Site-suitability difference and the expectation of species richness difference in stochastic two-subcommunities models

Ryosuke Iritani [*†‡§]

January 11, 2023

Abstract

Understanding how community assembly during the initial faunization phases determines the difference in species richness is a fundamental question in ecology. However, there are few predictions for when and to what degree the differences in species richness between subcommunities will emerge. Here, we investigate the expectation of richness difference in a pair of subcommunities, assuming that species may have different and independent presence probabilities. We then introduce several indices and examine how expected richness difference is determined by the indices. We found that (i) species differences (the average of species presence probabilities in two subcommunities divided by the total presence probability) have inconsistent effects on richness difference; (ii) the degree of spatial heterogeneity (average of differences in species presence probabilities in two subcommunities across species) may, but not always, have a good predictive ability for richness difference; and (iii) the absolute difference in average presence probabilities (site-suitability difference) has robust predictive ability for richness difference unless richness difference is very small. This work provides a theoretical framework for predicting and analyzing species richness difference from presence-absence data based on null models.

Key words: Heterogeneity; Nestedness; Null models; Species richness; Stochasticity.

*RIKEN Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS), RIKEN, Wako, Saitama, Japan;
†Department of Biology, Faculty of Sciences, The University of Tokyo
‡lambtani@gmail.com
§ORCID iD:0000-0002-2396-1109

1 Introduction

Why, and how much, does one area have richer species number than does another? Addressing this question and predicting the patterns of richness difference are ultimate goals in ecology [1–3]. Richness difference (the absolute difference between two local species-richnesses) is caused by various assemblage processes such as distance, area, species differences in colonization ability, and environmental filtering [4–7]. Knowledge in richness differences provides informative facets, for example, for conservation planning of biodiversity hotspots [8, 9], the global pattern of richness gradient [10–12], and spatial variations in biodiversity (beta-diversity; [13, 14]). It is thus of great importance to predict how richness difference is determined by various biotic and abiotic factors.

A challenging problem in predicting richness difference results from heterogeneous factors in natural communities [15–18]. Such factors include distance, degree of fragmentation, disturbance, and colonization ability [11, 12]. Surprisingly few theories exist to make predictions for richness difference. Iritani *et al.* [19], by handling species-incidence (presence or absence) as a probabilistic variable [2], studied how species difference and spatial heterogeneity affect beta-diversity, but did not consider richness difference. Similarly, based on the assumption that species incidence is independent, Calabrese *et al.* [20] and Carmona & Pärtel [21] took the probabilistic approach to estimate the species richness and its stochastic variations. However, neither did their studies consider richness difference. Taken together, the current methodological tools have not yet succeeded in providing theoretical predictions of how richness difference can be shaped by stochastic processes when there are species difference and spatial heterogeneity.

In this article, we derive an exact, analytical formula for richness difference between two subcommunities, assuming that: (i) species incidence is independent across space and species, and (ii) presence probabilities of species in two subcommunities are allowed to take differential values. Then we use three predictive indices (predictors) and examine their predictive abilities for richness difference. The present theoretical work provides useful guide to predict how richness difference can occur in natural and experimental environments, especially by nullifying correlations among species incidence.

2 Method and result

2.1 General analysis

We write $x_{i,j}$ for the binary state of species i's incidence in site j ($p_{i,j} = 1$ if i is present in j, otherwise absent $p_{i,j} = 0$), and $p_{i,j}$ for the probability that species i is present in site j. The probabilities $p_{i,j}$ may be, for example, estimated from long time-series data of species incidence (e.g., BioTIME [22]). We assume that a species-pool has a limited size, such that at most S species can be present in at least one of the sites (more technically, the species pool consists of species i that can inhabit at least one of the sites with a strictly positive probability, $\min\left\{p_{i,1}, p_{i,2}\right\} > 0$). The symbols are summarized in Appendix A.

The species richness, or alpha-diversity, in site j is given by:

$$\alpha_j(\mathbf{X}) := \sum_{i=1}^{S} x_{i,j}, \tag{1}$$

which is a nonnegative integer value, $\alpha_j = 0, 1, \ldots, S$. In statistics, α_j is said to follow Poisson-Binomial distribution with parameters $\mathbf{p}_j := \left(p_{1,j}, p_{2,j}, \ldots, p_{S,j}\right)$ [23]. The expected value of alpha-diversity is given by:

$$\mathbf{E}\left[\alpha_j\right] = \sum_{i=1}^{S} p_{i,j}. \tag{2}$$

Denoting by $\Lambda(\mathbf{X}) := \min\left\{\alpha_1(\mathbf{X}), \alpha_2(\mathbf{X})\right\}$ the minimum of the species richness, we can write down its expected value, $\mathbf{E}[\Lambda]$, in an analytically closed form (Appendix B). Using $\mathbf{E}[\Lambda]$ and writing $D(\mathbf{X}) := \left|\alpha_1(\mathbf{X}) - \alpha_2(\mathbf{X})\right|$ for the difference

of species richness between two subcommunities, we use the identity $|\alpha_1 - \alpha_2| \equiv \alpha_1 + \alpha_2 - 2\min\{\alpha_1, \alpha_2\}$ to get:

$$\mathbf{E}[D] = \mathbf{E}\left[\ |\alpha_1 - \alpha_2|\ \right] = \mathbf{E}[\alpha_1 + \alpha_2] - 2\mathbf{E}[\Lambda]. \tag{3}$$

2.2 Predictors

We use three parameters to characterize the variations in the presence probabilities, $p_{i,j}$. First is species difference denoted w, which is defined as the proportion of the average taken for two sites of absolute deviation of the presence probabilities from within-site average presence probability to the sum of presence probabilities for all species for two sites [19]. Second is spatial heterogeneity denoted h, which is defined as the average, taken across species, of the difference of presence probabilities for two sites. These are given by:

$$\begin{aligned} w &:= \frac{1}{\mu_1 + \mu_2} \sum_{i=1}^{S} \frac{1}{2S} \sum_{j=1}^{2} \left| p_{i,j} - \mu_j \right|, \\ h &:= \frac{1}{S} \sum_{i=1}^{S} \left| p_{i,1} - p_{i,2} \right|, \end{aligned} \tag{4}$$

where $\mu_j := \sum_{i=1}^{S} p_{i,j}$ represents the average presence probability of species in site j [19]. Species difference is zero $w = 0$ if and only if all species are identical in the presence probabilities. Finally, we propose the following quantity ("site-suitability difference"):

$$\delta := \frac{1}{S} \left| \sum_{i=1}^{S} \left(p_{i,1} - p_{i,2} \right) \right| = \frac{1}{S} \left| \mathbf{E}[\alpha_1] - \mathbf{E}[\alpha_2] \right| = \left| \mu_1 - \mu_2 \right|, \tag{5}$$

which represents the absolute difference in average presence probabilities in sites 1 and 2. Note that $\delta \leq h$ by the triangle inequality (in which the equality, $h = \delta$, occurs when one site is predominantly more suitable than the other: $p_{i,1} \geq p_{i,2}$ for all $i = 1, ..., S$ or either $p_{i,1} \leq p_{i,2}$ for all $i = 1, ..., S$), and also that $S \cdot \delta \leq \mathbf{E}[D]$ by Jensen's inequality [24].

2.3 Simulations

To examine how richness difference depends on these three indices, we carried out stochastic simulations based on the following scheme: (1) setting $S = 16$, we choose a pair of probability distributions to generate $\mathbf{p}_j = \left(p_{1,j}, ..., p_{S,j}\right)$ for site 1 and 2; we refer to such distributions as 'incidence-generators' in this article. The number of simulations is 300. (2) We compute the species difference, spatial heterogeneity, site-suitability difference, and expected richness difference, and then plot scatter plots.

We illustrate the theoretical result using three sets of the beta distributions for incidence-generators (and we write "LHS ~ Beta[•,•]" to mean that LHS follows the Beta distribution with parameters indicated by [•, •]).

(1) "Random model": $p_{i,1}$ ~ Beta[1.10, 1.09] and $p_{i,2}$ ~ Beta[1.09, 1.10];

(2) "Overlap model": $p_{i,1}$ ~ Beta[1.28,0.96] and $p_{i,2}$ ~ Beta[0.96,1.28]; and

(3) "Dominance model": $p_{i,1}$ ~ Beta[1.6,0.4] and $p_{i,2}$ ~ Beta[0.4,1.6].

For the random model, in which there is no clear small-large relationship between $p_{i,1}$ and $p_{i,2}$, neither species difference nor spatial heterogeneity predict the richness difference accurately, but site-suitability difference has a monotonic, and very consistent, effect of richness difference (Figure 1).

For the overlap model, in which $p_{i,2}$ is likely to be larger than $p_{i,1}$ compared to the random model case, richness difference has no clear dependence on species difference. The predictive abilities of both spatial heterogeneity and site-suitability for richness difference are improved over the random model case (Figure 1A and B), but the site-suitability's predictive ability decreases as site-suitability and richness difference are smaller (Figure 1B).

For the dominance model, in which $p_{i,2}$ is very likely to be larger than $p_{i,1}$, we find that whereas species difference have a low predictive ability, spatial heterogeneity and site-suitability difference both almost perfectly (and linearly) predict richness difference; it should be however noted that this coincidence is in part due to the equality $\delta = h$ for the

dominance model (Figure 1C). We found that these predictions are robust against changing the incidence-generators (the uniform, instead of beta distributions) as well as limited observations of incidence-tables (Appendix C).

In sum, we examined how accurately species difference index, spatial heterogeneity index, and site-suitability difference index can predict the expectation of richness difference. The predictive abilities of these indices strongly depend on the nature of probability distributions for species incidence. When species' presence probabilities in two sites are very similar as in the random model, it is difficult to predict the richness difference, but there seems to be a governing relation of site-suitability difference with richness difference. As the probability distributions for the two sites become dissimilar (from random, overlap, to dominance model), the predictive abilities of the site-suitability and spatial heterogeneity are improved.

3 Discussion

We investigated how accurately the expectation of richness difference can be predicted by three indices. These indices can be estimated from long time-series data of species incidence. We found that the expectation of richness difference has a complicated, and almost intractable, analytical form. While species differences and site-heterogeneity both have limited predictive ability of the expected richness difference, a site-suitability index, defined as the average of the absolute difference of alpha diversity between two sites, provides a good predictor. However, the predictive ability of the index decreases as the site-suitability index becomes small. The dependence of richness difference on site-suitability difference seems to be positive and nonlinear (though asymptotically linear).

The predictive ability of the spatial heterogeneity for richness difference is sensitive to the probability distributions with which species are present in each site. Specifically, if one site is a priori known to have predominantly larger species richness than does the other site, then spatial heterogeneity estimated from long-time series data of species presence-absence can be used as a measure of richness difference. On the other hand, without such a priori knowledge, i.e., if the average probabilities of species to be present in one site and in the other are similar, it would be difficult to statistically infer richness difference from site-heterogeneity measure alone. This is because site-heterogeneity measure is invariant with the permutation of sites for each species: exchanging $p_{i,1}$ and $p_{i,2}$ for each species i does not change the value of h, and thus does not detect absolute richness difference.

In addition, our key finding predicts that richness difference may be adequately predicted by the site-suitability index, all things being equal. Biologically, despite the presence of heterogeneity, the absolute difference in alpha-diversity provides a good estimate for the expected absolute difference in species richness. However, the fit becomes less accurate when the site-suitability and richness difference are both low, suggesting that it is difficult to estimate the richness difference from empirical data that typically contains a limited number of unique species. We reason this result as follows. First of all, the absolute value and its original value are equivalent (namely $|x| = x$) whenever the latter x is positive ('positivity invariance rule'). When site-suitability difference is large, it is immediately clear that one site (say site 2) tends to have predominantly higher alpha-diversity as does the other (site 1; dominance model case, Figure 1C). In this case, $\alpha_2 - \alpha_1$ is likely positive with a very high probability, so that the positivity-invariance rule applies and the original value provides the good estimate, although with some error associated with the case $\alpha_2 < \alpha_1$. In contrast, if site-suitability difference is small, then two sites are likely to have similar species richness (as in random model case; Figure 1A); in this case, the positivity-invariance rule does not apply because the sign of $\alpha_1 - \alpha_2$ varies. Biologically, site-suitability difference being zero likely occurs when two sites are homogeneous. Thus, heterogeneity may lower the predictive ability of site-suitability difference for richness difference.

Our model can be used as a framework of environmentally constrained null-model analyses of richness difference [15]. Despite the intuition that incorporating species differences and spatial heterogeneity matters, the present work showed that the effects of such heterogeneity are overall small unless the absolute difference of the expectations of alpha-diversity in two sites is low. Specifically, by averaging long time-series data of incidence (e.g., [22]; also see Appendix C3), one may estimate presence probabilities, assuming that presence probabilities are at dynamical equilibrium [2]. In this vein, deviations of observed richness difference from the expectation of the present work provide a signal of the effect of certain nullified factors. Since we assumed that species incidence are uncorrelated

and also that two sites are uncorrelated, such deviation can be explained by these factors.

Our work has further implications for predicting beta-diversity, because richness difference is a subcomponent beta-diversity (spatial variations in species richness; [6, 25]). Specifically, Baselga [6] proposed to partition incidence-based beta-diversity measures, e.g., Jaccard or Sørensen's [26–28], into additive components of richness difference and turnover. Iritani *et al.* [19] derived the exact formula for the expectation of Jaccard dissimilarity in the same setting as the current work's, but further methodology has yet to be developed for the additive partitioning scheme proposed by Baselga [6, 25]. Future studies may apply the methodology to the additive partitioning frameworks and thereby advance the theory of beta-diversity for the heterogeneous world.

The present study has several avenues for generalizations, e.g., considering multiple-sites case, relaxing species-independence assumption, and analyzing abundance-based nestedness components. Future studies may aim to make testable predictions for, and improve our understanding of, the repeated patterns of spatial biodiversity [29].

In conclusion, we derived the analytical formula for the expected richness-difference based on the usual assumption that species incidences are uncorrelated across space. Our stochastic simulations showed that richness difference is very well explained by site-suitability, but the explanatory power is reduced when site-suitability index is small. Richness difference is a fundamental quantity and has the potential for future studies to produce interesting hypotheses of beta-diversity patterns.

Acknowledgement and declaration

We thank Naoto Shinohara for helpful comments; and JSPS KAKENHI (grant numbers 19K22457, 19K23768 and 20K15882) for funding. We have no competing interests.

References

[1] Darlington Junior, P. J. *Zoogeography: the geographical distribution of animals*. Wiley, 1957, pp. 675–675.

[2] MacArthur, R. H. & Wilson, E. O. "An equilibrium theory of insular zoogeography". *Evolution* **17**.4 (1963), p. 373. DOI: 10.2307/2407089.

[3] Ulrich, W. & Almeida-Neto, M. "On the meanings of nestedness: back to the basics". *Ecography* **35**.10 (2012), pp. 865–871. DOI: 10.1111/j.1600-0587.2012.07671.x.

[4] Carvalho, J. C., Cardoso, P., & Gomes, P. "Determining the relative roles of species replacement and species richness differences in generating beta-diversity patterns". *Global Ecology and Biogeography* **21**.7 (2011), pp. 760–771. DOI: 10.1111/j.1466-8238.2011.00694.x.

[5] Podani, J. & Schmera, D. "A new conceptual and methodological framework for exploring and explaining pattern in presence - absence data". *Oikos* **120**.11 (2011), pp. 1625–1638. DOI: 10.1111/j.1600-0706.2011.19451.x.

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

[7] Qian, H., Ricklefs, R. E., & White, P. S. "Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America". *Ecology Letters* **8**.1 (2004), pp. 15–22. DOI: 10.1111/j.1461-0248.2004.00682.x.

[8] Ferrier, S., Powell, G. V., Richardson, K. S., Manion, G., Overton, J. M., Allnutt, T. F., Cameron, S. E., Mantle, K., Burgess, N. D., Faith, D. P., *et al.* "Mapping More of Terrestrial Biodiversity for Global Conservation Assessment". *BioScience* **54**.12 (2004), p. 1101. DOI: 10.1641/0006-3568(2004)054[1101:mmotbf]2.0.co;2.

[9] Hurlbert, A. H. & Jetz, W. "Species richness, hotspots, and the scale dependence of range maps in ecology and conservation". *Proceedings of the National Academy of Sciences* **104**.33 (2007), pp. 13384–13389. DOI: 10.1073/pnas.0704469104.

[10] Pianka, E. R. "Latitudinal gradients in species diversity". *Trends in Ecology & Evolution* **4**.8 (1989), p. 223. DOI: 10.1016/0169-5347(89)90163-8.

[11] Gaston, K. J. "Global patterns in biodiversity". *Nature* **405**.6783 (2000), pp. 220–227. DOI: 10.1038/35012228.

[12] Hillebrand, H. "On the Generality of the Latitudinal Diversity Gradient". *The American Naturalist* **163**.2 (2004), pp. 192–211. DOI: 10.1086/381004.

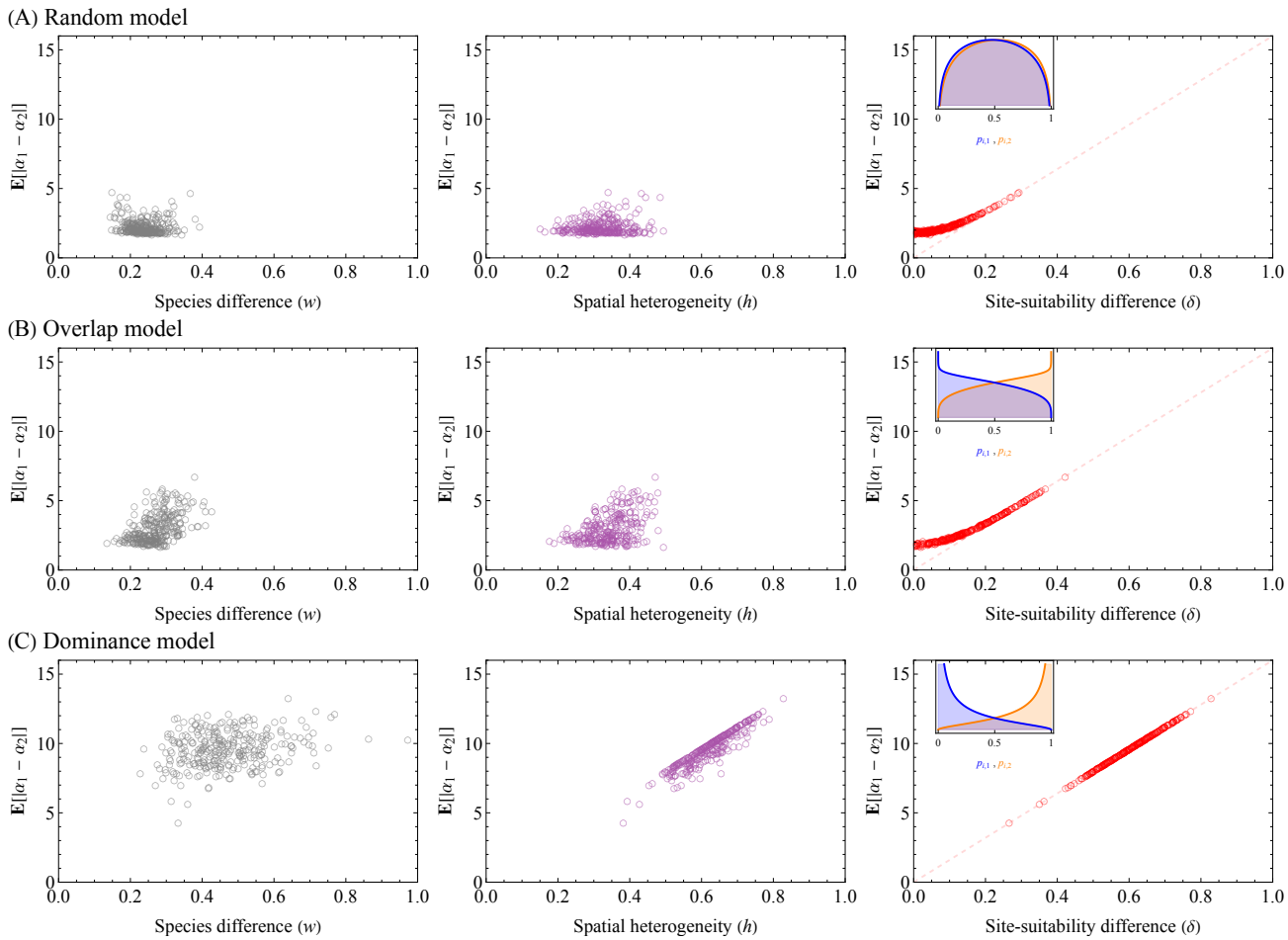


Figure 1: The performance of species difference (left), spatial heterogeneity (middle), and site-suitability difference (right), for predicting the expectation of richness difference, in three models producing different probability distributions of $\mathbf{p}_1$ and $\mathbf{p}_2$. In the right panels, the dotted segment is a line with slope S. Inset panels are the incidence-generators (i.e., the probability density functions) of $\mathbf{p}_1$ (blue) and $\mathbf{p}_2$ (orange). (A) When $\mathbf{p}_1$ and $\mathbf{p}_2$ are drawn from nearly indentical distributions (incidence-generator), richness difference is systematically small because the alpha diversity is likely similar between two sites. Neither can three predictors perform well, although site-suitability difference seems to have a formal relationship with richness difference. (B) When the two incidence-generators are partially (but not fully) overlapped, the predictive ability of site-suitability difference for species richness becomes better, whereas the other two indices do not show consistent patterns. (C) When one site is dominantly more likely to be inhabited by species than is the other, the predictive ability of site-suitability difference for the species becomes almost perfect. Parameters: $S = 16$; the number of simulations (i.e., circles) in each of panels (A–C): 300.

[13] Whittaker, R. H. "Vegetation of the Siskiyou Mountains, Oregon and California". *Ecological Monographs* **30**.3 (1960), pp. 279–338. DOI: 10.2307/1943563.

[14] 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. "Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist". *Ecology Letters* **14**.1 (2010), pp. 19–28. DOI: 10.1111/j.1461-0248.2010.01552.x.

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

[16] Willig, M., Kaufman, D., & Stevens, R. "Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis". *Annual Review of Ecology, Evolution, and Systematics* **34**.1 (2003), pp. 273–309. DOI: 10.1146/annurev.ecolsys.34.012103.144032.

[17] Connolly, S. R. "Macroecological theory and the analysis of species richness gradients". *Marine macroecology*. University of Chicago Press, 2009, pp. 279–309. DOI: 10.7208/chicago/9780226904146.003.0011.

[18] Gotelli, N. J., Anderson, M. J., Arita, H. T., Chao, A., Colwell, R. K., Connolly, S. R., Currie, D. J., Dunn, R. R., Graves, G. R., Green, J. L., Grytnes, J.-A., Jiang, Y.-H., Jetz, W., Lyons, S. K., McCain, C. M., Magurran, A. E., Rahbek, C., Rangel, T. F., Soberón, J., Webb, C. O., & Willig, M. R. "Patterns and causes of species richness: a general simulation model for macroecology". *Ecology Letters* **12**.9 (2009), pp. 873–886. DOI: 10.1111/j.1461-0248.2009.01353.x.

[19] Iritani, R., Ontiveros, V. L. J., Alonso, D., Capitan, J. A., Godsoe, W., & Tatsumi, S. "Jaccard dissimilarity in stochastic community models based on the species-independence assumption" (2022). DOI: 10.1101/2022.12.13.520233.

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

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

[22] Dornelas, M., Magurran, A. E., Buckland, S. T., Chao, A., Chazdon, R. L., Colwell, R. K., Curtis, T., Gaston, K. J., Gotelli, N. J., Kosnik, M. A., McGill, B., McCune, J. L., Morlon, H., Mumby, P. J., Øvreås, L., Studeny, A., & Vellend, M. "Quantifying temporal change in biodiversity: challenges and opportunities". *Proceedings of the Royal Society B: Biological Sciences* **280**.1750 (2013), p. 20121931. DOI: 10.1098/rspb.2012.1931.

[23] Fernandez, M. & Williams, S. "Closed-form expression for the Poisson-binomial probability density function". *IEEE Transactions on Aerospace and Electronic Systems* **46**.2 (2010), pp. 803–817. DOI: 10.1109/taes.2010.5461658.

[24] Ruel, J. J. & Ayres, M. P. "Jensen's inequality predicts effects of environmental variation". *Trends in Ecology & Evolution* **14**.9 (1999), pp. 361–366. DOI: 10.1016/s0169-5347(99)01664-x.

[25] Baselga, A. "The relationship between species replacement, dissimilarity derived from nestedness, and nestedness". *Global Ecology and Biogeography* **21**.12 (2012), pp. 1223–1232. DOI: 10.1111/j.1466-8238.2011.00756.x.

[26] Jaccard, P. "Nouvelles recherches sur la distribution florale". *Bull. Soc. Vaud. Sci. Nat.* **44** (1908), pp. 223–270.

[27] Jaccard, P. "The Distribution of the flora in the alpine zone". *New Phytologist* **11**.2 (1912), pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

[28] Sørensen, T. A. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". *Biol. Skar.* **5** (1948), pp. 1–34.

[29] MacArthur, R. H. *Geographical ecology: patterns in the distribution of species*. Princeton University Press, 1984.

Appendices

A Notation

- i: label for species
- S: the maximum number of species that can be present in at least one of the sites
- j: label for site
- x_{ij}: incidence of species i in site j

- $\mathbf{P}[\text{proposition}]$: the probability that the proposition holds true
- $\mathbf{X}$: matrix (table) with elements $x_{i,j}$
- $p_{i,j}$: the probability of species i present in site j
- $\mathbf{E}[]$: expectation over all possible $\mathbf{X}$s
- $\alpha_j(\mathbf{X}) := \sum_{i=1}^{S} x_{i,j}$: the species richness in site j
- $\mu_j := \sum_{i=1}^{S} p_{i,j} \,/\, S$: the average presence probability of speicies in site j
- $D(\mathbf{X}) := \left|\alpha_1(\mathbf{X}) - \alpha_2(\mathbf{X})\right|$: the absolute difference of richness

B Expectation of richness difference

In general, for any nonnegative integer-valued stochastic variable z, the following identity holds true:

$$\mathbf{E}[z] \equiv \sum_{k\geq 1} \mathbf{P}[z \geq k]. \tag{B1}$$

If we take $z = \min\{z_1, z_2\}$, that is, if we aim to calculate the minimum of two independent random variables, then:

$$\begin{aligned}\mathbf{E}\left[\min\{z_1, z_2\}\right] &= \sum_{k\geq 1} \mathbf{P}\left[\min\{z_1, z_2\} \geq k\right] \\ &= \sum_{k\geq 1} \prod_{j=1}^{2} \mathbf{P}\left[z_j \geq k\right] \\ &= \sum_{k\geq 1} \prod_{j=1}^{2} \left(1 - \mathbf{P}\left[0 \leq z_j \leq k-1\right]\right).\end{aligned} \tag{B2}$$

By taking $z_j = \alpha_j$, the formula leads to:

$$\mathbf{E}\left[\min\{\alpha_1, \alpha_2\}\right] = \sum_{k\geq 1} \prod_{j=1}^{2} \left(1 - \sum_{n=0}^{k-1} \mathbf{P}\left[\alpha_j = n\right]\right). \tag{B3}$$

Since α_j follows the Poisson-Binomial distribution (PBD) with parameters $\mathbf{p}_j\left(:= \left(p_{1,j}, \ldots, p_{S,j}\right)\right)$,

$$\mathbf{P}\left[\alpha_j = n\right] = \frac{1}{S+1} \sum_{\ell=0}^{S} \mathrm{e}^{-\frac{2\pi\imath n\ell}{S+1}} \prod_{m=1}^{S} \left(p_{m,j}\mathrm{e}^{\frac{2\pi\imath\ell}{S+1}} + 1 - p_{m,j}\right) \tag{B4}$$

[23], where $\imath$ represents the imaginary unit. Therefore,

$$\mathbf{P}\left[0 \leq \alpha_j \leq k-1\right] = \sum_{n=0}^{k-1} \mathbf{P}\left[\alpha_j = n\right] = \frac{1}{S+1} \sum_{\ell=0}^{S} \frac{1 - \mathrm{e}^{-\frac{2\pi\imath k\ell}{S+1}}}{1 - \mathrm{e}^{-\frac{2\pi\imath\ell}{S+1}}} \prod_{m=1}^{S} \left(p_{m,j}\mathrm{e}^{\frac{2\pi\imath\ell}{S+1}} + 1 - p_{m,j}\right). \tag{B5}$$

Note that the factor with $\ell = 0$ in the summand is a removal singularity for the fraction. Inserting the last equation into Equation (B3) supplies the expectation of the minimum richness. Using the identity $\alpha_1 + \alpha_2 - 2\min\{\alpha_1, \alpha_2\} \equiv |\alpha_1 - \alpha_2|$ and taking the expectation in both sides, we get the expectation of richness difference.

C Robustness

C.1 The uniform distributions

We choose the uniform distributions as the incidence-generators of $\mathbf{p_j}$. We found that the predictions obtained with the beta distributions are robust against this change (Figure SI-1).

C.2 Averaging: Binomial approximation

As before we first generate $\mathbf{p}_j$ using the same scheme and then calculate μ_j, and list it in a vector with length S: $\widehat{\mathbf{p}_1} := (\mu_1, ..., \mu_1)$ and $\widehat{\mathbf{p}_2} := (\mu_2, ..., \mu_2)$. Then calculate the exact expectation $\mathbf{E}[D]$ (using $\mathbf{p}_1$ and $\mathbf{p}_2$) and the approximated expectation based on the approximation (using $\widehat{\mathbf{p}_1}$ and $\widehat{\mathbf{p}_2}$). We found that the approximation tends to underestimate the true expectation, but the error becomes lower as the incidence-generators become more dissimilar (Figure SI-2).

C.3 Averaging the raw values of δ and D

So far, we used the expectation of richness difference and site-suitability from $p_{i,j}$s, and compared their values. In real situations, however, we estimate the mean richness difference and mean site-suitability both from incidence tables and compare them. That is, the comparison of the true expectations ($\mathbf{E}[D]$ versus δ) and that of means (average of D versus average of δ) are different.

Here we assess whether the difference is large. As before, we generate $\mathbf{p}_j$ from incidence-generators:

$$\mathcal{P}_{(k)} = \left(\mathbf{p}_{1;(k)}, \mathbf{p}_{2;(k)}\right), \tag{C6}$$

each of which is drawn from a fixed probability distribution (beta distribution), where k represents the simulation number, and incidence-tables ('empirical tables') from $\mathbf{p}_j$, with the number of the empirical tables, N, set to 300 (note that with $S = 16$ we have $2^{32} \approx 10^{10}$ possible incidence-tables, far larger than 300):

$$\mathcal{X}_{(k)} = \left\{\widehat{\mathbf{X}}^{[1]}_{(k)}, \widehat{\mathbf{X}}^{[2]}_{(k)}, ..., \widehat{\mathbf{X}}^{[300]}_{(k)}\right\}, \tag{C7}$$

where each $\mathbf{X}^{[\ell]}_{(k)}$ for $\ell = 1, ..., 300$ is drawn from the common, joint Beta distribution. We then compute (i) the true expectation of richness difference as well as (ii) the average of the absolute values of the richness difference for the empirical tables:

$$\mathcal{D}_{(k)} = \left\{D\left(\widehat{\mathbf{X}}^{[1]}_{(k)}\right), D\left(\widehat{\mathbf{X}}^{[2]}_{(k)}\right), ..., D\left(\widehat{\mathbf{X}}^{[300]}_{(k)}\right)\right\} \tag{C8}$$

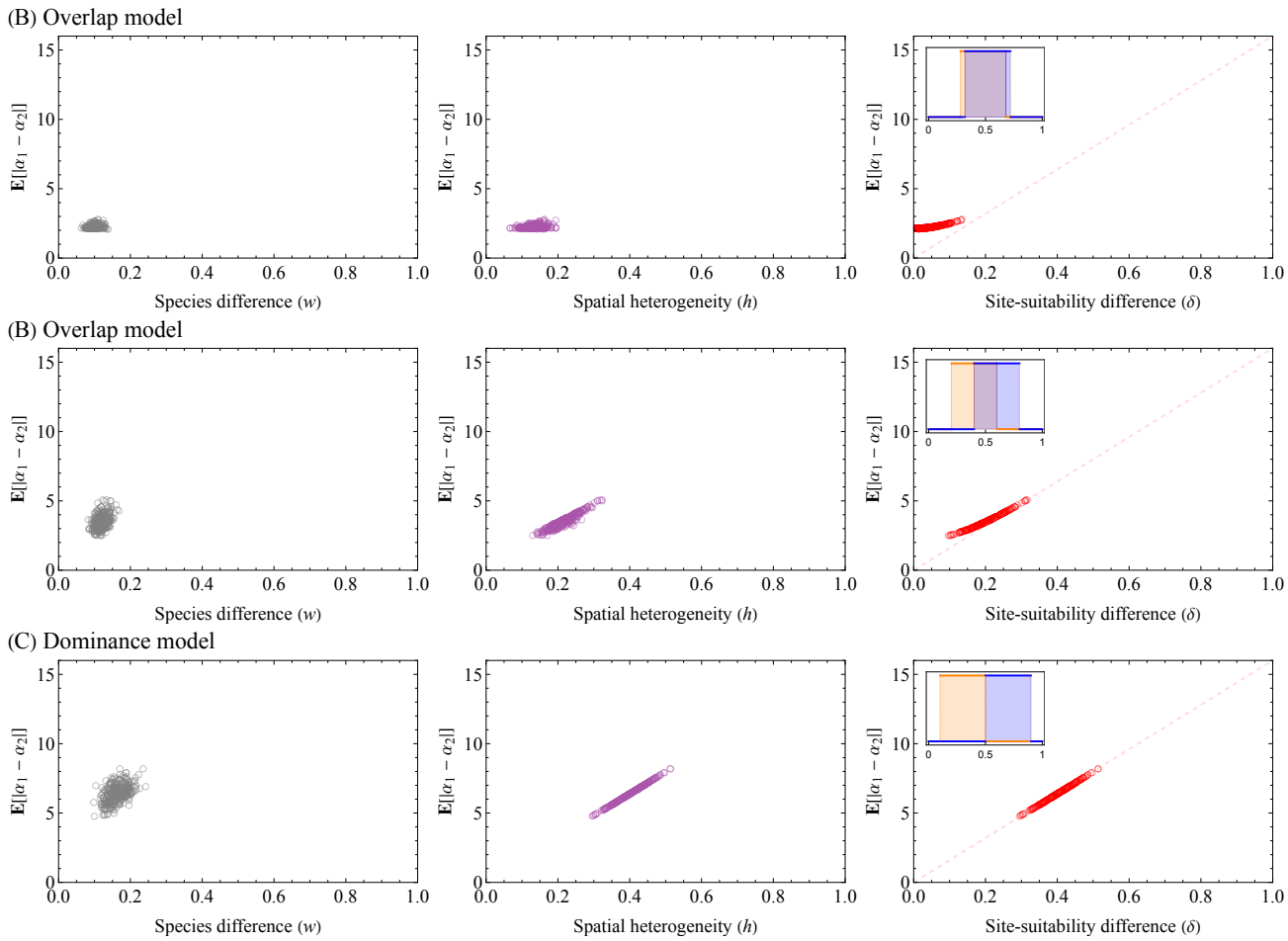
We then get:

$$\begin{aligned} \overline{D_{(k)}} &:= \frac{1}{300}\sum_{\ell=1}^{300} D\left(\widehat{\mathbf{X}}^{[\ell]}_{(k)}\right), \\ \overline{\delta_{(k)}} &:= \frac{1}{300}\left|\sum_{\ell=1}^{300} \frac{\alpha_1\left(\widehat{\mathbf{X}}^{[\ell]}_{(k)}\right) - \alpha_2\left(\widehat{\mathbf{X}}^{[\ell]}_{(k)}\right)}{S}\right|, \end{aligned} \tag{C9}$$

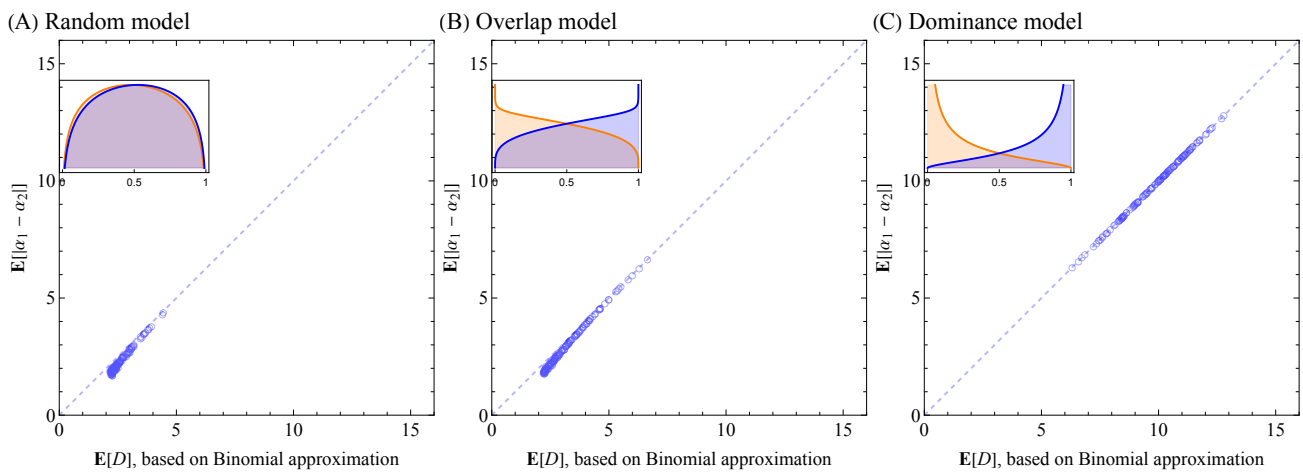
We repeat this procedure 200 times and compared $\overline{D_{(k)}}$ and $\overline{\delta_{(k)}}$ ($k = 1, 2, ..., 200$) by depicting a scatter plot (Figure SI-3).

Literatures cited in Appendices

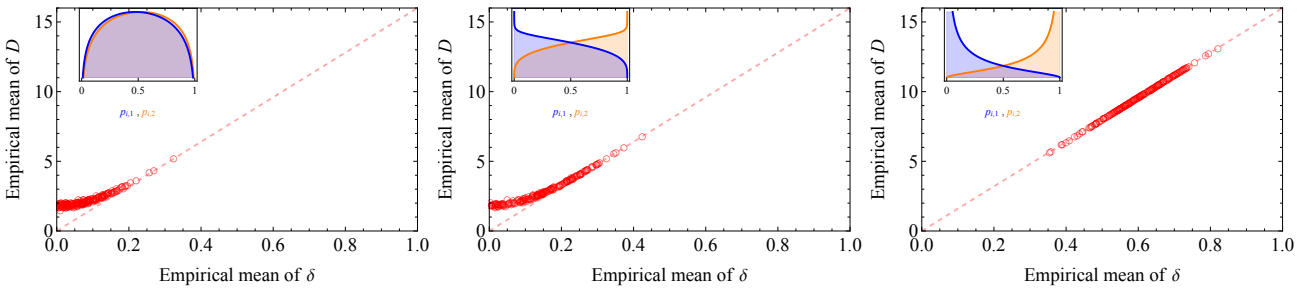
[23] Fernandez, M. & Williams, S. "Closed-form expression for the Poisson-binomial probability density function". *IEEE Transactions on Aerospace and Electronic Systems* **46**.2 (2010), pp. 803–817. DOI: 10.1109/taes.2010.5461658.



SI Figure SI-1: The same figure as the one in the main text, except the choice of the incidence-generators. Uniform distributions on: (A) $[0.28, 0.68]$ and $[0.32, 0.72]$; (B) $[0.2, 0.6]$ and $[0.4, 0.8]$; (C) $[0.1, 0.5]$ and $[0.5, 0.9]$.



SI Figure SI-2: Approximation, based on the averaging method.



SI Figure SI-3: Plotting the relation between (i) abscissa: the average value of δ estimated from 300 incidence tables generated by a common probability distribution, versus (ii) ordinate: the average value of D from the 300 incidence tables generated by a common probability distribution. Each circle corresponds to a single simulation run (200 runs).