix}, \tag{B44}$$

which can be numerically implemented with Hadamard product.

(3) Compute the followings:
  - $\mathbf{E}\left[\beta^{\mathrm{J}} \middle| \gamma > 0\right]$ using Eqn (A17);
  - β^{J}_{-} in a brute-force approach;
  - β^{J}_{+} in a brute-force approach;
  - $\beta^{\mathrm{J}}_{\mathrm{heur}}$ in a brute-force approach.

(4) Repeat the procedure 100 times and generated the probability distributions of those four quantities (but note that they are stochastic variables because we used the Beta distribution to generate $\{\, p_{i,j} \,\}$).

(5) We plot the distribution of the expectations of Jaccard dissimilarity measures (approximations and exact value).



SI Figure 1: We assessed the probability distributions of the exact and approximated Jaccard expectations.

Appendix C Well-definedness

Here we clarify some concepts that have been left unspecified in the main text.

Species pool

We first suppose that species i is *presentable* in the focal metacommunity if strictly $d_{i,\circ} < 1$. The collection of all presentable species is called as *species pool*. Its cardinality (i.e., the number of members of the species pool) is referred to as *species pool size* and denoted as S. By defining these, we exclude never presentable species from the species pool, $p_{i,1} = p_{i,2} = 0$.

Bounds of w and h

species difference

The species difference index is Schur-convex (Arnold 2012; McVinish & Lester 2020), thus taking the minimum zero when all species are equally likely to be present in each site; that is whenever:

$$p_{1,j} = \cdots = p_{S,j} > 0 \tag{C45}$$

for $j = 1, 2$.

To determine an upper bound (if any), we conventionally assume that for all species i, there is a minimum value of the presence probabilities $p_{i,j} \geq \varepsilon$. Then by the property of Schur-convexity of Hoover index, the largest difference (inequality) occurs when:

$$\begin{aligned} p_{1,1} = 1, \quad p_{i,1} \equiv \varepsilon > 0, {}^{\forall} i \geq 2, \\ p_{1,2} = 1, \quad p_{i,2} \equiv \varepsilon > 0, {}^{\forall} i \geq 2, \end{aligned} \tag{C46}$$

with species-wise permutation permitted for each j. Substituting this into w gives the maximum value, which reads:

$$\max\{\, w \,\} = \frac{2\,(1-\varepsilon)\,(S-1)}{S\,\big(\varepsilon\,(S-1)+1\big)} < 2. \tag{C47}$$

Spatial heterogeneity

If and only if $p_{i,1} \equiv p_{i,2}$ for each i, spatial heterogeneity h achieves the minimum of zero. If and only if $\left|p_{i,1} - p_{i,2}\right| = 1 - \varepsilon$, spatial heterogeneity h achieves the maximum of $1 - \varepsilon$.

Axioms

Species difference w has to satisfy a number of axioms in order to be a "difference" measure among species, as does h in order to measure heterogeneity between two sites. We adopt Routledge's (1983), Jost's (2006, 2007), Chao & Ricotta's (2019), and Leinster's (2021) approach to develop appropriate indices. Note that the presence probabilities do not represent abundances.

requirement 1 The first requirement for w is that, if, and only if, $p_{i,1}$s are all equal and $p_{i,2}$ are all equal, then $w = 0$.

requirement 2 Second, w is Schur-convex (or Schur-increasing; Marshall *et al.* 1979, Chapter 1).

requirement 3 The requirement for heterogeneity h is that it is a distance function between $\mathbf{p}_{\circ,1}$ and $\mathbf{p}_{\circ,2}$.

Appendix D Variance of Jaccard dissimilarity

Same method as the mean

To compute the variance, we use the identity for a pair of positive quantities $\nu_1, \nu_2 > 0$:

$$\frac{\nu_1^2}{\nu_2^2} \equiv -\int_0^{+\infty} \frac{\partial^3}{\partial\xi\partial\eta^2} \exp\left(\nu_1\eta - (\nu_2\theta + \xi\theta)\right) \mathrm{d}\theta\bigg|_{\xi=\eta=0}. \tag{D48}$$

One may preferably differentiate the quantity before integration (otherwise, erroneous calculation is possible).

For Jaccard dissimilarity, we choose $\nu_1 = \sum_{i=1}^{S}\left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2}\right)$, which represents the number of unique species, and $\nu_2 = \sum_{i=1}^{S}\left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2}\right)$, which represents the number of present species (gamma diversity). That is:

$${\beta_{\mathbf{X}}^{\mathrm{J}}}^2 = -\int_0^{+\infty} \frac{\partial^3}{\partial\xi\partial\eta^2} \exp\left(\eta\sum_{i=1}^{S}\left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2}\right)\right) \exp\left(-\theta\sum_{i=1}^{S}\left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2}\right)\right) \mathrm{d}\theta\bigg|_{\xi=\eta=0}. \tag{D49}$$

The expectation of ${\beta_{\mathbf{X}}^{\mathrm{J}}}^2$ is given by:

$$\mathbf{E}\left[{\beta^{\mathrm{J}}}^2\right] = -\int_0^{+\infty} \frac{\partial^3}{\partial\xi\partial\eta^2} \mathrm{e}^{-\xi\theta} \prod_{i=1}^{S}\sum_{\mathbf{x}_{i,\circ}} P_{\mathbf{X}} \exp\left(\eta\left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2}\right) - \theta\left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2}\right)\right) \mathrm{d}\theta\bigg|_{\xi=\eta=0} \tag{D50}$$

By evaluating the Boolean variable,

$$\eta\left(x_{i,1} + x_{i,2} - 2x_{i,1}x_{i,2}\right) - \theta\left(x_{i,1} + x_{i,2} - x_{i,1}x_{i,2}\right) = \begin{cases} 0, & \mathbf{x}_{i,\circ} = (0,0); \\ \eta - \theta, & \mathbf{x}_{i,\circ} = (0,1); \\ \eta - \theta, & \mathbf{x}_{i,\circ} = (1,0); \\ -\theta, & \mathbf{x}_{i,\circ} = (1,1); \end{cases} \tag{D51}$$

the resulting expression reads:

$$\mathbf{E}\left[{\beta^{\mathrm{J}}}^2\right] = -\int_0^{+\infty} \frac{\partial^3}{\partial\xi\partial\eta^2}\,\mathrm{e}^{-\xi\theta}\prod_{i=1}^{S}\left(d_{i,\circ}+b_{i,\circ}\mathrm{e}^{-\theta}+\left(1-d_{i,\circ}-b_{i,\circ}\right)\mathrm{e}^{\eta-\theta}\right)\mathrm{d}\theta\Bigg|_{\xi=\eta=0}. \tag{D52}$$

This is the most general expression for the second moment of the Jccard dissimilarity. For brevity we write $u_{i,\circ} := 1 - b_{i,\circ} - d_{i,\circ}$ for the probability that species i is unique (neither double-absent nor common); also, we write $\tau_\ell(\theta) := d_\ell + (1-d_\ell)\,\mathrm{e}^{-\theta}$ for the moment generating function of the probability that species ℓ is present in at least one of the sites, $1 - d_{\ell,\circ}$; write $\psi_i(\theta,\eta) := d_{i,\circ} + b_{i,\circ}\mathrm{e}^{-\theta} + u_{i,\circ}\mathrm{e}^{\eta-\theta}$, thus with $\psi_i(\theta,0) \equiv \tau_i(\theta)$.

Leibniz rule for the second η-derivatives is given by:

$$\frac{\partial^2}{\partial\eta^2}\prod_{i=1}^{S}\psi_i(\theta,\eta)\Bigg|_{\eta=0} = \left(\sum_{i=1}^{S}u_{i,\circ}\mathrm{e}^{-\theta}\prod_{\ell(\neq i)}^{S}\tau_\ell(\theta)\right) + \sum_{i=1}^{S}u_{i,\circ}\mathrm{e}^{-\theta}\sum_{\substack{k=1;\\k\neq i}}^{S}u_k\mathrm{e}^{-\theta}\prod_{\substack{\ell=1;\\\ell\neq i,k}}^{S}\tau_\ell(\theta)\,, \tag{D53}$$

using which we get:

$$\mathbf{E}\left[{\beta^{\mathrm{J}}}^2\right] = -\frac{\partial}{\partial\xi}\int_0^{+\infty}\mathrm{e}^{-\xi\theta}\left(\sum_{i=1}^{S}u_{i,\circ}\mathrm{e}^{-\theta}\prod_{\ell(\neq i)}^{S}\tau_\ell(\theta) + \sum_{i=1}^{S}u_{i,\circ}\mathrm{e}^{-\theta}\sum_{\substack{k=1;\\k\neq i}}^{S}u_k\mathrm{e}^{-\theta}\prod_{\substack{\ell=1;\\\ell\neq i,k}}^{S}\tau_\ell(\theta)\right)\mathrm{d}\theta. \tag{D54}$$

We can evaluate this integral as did we before. However, the resulting equation is heavily complicated (involving, e.g., Harmonic numbers) and computationally expensive.

Approximating variance using Hubbard-Stratonovich transformation

Here, we take a different approach to evaluate the variance. We use the identity:

$$\frac{\nu_1^2}{\nu_2^2} \equiv -\frac{\partial}{\partial\xi}\int_0^{\infty}\mathrm{e}^{-\left(\nu_2+\nu_1^2\xi\right)\theta}\,\mathrm{d}\theta\Bigg|_{\xi=0} \tag{D55}$$

for $0 \le \nu_1 \le \nu_2$, as well as the Hubbard-Stratonovich transformation (Hubbard 1959):

$$\mathrm{e}^{-\xi\theta\lambda^2} \equiv \frac{1}{\sqrt{4\pi\xi\theta}}\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}-\imath\lambda\zeta}\,\mathrm{d}\zeta \tag{D56}$$

where $\imath$ represents the imaginary unit. Combining the identities gives:

$$\begin{aligned}
\left(\beta^{\mathrm{J}}\right)^2 &= \left(\frac{\sum_{i=1}^{S}x_{i,1}+x_{i,2}-2x_{i,1}x_{i,2}}{\sum_{i=1}^{S}x_{i,1}+x_{i,2}-x_{i,1}x_{i,2}}\right)^2 \\
&\overset{\text{Eqn (D55)}}{=} -\int_0^{\infty}\exp\left(-\left(\sum_{i=1}^{S}\left(x_{i,1}+x_{i,2}-x_{i,1}x_{i,2}\right)+\xi\left(\sum_{i=1}^{S}x_{i,1}+x_{i,2}-2x_{i,1}x_{i,2}\right)^2\right)\theta\right)\mathrm{d}\theta \\
&= -\int_0^{\infty}\mathrm{d}\theta\int_{-\infty}^{\infty}\mathrm{d}\zeta\,\exp\left(-\theta\left(\sum_{i=1}^{S}\left(x_{i,1}+x_{i,2}-x_{i,1}x_{i,2}\right)\right)\right)\frac{1}{\sqrt{4\pi\xi\theta}}\exp\left(-\frac{\zeta^2}{4\xi\theta}-\imath\zeta\sum_{i=1}^{S}\left(x_{i,1}+x_{i,2}-2x_{i,1}x_{i,2}\right)\right)
\end{aligned} \tag{D57}$$

Let us evaluate the Boolean variable:

$$-\theta\left(x_{i,1}+x_{i,2}-x_{i,1}x_{i,2}\right)-\imath\zeta\left(x_{i,1}+x_{i,2}-2x_{i,1}x_{i,2}\right) = \begin{cases}0 & x_{i,1}=x_{i,2}=0;\\ -\theta-\imath\zeta & x_{i,1}+x_{i,2}=1;\\ -\theta & x_{i,1}=x_{i,2}=1;\end{cases} \tag{D58}$$

then we get:

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] = -\frac{\partial}{\partial\xi}\int_0^\infty \mathrm{d}\theta \int_{-\infty}^{\infty}\mathrm{d}\zeta \mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}}\frac{1}{\sqrt{4\pi\xi\theta}}\prod_{i=1}^{S}\left(d_{i,\circ} + \left(1 - d_{i,\circ} - b_{i,\circ}\right)\mathrm{e}^{-\theta-\imath\zeta} + b_{i,\circ}\mathrm{e}^{-\theta}\right). \tag{D59}$$

If we approximate the product as:

$$\prod_{i=1}^{S}\left(d_{i,\circ} + \left(1 - d_{i,\circ} - b_{i,\circ}\right)\mathrm{e}^{-\theta-\imath\zeta} + b_{i,\circ}\mathrm{e}^{-\theta}\right) \approx \left(d_{*,\circ} + u_{*,\circ}\mathrm{e}^{-\theta-\imath\zeta} + b_{*,\circ}\mathrm{e}^{-\theta}\right)^S, \tag{D60}$$

where the $*$-subscripted quantities are the arithmetic means, over $i \in \{1, ..., S\}$, of the corresponding quantities, i.e, $d_{*,\circ} := \frac{1}{S}\sum_{i=1}^{S} d_{i,\circ}$, $b_{*,\circ} := \frac{1}{S}\sum_{i=1}^{S} b_{i,\circ}$, and $u_{*,\circ} := 1 - d_{*,\circ} - b_{*,\circ}$, then the expected value is approximated by

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx -\frac{\partial}{\partial\xi}\int_0^\infty \mathrm{d}\theta \int_{-\infty}^{\infty}\mathrm{d}\zeta \mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}}\frac{1}{\sqrt{4\pi\xi\theta}}\left(d_{*,\circ} + u_{*,\circ}\mathrm{e}^{-\theta-\imath\zeta} + b_{*,\circ}\mathrm{e}^{-\theta}\right)^S \tag{D61}$$

evaluated at $\xi = 0$.

Interchanging the order of the derivative and the double integral, we get

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx -\int_0^\infty \mathrm{d}\theta \int_{-\infty}^{\infty}\mathrm{d}\zeta \frac{\partial}{\partial\xi}\left(\frac{1}{\sqrt{4\pi\xi\theta}}\mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}}\right)\mathrm{e}^{S\log\left(d_{*,\circ}+u_{*,\circ}\mathrm{e}^{-\theta-\imath\zeta}+b_{*,\circ}\mathrm{e}^{-\theta}\right)}. \tag{D62}$$

In the limit $\xi \to 0$, the function $\frac{\partial}{\partial\xi}\left(\frac{1}{\sqrt{4\pi\xi\theta}}\mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}}\right) = \frac{\zeta^2-2\xi\theta}{8\sqrt{\pi}(\xi\theta)^{5/2}}\mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}}$ is very peaked about $\zeta = 0$. Therefore, we expect the integrand to be nicely approximated if we substitute the logarithm by its series expansion about $\zeta = 0$,

$$\log\left(\kappa_1 + \kappa_2\mathrm{e}^{-\imath\zeta}\right) \approx \log(\kappa_1 + \kappa_2) - \imath\frac{b\zeta}{\kappa_1 + \kappa_2} - \frac{\kappa_1\kappa_2\zeta^2}{2(\kappa_1 + \kappa_2)^2} + \mathcal{O}(\zeta^3), \tag{D63}$$

with $\kappa_1 := d_{*,\circ} + b_{*,\circ}\mathrm{e}^{-\theta}$ and $\kappa_2 := u_{*,\circ}\mathrm{e}^{-\theta}$. Inserting this second approximation into Eqn (D62) we get

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx -\int_0^\infty \mathrm{d}\theta\left(d_{*,\circ} + (1 - d_{*,\circ})\mathrm{e}^{-\theta}\right)^S\frac{\partial}{\partial\xi}\frac{1}{\sqrt{4\pi\xi\theta}}\int_{-\infty}^{\infty}\mathrm{d}\zeta \mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}-\imath\frac{S\kappa_2\zeta}{\kappa_1+\kappa_2}-\frac{S\kappa_1\kappa_2\zeta^2}{2(\kappa_1+\kappa_2)^2}}, \tag{D64}$$

which, again, has to be evaluated at $\xi = 0$. The integral over ζ can be evaluated as

$$\frac{1}{\sqrt{4\pi\xi\theta}}\int_{-\infty}^{\infty}\mathrm{d}\zeta \mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}-\imath\frac{S\kappa_2\zeta}{\kappa_1+\kappa_2}-\frac{S\kappa_1\kappa_2\zeta^2}{2(\kappa_1+\kappa_2)^2}} = \frac{\kappa_1 + \kappa_2}{\sqrt{(\kappa_1 + \kappa_2)^2 + 2S\kappa_1\kappa_2\xi\theta}}\mathrm{e}^{-\frac{\xi\theta(\kappa_2 S)^2}{(\kappa_1+\kappa_2)^2+2S\kappa_1\kappa_2\xi\theta}}. \tag{D65}$$

Now, we can take the derivative with respect to ξ and evaluate it at $\xi = 0$ to get

$$\left.\frac{\partial}{\partial\xi}\frac{1}{\sqrt{4\pi\xi\theta}}\int_{-\infty}^{\infty}\mathrm{d}\zeta \mathrm{e}^{-\frac{\zeta^2}{4\xi\theta}-\imath\frac{S\kappa_2\zeta}{\kappa_1+\kappa_2}-\frac{S\kappa_1\kappa_2\zeta^2}{2(\kappa_1+\kappa_2)^2}}\right|_{\xi=0} = -\frac{S\kappa_2\theta(\kappa_1 + \kappa_2 S)}{(\kappa_1 + \kappa_2)^2}. \tag{D66}$$

Therefore, inserting this expression into Eqn (D64) and replacing κ_1 and κ_2 by their expressions in terms of $d_{*,\circ}$, $b_{*,\circ}$, $u_{*,\circ}$, and θ, we obtain

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx Su_{*,\circ}\int_0^\infty \mathrm{d}\theta \mathrm{e}^{-\theta}\left(d_{*,\circ} + b_{*,\circ}\mathrm{e}^{-\theta} + Su_{*,\circ}\mathrm{e}^{-\theta}\right)\left(d_{*,\circ} + (1 - d_{*,\circ})\mathrm{e}^{-\theta}\right)^{S-2}\theta. \tag{D67}$$

Changing to the variable $z = e^{-\theta}$ yields

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx Su_{*,\circ}\int_0^1 \mathrm{d}z(-\log z)\left(d_{*,\circ} + b_{*,\circ}z + Su_{*,\circ}z\right)\left(d_{*,\circ} + (1 - d_{*,\circ})z\right)^{S-2}. \tag{D68}$$

We now use the binomial expansion $\left(d_{*,\circ}+(1-d_{*,\circ})z\right)^{S-2} = d_{*,\circ}^{S-2}\sum_{k=0}^{S-2}\binom{S-2}{k}\left(\frac{(1-d_{*,\circ})}{d_{*,\circ}}z\right)^{k}$ to get

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx S u_{*,\circ} d_{*,\circ}^{S-2} \sum_{k=0}^{S-2} \binom{S-2}{k} \left(\frac{1-d_{*,\circ}}{d_{*,\circ}}\right)^{k} \int_0^1 \mathrm{d}z(-\log z)\left(d_{*,\circ}+b_{*,\circ}z+S u_{*,\circ}z\right)z^k, \tag{D69}$$

which, upon evaluation of the integral, yields

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx S u_{*,\circ} d_{*,\circ}^{S-2} \sum_{k=0}^{S-2} \binom{S-2}{k} \left(\frac{1-d_{*,\circ}}{d_{*,\circ}}\right)^{k} \left(\frac{d_{*,\circ}}{(k+1)^2}+\frac{b_{*,\circ}+S u_{*,\circ}}{(k+2)^2}\right). \tag{D70}$$

The sum above can be expressed in terms of generalized hypergeometric functions ${}_pF_q\left(\{A\},\{B\};Z\right)$ as

$$\begin{aligned}\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx{}& \frac{u_{*,\circ}\left(b_{*,\circ}+S u_{*,\circ}\right)\left(1-d_{*,\circ}^{S}\right)}{(S-1)\left(1-d_{*,\circ}\right)^2} - \frac{S u_{*,\circ} d_{*,\circ}^{S-1}\left(b_{*,\circ}+S u_{*,\circ}\right)}{(S-1)\left(1-d_{*,\circ}\right)}\,{}_3F_2\left(\{1,1,1-S\},\{2,2\};1-\frac{1}{d_{*,\circ}}\right)\\ &+ S u_{*,\circ} d_{*,\circ}^{S-1}\,{}_3F_2\left(\{1,1,2-S\},\{2,2\};1-\frac{1}{d_{*,\circ}}\right).\end{aligned} \tag{D71}$$

As a consequence, we find the following approximation for the variance,

$$\begin{aligned}\mathbf{V}\left[\beta^{\mathrm{J}}\right] \approx{}& \frac{u_{*,\circ}\left(b_{*,\circ}+S u_{*,\circ}\right)\left(1-d_{*,\circ}^{S}\right)}{(S-1)\left(1-d_{*,\circ}\right)^2} - \frac{S u_{*,\circ} d_{*,\circ}^{S-1}\left(b_{*,\circ}+S u_{*,\circ}\right)}{(S-1)\left(1-d_{*,\circ}\right)}\,{}_3F_2\left(\{1,1,1-S\},\{2,2\};1-\frac{1}{d_{*,\circ}}\right)\\ &+ S u_{*,\circ} d_{*,\circ}^{S-1}\,{}_3F_2\left(\{1,1,2-S\},\{2,2\};1-\frac{1}{d_{*,\circ}}\right) - \left(\frac{u_{*,\circ}}{1-d_{*,\circ}}\right)^2,\end{aligned} \tag{D72}$$

where we have approximated the expectation $\mathbf{E}\left[\beta^{\mathrm{J}}\right]^2$ with the square of our heuristic approximation,

$$\mathbf{E}\left[\beta^{\mathrm{J}}\right]^2 \approx \left(\beta^{\mathrm{J}}_{\mathrm{heur}}\right)^2 = \left(\frac{\sum_{i=1}^{S}\left(1-b_{i,\circ}-d_{i,\circ}\right)}{\sum_{i=1}^{S}\left(1-d_{i,\circ}\right)}\right)^2 = \left(\frac{u_{*,\circ}}{1-d_{*,\circ}}\right)^2. \tag{D73}$$

The analytical approximation obtained in Eqn (D72) yields always averaged standard deviation relative errors less than 10%. In most of the cases relative errors for the standard deviation, averaged over realizations of incidence vectors, are only about 2%.

Leading term in the limit of large S

In order to get more insight about the dependence with S in the limit $S \to \infty$, we have computed an asymptotic expansion of the variance to get the leading term in the series expansion on S. First let us write Eqn (D67) as

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx S u_{*,\circ} \int_0^{\infty} \mathrm{d}\theta \mathrm{e}^{-\theta}\theta\left(d_{*,\circ}+b_{*,\circ}\mathrm{e}^{-\theta}+S u_{*,\circ}\mathrm{e}^{-\theta}\right)\mathrm{e}^{(S-2)\log\left(d_{*,\circ}+(1-d_{*,\circ})\mathrm{e}^{-\theta}\right)}. \tag{D74}$$

In the limit of large S, the exponential function will be very peaked at the maximum of the function $\log\left(d_{*,\circ}+(1-d_{*,\circ})\mathrm{e}^{-\theta}\right)$. So we expect to have a good approximation in the limit $S \to \infty$ if we replace the logarithm by its series expansion,

$$\log\left(d_{*,\circ}+(1-d_{*,\circ})\mathrm{e}^{-\theta}\right) \approx -(1-d_{*,\circ})\theta + \mathcal{O}(\theta^2), \tag{D75}$$

about the point at which the maximum is reached, i.e, $\theta = 0$. Then, for large S, Eqn (D67) will be nicely approximated by

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx S u_{*,\circ} \int_0^\infty \mathrm{d}\theta \mathrm{e}^{-\theta}\theta \left(d_{*,\circ} + b_{*,\circ}\mathrm{e}^{-\theta} + S u_{*,\circ}\mathrm{e}^{-\theta}\right) \mathrm{e}^{-(S-2)(1-d)\theta}. \tag{D76}$$

This integral can be actually evaluated to give

$$\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx S u_{*,\circ} \left(\frac{d_{*,\circ}}{\left(1 + (S-2)(1-d_{*,\circ})\right)^2} + \frac{b_{*,\circ} + S u_{*,\circ}}{\left(2 + (S-2)(1-d_{*,\circ})\right)^2} \right) \tag{D77}$$

plus subleading terms in S. Here we observe that our approximation for $\mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right]$ converges to the squared heuristic Jaccard measure approximation,

$$\lim_{S\to\infty} \mathbf{E}\left[\left(\beta^{\mathrm{J}}\right)^2\right] \approx \left(\frac{u_{*,\circ}}{1 - d_{*,\circ}}\right)^2, \tag{D78}$$

so, in the limit of large S we find the following leading term for the variance approximation:

$$\mathbf{V}\left[\beta^{\mathrm{J}}\right] \approx \frac{u_{*,\circ}\left((1-d_{*,\circ})(b_{*,\circ} + d_{*,\circ}) - 4 d_{*,\circ} u_{*,\circ}\right)}{(1-d_{*,\circ})^3 S}. \tag{D79}$$

The variance decreases as S^{-1} in the case of large number of species. This explains why our heuristic approximation works very well in that limit.

Appendix E SDM

Data was collected over a four-year period (1993-1996) in usually three visits per year (2 above the treeline) using a simplified territory mapping approach, and integrated in the Swiss breeding bird atlas at 1-by-1 km resolution (Schmid *et al.* 1998, 2018). The data source we used included environmental predictor variables corresponding to climate, topography and vegetation structure at the same spatial scale, but geographic coordinates were removed. Data can be found at Zurell *et al.* 2019b,a. We used the whole dataset to infer SDMs using only the climatic variables, as current and future values of these variables for Switzerland are available in worldclim (`www.worldclim.org`; Hijmans *et al.* 2005). For each species, we used an ensemble approach of, initially, four different algorithms: generalized linear models (GLMs), generalized additive models (GAMs), random forests (RFs) and boosted regression trees (BRTs). However, GLMs and GAMs produced unreliable projections and were subsequently excluded from our analyses. We then projected current and future incidences for each species in our ensemble approach. With those incidences, we calculated the expected dissimilarity provided by Eqn (2) at each location of Switzerland for the subcommunity of woodpeckers.

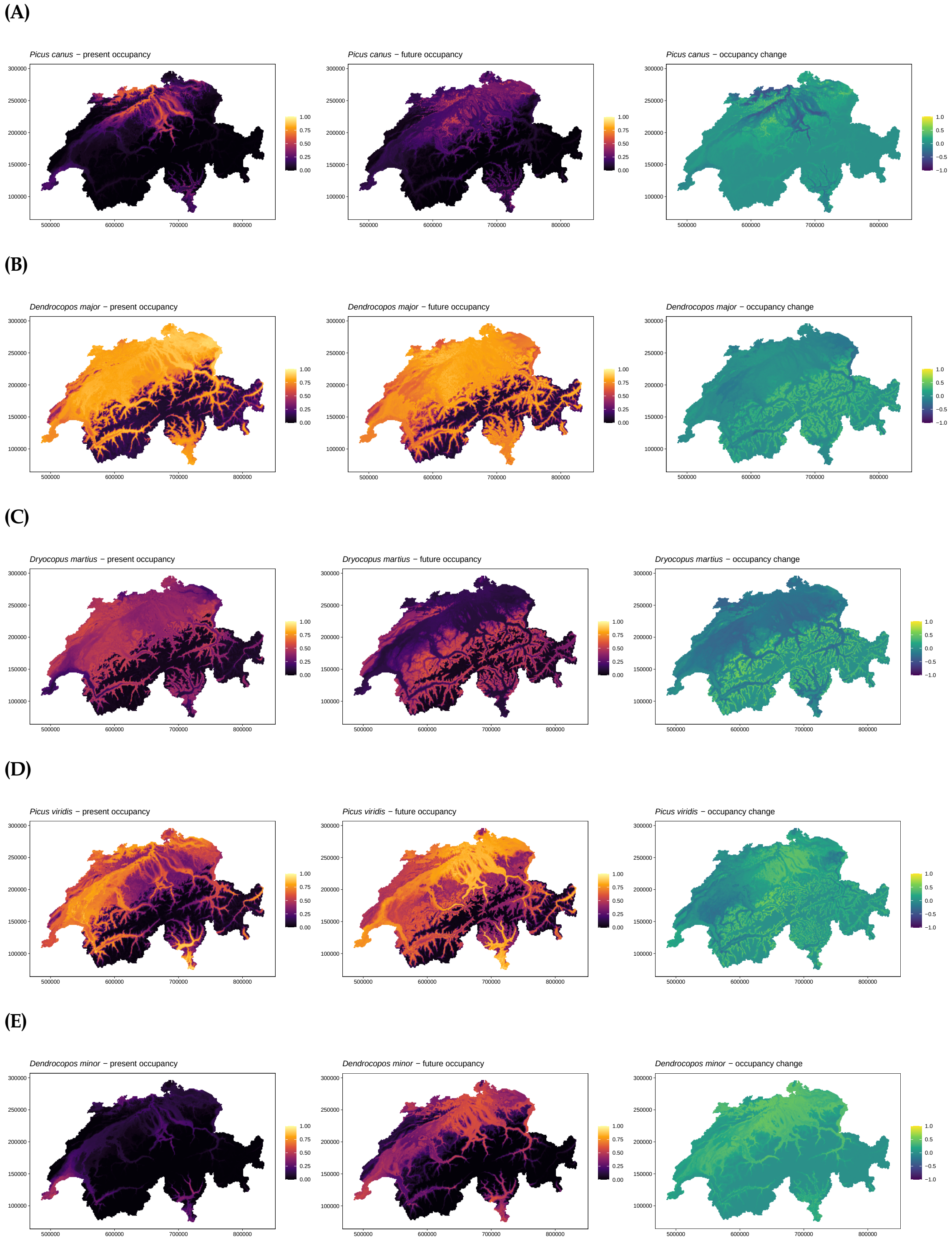
Literatures cited in Appendices

Arnold, B. C. (2012). *Majorization and the Lorenz order: A brief introduction*. **43**. Springer Science & Business Media (cit. on p. 27).

Chao, A. & Ricotta, C. (2019). Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, **100**.12. DOI: `10.1002/ecy.2852` (cit. on pp. 7, 28).

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**.15, pp. 1965–1978. DOI: `10.1002/joc.1276` (cit. on p. 32).

Hubbard, J. (1959). Calculation of partition functions. *Physical Review Letters*, **3**.2, pp. 77–78. DOI: `10.1103/physrevlett.3.77` (cit. on p. 29).


(A)
Picus canus − present occupancy
Picus canus − future occupancy
Picus canus − occupancy change
(B)
Dendrocopos major − present occupancy
Dendrocopos major − future occupancy
Dendrocopos major − occupancy change
(C)
Dryocopus martius − present occupancy
Dryocopus martius − future occupancy
Dryocopus martius − occupancy change
(D)
Picus viridis − present occupancy
Picus viridis − future occupancy
Picus viridis − occupancy change
(E)
Dendrocopos minor − present occupancy
Dendrocopos minor − future occupancy
Dendrocopos minor − occupancy change
300000
250000
200000
150000
100000
500000
600000
700000
800000
1.00
0.75
0.50
0.25
0.00
1.0
0.5
0.0
−0.5
−1.0

SI Figure 2: The presence probabilities: $p_{i,\text{current}}$, $p_{i,\text{future}}$, and $p_{i,\text{future}} - p_{i,\text{current}}$.

Jost, L. (2006). Entropy and diversity. *Oikos*, **113**.2, pp. 363–375. DOI: `10.1111/j.2006.0030-1299.14714.x` (cit. on p. 28).

— (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, **88**.10, pp. 2427–2439. DOI: `10.1890/06-1736.1` (cit. on p. 28).

Leinster, T. (2021). *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press (cit. on p. 28).

Lu, M., Vasseur, D., & Jetz, W. (2019). Beta diversity patterns derived from island biogeography theory. *The American Naturalist*, **194**.3, E52–E65. DOI: `10.1086/704181` (cit. on pp. 2, 6, 7, 11, 22).

Marshall, A. W., Olkin, I., & Arnold, B. C. (1979). *Inequalities: theory of majorization and its applications*. **143**. Springer (cit. on p. 28).

McVinish, R. & Lester, R. J. G. (2020). Measuring aggregation in parasite populations. *Journal of The Royal Society Interface*, **17**.165, p. 20190886. DOI: `10.1098/rsif.2019.0886` (cit. on p. 27).

Ontiveros, V. J., Capitán, J. A., Casamayor, E. O., & Alonso, D. (2021). The characteristic time of ecological communities. *Ecology*, **102**.2. DOI: `10.1002/ecy.3247` (cit. on pp. 3, 6, 7, 25).

Routledge, R. D. (1983). Evenness indices: are any admissible? *Oikos*, **40**.1, p. 149. DOI: `10.2307/3544211` (cit. on p. 28).

Schmid, H., M. Kestenholz, Knaus, P., L. Rey, & Sattler, T. (2018). The state of birds in Switzerland: special issue on the breeding bird atlas 2013–2016. en. DOI: `10.13140/RG.2.2.17872.15366` (cit. on pp. 10, 32).

Schmid, H., Posse, B., & Vogelwarte, S. S. (1998). *Schweizer Brutvogelatlas: Verbreitung der Brutvögel in der Schweiz und im Fürstentum Liechtenstein 1993-1996*. Schweizerische Vogelwarte (cit. on pp. 10, 32).

Van Lint, J. H. & Wilson, R. M. (2001). *A course in combinatorics*. Cambridge University Press (cit. on pp. 6, 23).

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2019a). *Data from: Testing species assemblage predictions from stacked and joint species distribution models*. en. DOI: `10.5061/DRYAD.K88V330` (cit. on p. 32).

— (2019b). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, **47**.1, pp. 101–113. DOI: `10.1111/jbi.13608` (cit. on pp. 10, 32).