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

Ryosuke Iritani *†‡§

January 4, 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 is little prediction for when and to what degree species richness will emerge. Here, we investigate the expectations of richness difference in a pair of subcommunities, assuming that species may have different and independent presence probabilities in a two-subcommunity system. We also explore predictive indices for richness difference. We found that (i) species differences (the average of species presence abilities in two subcommunities divided by the total presence probability) tend to have inconsistent effects on richness difference; (ii) strong spatial heterogeneity (average of differences in species presence probabilities in two subcommunities across species) may have good predictive power for the emergence of richness differences; and (iii) the absolute difference in average presence probabilities (site-suitability difference) has robust predictive power for richness differences. Depending on the circumstances, the predictive ability is so large that it can provide an accurate approximation, but the predictive ability tends to decrease as the site-suitability index becomes smaller. This finding suggests that, although it is tempting and challenging to incorporate full, heterogeneous factors into null models, the average suitability of site occupancy for any species may suffice to predict patterns of richness differences in the two subcommunities. This work also demonstrates that null-model analyses typically assuming species equivalence may be well-underpinned by robustness against heterogeneity. Overall, this work provides a theoretical framework for accurately predicting and analyzing species richness differences from presence-absence data using 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

Understanding the differences in species richness ('richness difference') is one of the fundamental problems in ecology [1, 2]. For presence-absence data, richness difference is defined as the absolute difference between two local species richnesses. The richness difference may be caused by various assemblage processes such as distance, area, species differences in colonization ability, and spatial heterogeneity [3–6]. Because richness difference provides essential information for the spatial biodiversity, it is becoming more and more important to predict factors shaping richness difference in the course of deterministic and stochastic community assemblage processes.

A challenging problem associated with predicting richness differences stems from complex, heterogeneous nature of the real world [7–10]. In the presence of heterogeneity, there are at least three factors that may contribute to richness differences. For example, spatial structure mediated by heterogeneity in spatial locations, distance, degree of fragmentation, and colonization ability can readily result in richness differences and subsequently form richness gradient [11, 12]. Surprisingly little theory exists as to how richness differences are influenced by such factors and how one can make accurate predictions for richness differences, both limiting our understanding of richness variations in nature. Iritani *et al.* [13], by regarding species incidence (presence or absence) as a probabilistic variable [14], studied how species difference and spatial heterogeneity affect compositional dissimilarity between two subcommunities (beta-diversity), but did not consider any properties of richness differences. Similarly, based on the assumption that species incidence is independent, Calabrese *et al.* [15] and Carmona & Pärtel [16] 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 analytical, exact formula for richness difference of species richness between two subcommunities, assuming that: (i) species incidence is independent across space and species, and (ii) differential values are taken among the presence probabilities of species i in subcommunity j (for $i = 1, \ldots, S$ and $j = 1, 2$). Then we examine three factors for their potential for good prediction of richness differences. The present theoretical work provides useful guide to predict how richness differences can occur in natural and experimental environments, especially by nullifying correlations among species incidence.

2 Method and result

2.1 General analysis

We write $p_{i,j}$ for the probability that species i is present in subcommunity $j(= 1, 2)$, and $x_{i,j}$ for the binary state of species i's incidence in subcommunity j: if species i is present in subcommunity j, we write $x_{i,j} = 1$, otherwise absent $x_{i,j} = 0$. We assume that species-pool is limited, such that the number S of species in total 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, $\min\{p_{i,1}, p_{i,2}\} > 0$). The symbols are summarized in Appendix A.

The species richness, or alpha-diversity, in j can be written as:

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

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 $\mathbf{p}_j := (p_{1,j}, p_{2,j}, \ldots, p_{S,j})$ [17]. The expected value of alpha-diversity over all possible $\mathbf{X}$s is given by:

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

The minimum of the species richness is defined by:

$$\mu_-(\mathbf{X}) := \min\left\{\alpha_1(\mathbf{X}), \alpha_2(\mathbf{X})\right\}. \tag{3}$$

We can write down its expected value, $\mathbf{E}\left[\mu_-\right]$, in a complicated form (Appendix B).

The expectation of the minimum of species richness for two subcommunities is immediately translated down to the expectation of the difference in species richness between two subcommunities. Let $D(\mathbf{X}) := \left|\alpha_1(\mathbf{X}) - \alpha_2(\mathbf{X})\right|$ be the difference of species richness between two subcommunities. Using the identity $|\alpha_1 - \alpha_2| \equiv \alpha_1 + \alpha_2 - 2\min\{\alpha_1, \alpha_2\}$ yields:

$$\mathbf{E}[D] = \mathbf{E}\left[\ |\alpha_1 - \alpha_2|\ \right] = \mathbf{E}[\alpha_1 + \alpha_2] - 2\mathbf{E}\left[\mu_-\right]. \tag{4}$$

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 [13]. 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:

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

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

where $\mu_j := \sum_{i=1}^{S} p_{i,j}$ represents the average presence probability of species in site j [13]. Finally, we propose the following quantity:

$$\delta := \frac{1}{S}\left|\sum_{i=1}^{S} (p_{i,1} - p_{i,2})\right| = \frac{1}{S}\left|\mathbf{E}[\alpha_1] - \mathbf{E}[\alpha_2]\right| = \left|\mu_1 - \mu_2\right|, \tag{7}$$

which represents the absolute difference of alpha-diversity and describes which site is on average more likely to be inhabited. Note that $\delta \le h$ by the triangle inequality and also that $\delta \le \mathbf{E}[D]$ by Jensen's inequality [18]. in which the equality ($h = \delta$) occurs when one site is predominantly more suitable than the other: $p_{i,1} > p_{i,2}$ for all $i = 1, \ldots, S$ or either $p_{i,1} < p_{i,2}$ for all $i = 1, \ldots, S$.

2.3 Stochastic simulations

To illustrate the dependence of richness difference against these three quantities, we carried out stochastic simulations (Monte Carlo simulations), based on the following scheme. (1) Setting $S = 16$, choose two probability distributions to generate $\mathbf{p}_j = \left(p_{1,j}, \ldots, p_{S,j}\right)$; 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 first find that whereas neither species difference and spatial heterogeneity on their own may provide clear predictions for richness difference, the site suitability difference best explains richness differences. This prediction seems to strongly depend on the probabilistic distributions to generate the values of $p_{i,j}$.

Second, 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, neither species difference nor spatial heterogeneity tend to predict the richness difference accurately, but site-suitability difference can generate a monotonic dependence of richness difference (Figure 1A).

For the overlap model, 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 pattern, but the site-suitability's predictive ability decreases with decreased site-suitability and richness difference (Figure 1B).

The dominance model predicts that, whereas species difference tends to 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 due to the equality $\delta = h$ for the dominance model (Figure 1C).

To examine the robustness of the theoretical predictions, we changed the incidence-generators (e.g., the uniform distributions) and found the result robust (Appendix C1). Also, we assessed if the result of the expected richness difference based on the Poisson-Binomial distributions with $\mathbf{p}_1$ and $\mathbf{p}_2$ can be approximated by that of the expected richness difference based on the binomial distributions with $(\mu_1, \mu_1 \ldots, \mu_2)$ and $(\mu_1, \mu_1 \ldots, \mu_2)$ (respectively), showing that the approximated expectation of species richness tends to underestimate, but agrees well with, the exact. The accuracy becomes better especially when the incidence-generators are very dissimilar (e.g., in the dominance model). Finally, we checked whether the results are robust against the imperfect estimation of $p_{i,j}$. Specifically, using the same scheme as above, we generate $\mathbf{p}_j$ from incidence-generators, and incidence-tables ('empirical tables') from $\mathbf{p}_j$, with the number of the empirical tables 300 (note that with $S = 16$ we have $2^{32} \approx 10^{10}$ possible incidence-tables, far larger than 300). From these empirical tables, we compute (i) the average of richness difference ('empirical mean of D', which is the unbiased estimator of the true expectation of D) as well as (ii) the average of the absolute value of the richness difference ('empirical mean of δ', which is the unbiased estimator of the true site-suitability difference index). We repeat this procedure 200 times and compared (i) and (ii) by depicting a scatter plot. We found that the relation between the empirical mean of D and empirical mean of δ is very similar as that between the exact expectation of D and the original δ.

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 tend to be improved.

3 Discussion

Understanding and predicting differences in species richness have been of great interest in community ecology. Using probabilistic approach, we investigated how 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 tend to 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 degree of spatial heterogeneity can have a good predictive power for richness difference, but the predictiveness is sensitive to the probability distributions with which species are present in each site. Specifically, if one site is a priori known to be likely inhabited by predominantly more species over 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 deduce richness difference from site-heterogeneity measure per se. This is because site-heterogeneity measure is invariant with the permutation of sites for each species: exchanging $p_{i,1}$

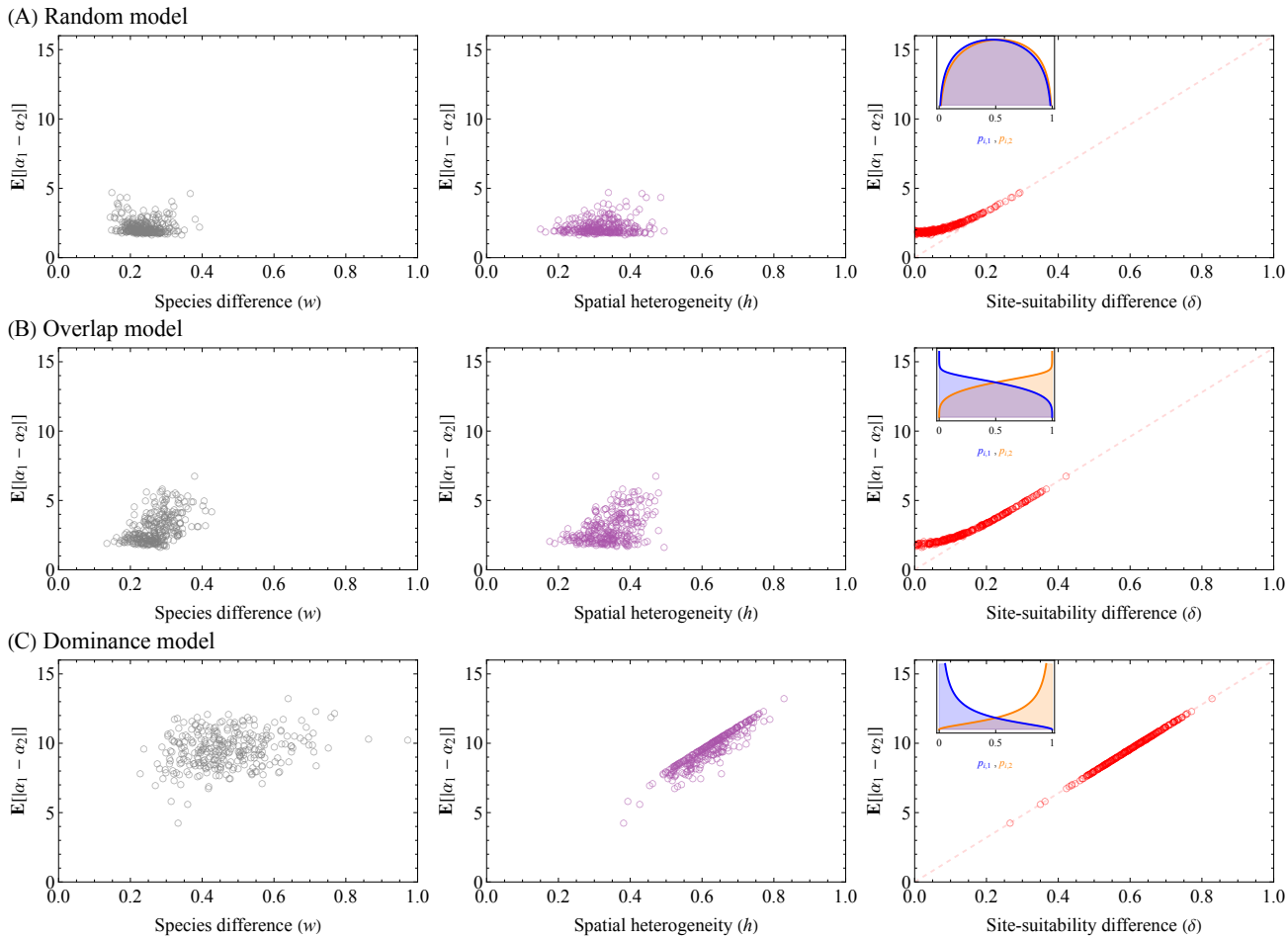


Figure 1: The performance of species difference (left), spatial heterogeneity (middle), and site-suitability difference (right), for predicting the expectation of richness difference, in three models producing different probability distributions of $\mathbf{p}_1$ and $\mathbf{p}_2$. In the right panels, 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 tends to be 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 tend to not show consistent patterns. (C) When one site is dominantly more likely to be inhabited by species, 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.

and $p_{i,2}$ for each species i does not change the value of h, and thus is not designed to detect richness difference.

In addition, our key finding predicts that richness difference may be adequately predicted by the site-suitability index, all things being equal. This mathematically suggests that it does not much matter exchanging the order of taking expectation and taking absolute value. 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. This result can be reasoned as follows. When site-suitability difference is large, it is immediately clear that one site is predominantly more likely to be inhabited by various species than the other, so that richness difference becomes more pronounced. In contrast, a small value of site-suitability means that two sites are equally likely to be inhabited by species and thus are similar, suggesting that the uncertainty as for the richness difference is large. In the extreme case where $p_{i,j}$ are all equal, which implies $\delta = 0$, the expectation of richness difference is still positive.

Our model can be used as a framework of environmentally constrained null-model analyses of richness difference [7]. 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., [19], one may estimate presence probabilities, assuming that presence probabilities are at dynamical equilibrium [14]. 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; [5, 20]). Specifically, Baselga [5] proposed to partition incidence-based beta-diversity measures, e.g., Jaccard or Sørensen's [21–23], into additive components of richness difference and turnoever. Iritani *et al.* [13] 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 [5, 20]. 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 limitations. First, the methodology is limited to the two-sites case, and therefore a multi-site scenario should be throughly considered. Second, the species-independence assumption, by which we consider that presence probabilities of each species are uncorrelated, may be less realistic, although it provides the null-model analytical framework. Third, related to the second limitation, more realistic spatial structure should be considered. Finally, generalizing the method to abundance-based nestedness components would be a challenging problem. That beings said, these limitations rather suggest that the current methodology has a number of research avenue to make testable predictions for, and improve our understanding of, the repeated patterns of spatial biodiversity [24].

In conclusion, we derived the analytical formula for the expected richness-difference based on the usual assumption that species incidences are uncorrelated across space. We performed stochastic simulations and found that richness difference is very well explained by site-suitability, but the explanatory power is reduced when site-suitability index is small, i.e., when the variation in spatial conditions for species presence versus absence 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 JSPS KAKENHI (grant numbers 19K22457, 19K23768 and 20K15882) for funding. We have no competing interests. The author RI is the only person as an author who contributed to the article.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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
- $x_{i,j}$: 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
- $\alpha_j(\mathbf{X}) \coloneqq \sum_{i=1}^{S} x_{i,j}$: the species richness in site j
- $\mu_j \coloneqq \sum_{i=1}^{S} p_{i,j} \,/\, S$: the average presence probability of speicies in site j
- $D(\mathbf{X}) \coloneqq \left|\alpha_1(\mathbf{X}) - \alpha_2(\mathbf{X})\right|$: 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:

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

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

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

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

$$\mathbf{E}\left[\min\{\alpha_1, \alpha_2\}\right] = \sum_{k\geq 1} \prod_{j=1}^{2} \left(1 - \sum_{n=0}^{k-1} \mathbf{P}\left[\alpha_j = n\right]\right). \tag{SI 3}$$

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

$$\mathbf{P}\left[\alpha_j = n\right] = \frac{1}{S+1} \left(\prod_{i=1}^{S} p_{i,j}\right) \sum_{\ell=0}^{S} \mathrm{e}^{-\frac{2\pi n \ell}{S+1}} \prod_{m=1}^{S} \left(\mathrm{e}^{\frac{2\pi i \ell}{S+1}} - \rho_{m,j}\right) \tag{SI 4}$$

[17], where $\imath$ represents the imaginary unit, and $\rho_{m,j} := -\left(1 - p_{m,j}\right) / p_{m,j}$. Therefore,

$$\mathbf{P}\left[0 \le \alpha_j \le k-1\right] = \sum_{n=0}^{k-1} \mathbf{P}\left[\alpha_j = n\right] = \frac{1}{S+1}\left(\prod_{i=1}^{S} p_{i,j}\right) \sum_{\ell=0}^{S} \frac{1 - \mathrm{e}^{-\frac{2\pi\imath k\ell}{S+1}}}{1 - \mathrm{e}^{-\frac{2\pi\imath\ell}{S+1}}} \prod_{m=1}^{S}\left(\mathrm{e}^{\frac{2\pi\imath\ell}{S+1}} - \rho_{m,j}\right). \tag{SI 5}$$

Note that the factor with $\ell = 0$ in the summand is a removal singularity for the fraction in the summand. Inserting this into Equation (SI 2) 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 gives 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 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: $\widehat{\mathbf{p}_1} := \left(\mu_1, \ldots, \mu_1\right)$ and $\widehat{\mathbf{p}_2} := \left(\mu_2, \ldots, \mu_2\right)$. 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 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 such differences are 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):

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

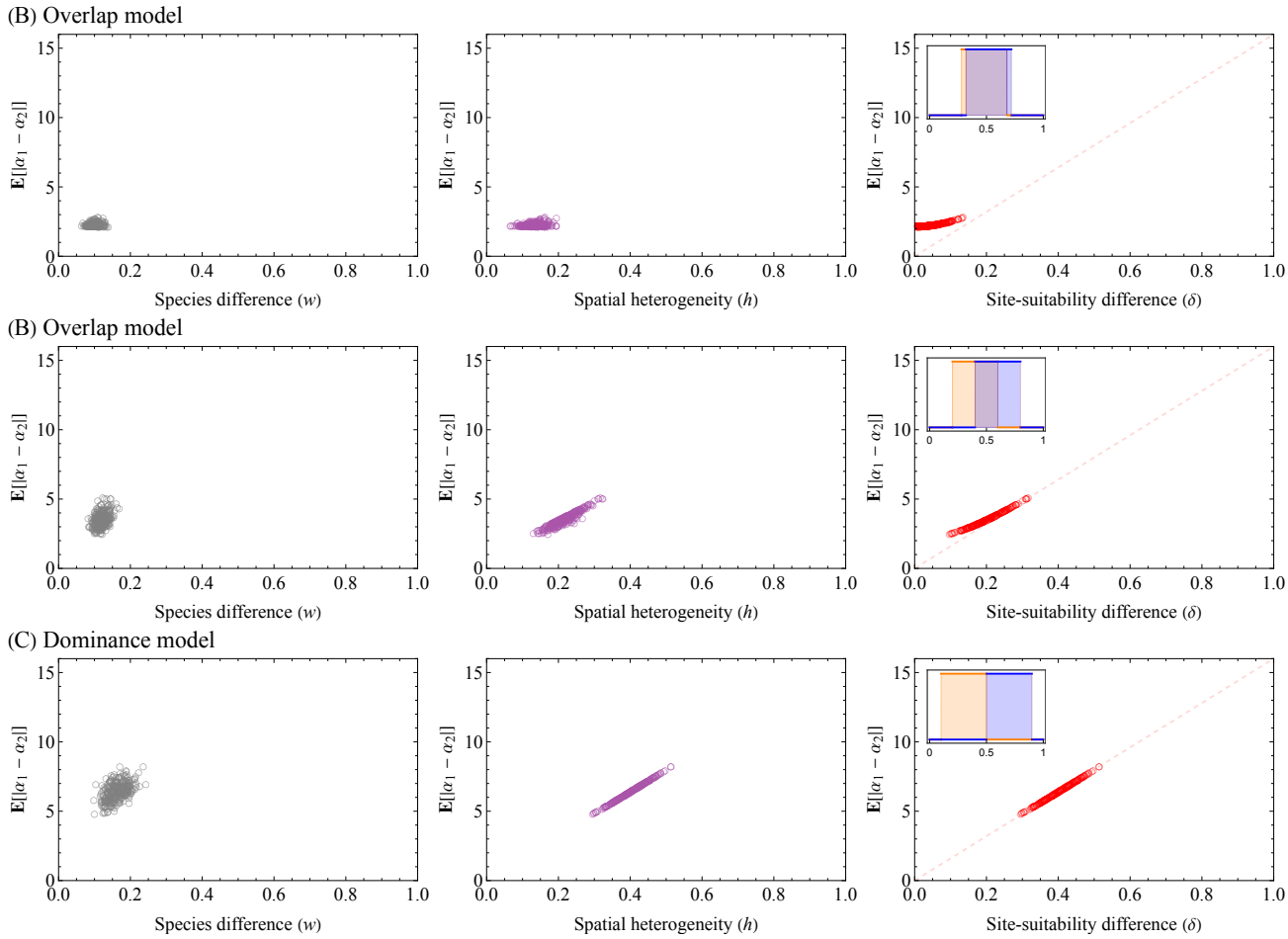
where each $\mathbf{X}_{(k)}^{[\ell]}$ for $\ell = 1, \ldots, 300$ is drawn from the common, joint Beta distribution $\mathcal{P}_{(k)}$. We then compute (i) the true expectation of richness difference as well as (ii) the average of the absolute values of the richness difference for the empirical tables:

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

We then get:

$$\overline{D_{(k)}} \coloneqq \frac{1}{300}\sum_{\ell=1}^{300} D\left(\widehat{\mathbf{X}}_{(k)}^{[\ell]}\right),$$
$$\overline{\delta_{(k)}} \coloneqq \frac{1}{300}\sum_{\ell=1}^{300} D\left(\widehat{\mathbf{X}}_{(k)}^{[\ell]}\right), \tag{C9}$$

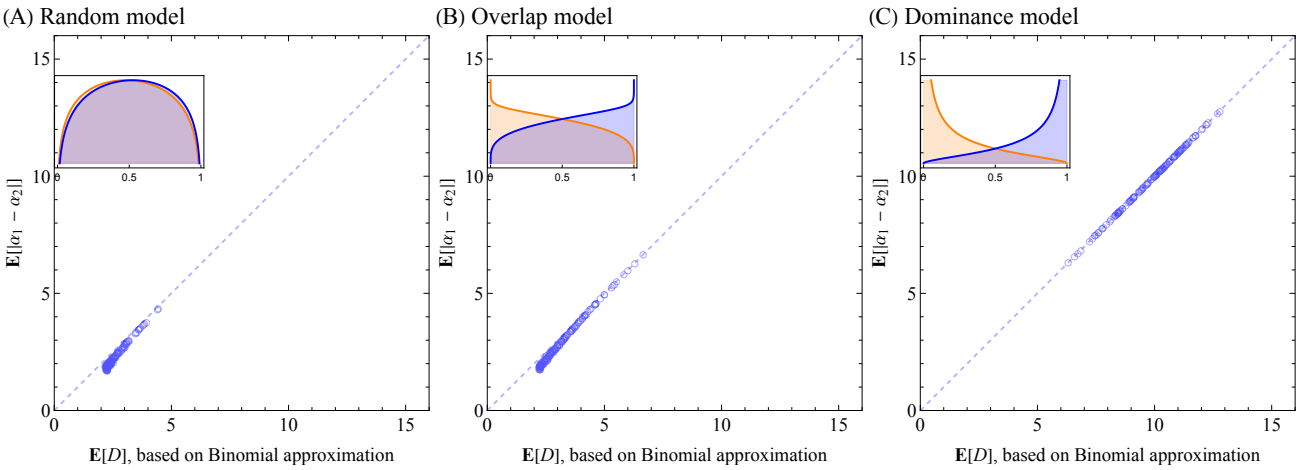
We repeat this procedure 200 times and compared (i) and (ii) by depicting a scatter plot (Figure 3).



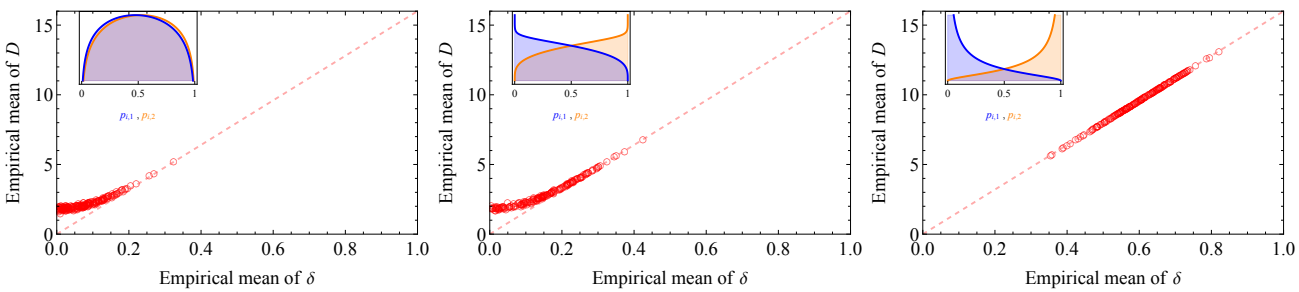
SI Figure 1: The same figure as the one in the main text, except the choice of the incidence-generators. Uniform distributions on: (A) $[0.28, 0.68]$ and $[0.32, 0.72]$; (B) $[0.2, 0.6]$ and $[0.4, 0.8]$; (C) $[0.1, 0.5]$ and $[0.5, 0.9]$.

Literatures cited in Appendices

[17] 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 2: Approximation, based on the averaging method.



SI Figure 3: Approximation, based on the averaging method.