A novel sequential sampling algorithm for the application of mechanistic models to species occurrence patterns

Nadiah P. Kristensen[1], Yong Chee Keita Sin[1], Hyee Shynn Lim[1], Frank E. Rheindt[1], and Ryan A. Chisholm[1]

https://doi.org/10.5802/placeholder.doi

Abstract

Aim: Species occurrence patterns are typically analysed using data-randomisation approaches, which reveal when observed patterns deviate from random expectation, but give little insight why. Mechanistic models, such as neutral models, have been proposed as an alternative null model, but they are computationally expensive. Here, we develop an efficient method to simulate such models and use it to explore likely mechanisms governing the occurrence patterns of birds on islands.

Location: Riau archipelago, Indonesia.

Taxon: Birds.

Methods: We used species richness and island area data to fit a niche–neutral model, where species obey neutral dynamics within non-overlapping discrete niches. We developed a sequential sampling algorithm that can efficiently generate randomised presence–absence matrices under the niche-neutral model and used mismatches to identify which mechanisms were potentially important to occurrence patterns.

Results: Birds were more segregated and less nested than expected from both data randomisation and the niche–neutral model. The niche–neutral model reproduced the mean relationship between island size and species richness, but it could not produce sufficient variability to account for the richness data. The model was brought into closer agreement with the data by allowing niche diversity to vary across islands (increased segregation) and allowing the per-capita immigration rate to vary across islands (decreased nestedness).

Main conclusion: While the species-area relationship could be explained by a model with constant per-capita immigration rates and number of niches across islands, inter-island heterogeneity was needed to explain higher-order metrics of nestedness and co-occurrence patterns. Our novel sequential sampling algorithm allowed us to explore different scenarios efficiently, and our approach may be useful for identifying the mechanisms structuring occurrence patterns in other systems.

Keywords: Co-occurrence, Segregation, Coalescence, Neutral Model, Birds

[1]Department of Biological Sciences, National University of Singapore, Singapore, Singapore.

Correspondence
nadiah@nadiah.org

1. Introduction

What determines the pattern of species occurrence across patches in a landscape? One method for identifying patterns of potential ecological interest is null-model analysis using randomised data (Gotelli, 2000; Zhang, 2020). One first computes a metric that summarises an occurrence pattern, then one compares the observed value to the distribution obtained when the data are randomised in various ways (Ulrich and Gotelli, 2007). For example, the 'fixed–fixed' randomisation swaps species presences and absences (1s and 0s in the presence–absence matrix) while holding fixed the number of sites where each species is observed (row sums) and the number of species observed at each site (column sums) (Strona et al., 2014). If the observed metric deviates significantly from the distribution in randomised data, then qualitative inferences can be drawn about the underlying mechanisms (e.g., Kohli et al., 2018; Meyer and Kalko, 2008). For example, if observed species co-occurrences are more segregated than in randomised data (Connor and Simberloff, 1979; Gotelli, 2000), then this may indicate strong inter-species competition (Diamond, 1975).

Data randomisation is versatile and imposes minimal data requirements (i.e., no information other than the presence–absence matrix itself is required); however, its use as a null model lacks a mechanistic foundation. In statistical hypothesis testing, the null should instead represent the same system with the ecological mechanisms of interest excluded (Bausman, 2018; Molina and Stone, 2020; Zhang, 2020). While data randomisation is useful for challenging intuitions about how unusual an observed pattern really is (Zhang, 2020), ecologists could benefit from a mechanistic model framework that allows more rigorous quantitative testing of the mechanisms driving species occurrence patterns.

One possible source of mechanistic null models is neutral theory (Bell, 2005; Molina and Stone, 2020; Munoz and Huneman, 2016; Rosindell et al., 2012). Neutral models assume an equivalence between species and focus on predicting broad statistical patterns of biodiversity (e.g., species-abundance distribution (Missa et al., 2016; Ofiţeru et al., 2010; Rosindell et al., 2012; Volkov et al., 2003)). In particular, the spatially implicit neutral model (Hubbell, 2001) takes a macroecological perspective (e.g., MacArthur and Wilson, 1967) in which communities are treated as subsets of a regional community and are formed by broad-scale stochastic processes: immigration and stochastic extinction due to drift (Munoz and Huneman, 2016). The advantages of the spatially implicit neutral model are that it is mechanistic (Alonso et al., 2006) and analytically and computationally tractable (Rosindell et al., 2012) to a degree that may make null-model testing feasible. There also exist a variety of "niche–neutral" models that incorporate some degree of niche mechanisms into the mathematical framework of neutral theory (e.g., Chisholm and Pacala, 2010, 2011).

A neutral model or a related niche model could serve as a baseline, representing expectations based on established processes, to help apportion causal impact between neutral and other factors towards producing the observed patterns (Bausman, 2018). Although the model used will necessarily be incomplete, it may still provide a causally relevant explanation for biodiversity in terms of the main processes involved (Odenbaugh, in press), for example, if co-occurrence patterns are mainly structured by dispersal and randomness (cf., Bell, 2001). Further, the ways in which the model predictions differ from observed data may reveal important mechanisms missing from the model (Marquet et al., 2014; Matthews and Whittaker, 2014; Rosindell et al.,

2012), and may provide a backdrop against which more complex mechanisms can be tested (Munoz and Huneman, 2016; Volkov et al., 2003).

One barrier to using a mechanistic model as a null model is that model parameterisation requires more data (e.g., Gotelli and McGill, 2006). Data randomisation requires only the data itself, i.e., the presence-absence matrices, whereas mechanistic macroecological models typically require estimates of other parameters, such as the immigration rate and metacommunity diversity (Adler, 2004; Chisholm et al., 2016; Wootton, 2005). Although a few studies have explored how occurrence patterns under neutral models differ from random (Bell, 2005; Gotelli and McGill, 2006; Ulrich, 2004; Ulrich et al., 2017), no study to date has used a neutral or related mechanistic model as a null model when analysing empirical data.

Here, we explore the potential of a niche–neutral model (Chisholm et al., 2016) to overcome the parameterisation problem and provide a baseline for exploring when species occurrence patterns can be explained by just a few simple mechanisms. Our application is to bird species on islands in an archipelago in Southeast Asia (Sin et al., 2021). The niche–neutral model used is an extension of Hubbell's spatially implicit neutral model. Each patch within the model comprises the same set of distinct equal-sized non-overlapping habitats or niches, and species obey neutral dynamics within each niche (Chisholm et al., 2016). To efficiently generate randomised presence–absence data from this model, we extend an existing sequential sampling scheme (Etienne, 2007). We found that the occurrence patterns were inconsistent with the null distributions generated by the most basic version of this model, and that more mechanistic detail was required to improve fidelity to the data. The approach thus leads to better mechanistic understanding of the patterns observed (Zhang, 2020).

2. Methods

2.1. Niche–neutral model

The niche–neutral model (Chisholm et al., 2016; Chisholm and Pacala, 2010, 2011) comprises a set of n^* distinct non-overlapping niches within each of which the community dynamics are neutral. In the local island community, the total number of individuals is equal to J, and in each time step an individual is chosen at random to die. The dead individual is replaced with probability m by an immigrant from the metacommunity, and with probability $1 - m$ it is replaced by the offspring of another randomly chosen individual that shares its niche.

The metacommunity is also divided into n^* niches, and the total number of individuals is equal to J_M. The metacommunity undergoes a similar birth–death dynamics to the island community, with the key difference that species input into the metacommunity occurs via point speciation, with per-capita probability ν, rather than via immigration. The island community dynamics operate on a much faster timescale than the metacommunity such that the metacommunity can be assumed static on timescales relevant to the local community. A composite parameter, named the fundamental biodiversity number $\theta = (J_M(1-\nu))/\nu$, characterises the diversity of the metacommunity at equilibrium.

The niche–neutral model explains the observed relationship between island area and species richness in archipelagic data (Chisholm et al., 2016), in particular the small-island effect (e.g., Lomolino and Weiser, 2001). In the model, equilibrium species richness generally increases with island area, driven by the effect of area on the immigration–extinction stochastic equilibrium, as in classic island biogeography theory (MacArthur and Wilson, 1967). But among small islands

the model departs from classic island biogeography theory by exhibiting a nearly flat phase that is consistent with the small-island effect seen in archipelagic data. In the small-island phase, equilibrium species richness is not driven by immigration–extinction dynamics but is instead roughly equal to the number of niches (Chisholm et al., 2016).

To compare the model to data, four parameters must be estimated (Chisholm et al., 2016): (1) the number of niches, n^*; (2) the per-capita immigration rate, m; (3) the fundamental biodiversity parameter, θ (Hubbell, 2001); and (4) the density of individuals per unit area, ρ. The last of these can often be obtained from independent density estimates for the species group concerned, but the other three parameters must typically be fitted to the archipleago richness data via least squares or a related method.

In the simplest version of the niche–neutral model, the niches are all assumed to have equal size, and all islands are assumed to have the same number of niches, n^*. The per-capita immigration rate m is assumed to either be constant with area (which implies that the total number of immigrants per unit time scales linearly with area), or to be proportional to one over the square root of area (which implies that the total number of immigrants per unit times scales linearly with island perimeter).

2.2. Sequential sampling scheme

To efficiently simulate presence–absence data from the niche–neutral model, we extended the work of Etienne (2005, 2007) to create a model-equivalent sequential sampling scheme. The sampling scheme generates a random sample of the species identity of every individual in the archipelago, and from this a presence–absence matrix can be derived. Although determining the species identity of every individual is more computationally expensive than data randomisation, it is more efficient than a forward-time simulation of the niche–neutral model because we only need to sample the present-day individuals plus their immigrant ancestors rather than account for every individual who was historically present.

Our sequential sampling scheme determines each individual's species identity by sampling events in the coalescence tree (Etienne and Olff, 2004; Rosindell et al., 2008). At each step backwards in time, a lineage may either coalesce with another lineage, indicating that all individuals both lineages share a common ancestor and thus belong to the same species, or it may conclude in a common ancestor that represents a new species. In the context of an island community, these new species represent immigrants from the metacommunity (mainland and other islands), while in the context of the metacommunity, they represent an originating point-speciation event. Thus, the species identity of all individuals in a lineage is determined when the lineage is traced back to a speciation event. The neutral dynamics within each niche of the niche–neutral model are equivalent to the coalescence model (Rosindell et al., 2008), which, in turn, is equivalent to a sequential sampling scheme (Branson, 1991). Thus, our sequential sampling scheme produces samples that are statistically identical to those of conventional forwards-time simulations but orders-of-magnitude more efficiently.

Our sampling scheme builds on three prior schemes, which we will discuss in order of increasing complexity: (1) the metacommunity, (2) the island community, and (3) multiple island communities.

To sample from the metacommunity, we model the coalescence process by drawing each individual's species-identity-determining common ancestor one at a time. This can be physically

represented by an urn model (Fig. 1a). The process is governed by the fundamental biodiversity number θ (Hubbell, 2001), which increases with community size and speciation rate and determines overall species diversity. Each draw corresponds to an event in the coalescence tree, either a coalescence (Eq. 1a) or a point-speciation event (Eq. 1b). For the jth individual drawn, the probability of it belonging to a previously sampled species is

$$P(j \text{ is a previously sampled species }) = \frac{j-1}{\theta+j-1}, \tag{1a}$$

while the probability of it belonging to a new species is

$$P(j \text{ is a new species}) = \frac{\theta}{\theta+j-1}. \tag{1b}$$

Eq. 1 is also known as the 'species generator' (Chapter 9, Hubbell, 2001; Rosindell et al., 2011), and it was derived first in the genetics literature from the analogous neutral genetic process (Lemma 2.1 Donnelly, 1986).

To simulate an island community, a model with two urns is used: one tracks coalescence events on the island, while the other tracks immigrants from the metacommunity (Etienne (2005); Fig. 1b). On the island, new species arrive by immigration instead of speciation, which is governed by an immigration rate $I = \frac{m(J-1)}{1-m}$ analogous to the role of θ previously. These immigrants are treated as a random sample from the metacommunity and modelled as described above (i.e., Fig. 1a). This simplifies reality by approximating the distribution of immigrant identities over time by a sample drawn at a single point in time (which is why the arrows representing immigration in Fig. 1b all start at the base of the immigrant-coalescence tree).

To simulate multiple islands, each island is represented by one urn, while a shared immigrant ancestors urn serves as the common source for all islands (Etienne (2007); Fig. 1c). This setup aligns with the niche–neutral model's assumption that islands are connected by immigration to the same metacommunity, but not each other (a necessary mathematical simplification). Although the sequence of draws from the ancestor urn corresponds to a temporal sequence of events in the coalescence tree, for a given final outcome, the probability of every temporal sequence is the same. As a result, Etienne (2007) was able to streamline the scheme by sampling the islands sequentially without altering the probability distribution of outcomes.

To simulate the complete niche–neutral model, the multiple-island simulations (see Fig. 1c) are replicated for each niche $1, \ldots, n^*$ independently (see SI A). Our new sequential sampling scheme also allows us to extend the scenario presented by Chisholm et al. (2016) by allowing the number of niches and the immigration rates to vary between islands.

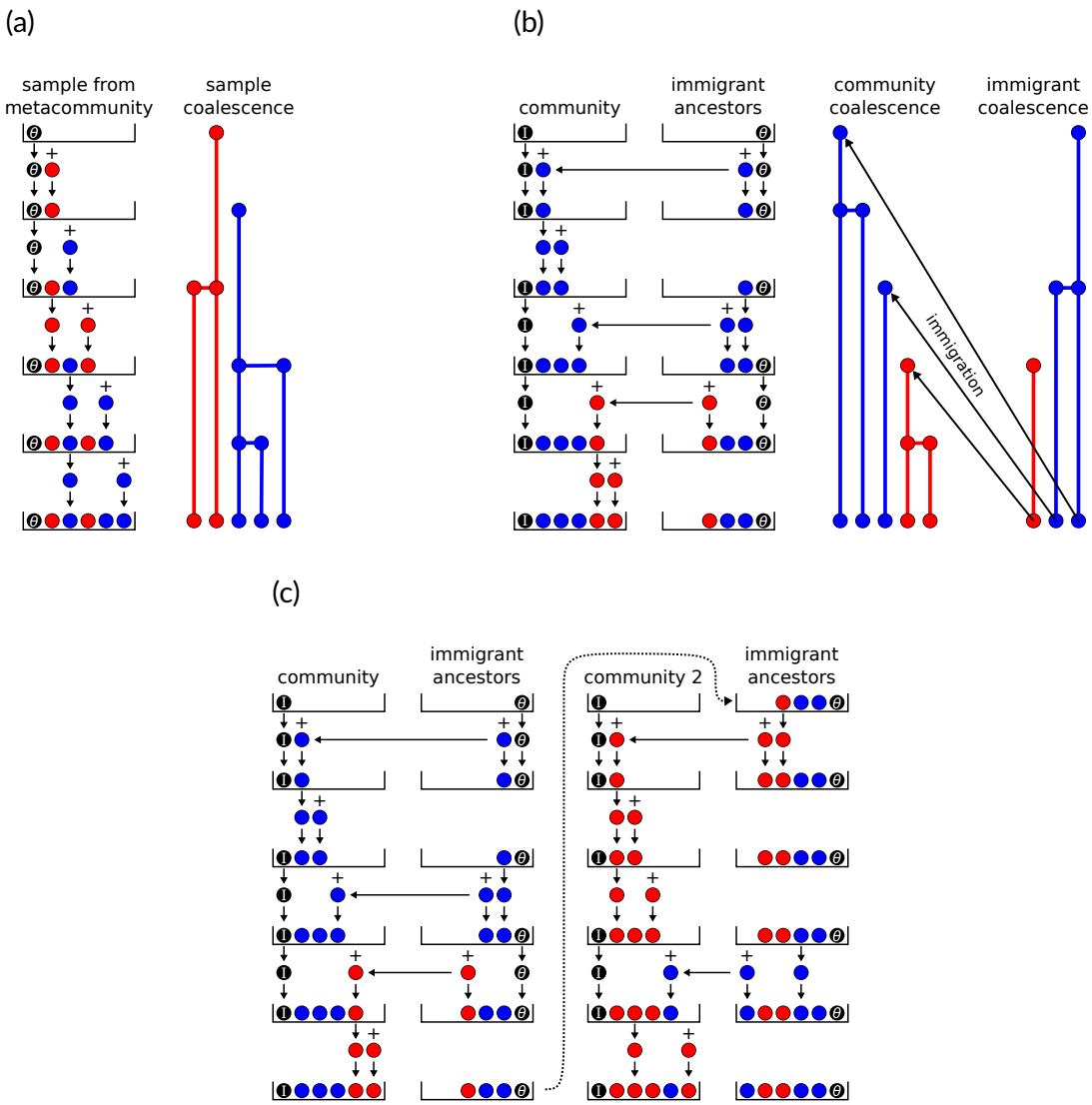

Figure 1 – Examples of coalescence models represented by urn models generating samples from (a) a metacommunity, (b) a local community, and (c) a collection of communities sharing a metacommunity (e.g., archipelago). Each coloured ball represents an individual with its colour indicating its species identity, and all coloured balls have equal weight 1 reflecting their equivalence in the neutral model. Black balls represent an event that generates a new species, either immigration with weight I or speciation with weight θ. **(a)** The black speciation ball θ is placed in the urn. At each step, a ball is drawn and returned, and a new ball is added depending on the colour of the ball drawn. If a coloured ball is drawn, that represents branching in the coalescence tree (Eq. 1a), and a ball of the same colour is added. If the black ball is drawn, that represents speciation (Eq. 1b), and a ball of a new colour is added. **(b)** A black immigration ball I is placed in the 'community' urn and speciation ball θ is placed into the 'immigrant ancestors' urn. At each step, the 'community' urn is sampled first. If a coloured ball is drawn, a ball of the same colour is added to the 'community' urn only. If the immigration ball is drawn, a draw from the 'immigrant ancestors' urn determines the colour of the ball to be added to both urns: a coloured ball adds the same colour to both urns, the speciation ball adds a new colour to both urns. **(c)** There is a single 'immigrant ancestors' urn and a separate 'community' urn for each local community. 'Community' urns are filled sequentially according to the same rules as (b). Once a community is filled, the 'immigrant ancestors' urn and its current state is passed to the next community.

2.3. Case study

We applied the methods above to the presence–absence matrix obtained from a comprehensive survey of birds on 23 islands in the Riau region, Indonesia, a subset of data previously published by Sin et al. (2021). The dataset included a substantial number of small islands ($<$10 km^2) to capture the small-island effect as well as the classic increase phase of the species–area curve (Chisholm and Fung, 2022). Surveys were conducted by traversing the islands (Sodhi et al., 2010) and excluded species that were introduced, nocturnal, aquatic, pelagic, coastal, and those not breeding on the islands (i.e., migratory species and flyby birds of prey). Species inventories were available for two islands to supplement observations (Lingga: Sutari (2004) and Gibson-Hill (1952); Durian: Sieber (1926)). Inventory completeness was assessed using species accumulation curves (Oksanen et al., 2019), and islands with poor sampling were excluded (Weigelt and Kreft, 2013).

To parameterise the niche–neutral model, we approximated the density of birds as $\rho =$ 1700 birds per km^2, which was chosen as a compromise between values reported in the literature: 1259 individuals/km^2 for resident birds in old secondary forests (Castelletta et al., 2005), and 2115 individuals/km^2 average density for primary rainforest habitats (Sheldon and Styring, 2011). The other three model parameters — number of niches n^*, immigration rate m, and fundamental biodiversity number θ — were fitted to the island species–area data using the methods described in Chisholm et al. (2016).

First, we created a homogeneous model with parameter values held constant across islands. We then identified discrepancies between the data and model predictions to pinpoint which excluded mechanisms were likely important in producing the observed pattern. Using this information, we refined the model by allowing m and n^* to vary across islands.

To quantify co-occurrence patterns in both our data and models, we used the C-score metric (Stone and Roberts, 1990) calculated using the `EcoSimR` package for `R` (Gotelli et al., 2015). We quantified nestedness using the NODF metric (Almeida-Neto et al., 2008), which we calculated using code adapted from Strona et al. (2018). In arranging the presence–absence matrix, species rows were sorted by decreasing fill, while columns representing islands were ordered by island size, which reflected the anticipated effect of island area on species richness.

To increase comparability between real and simulated data, we also computed each metric's standardised effect size (SES) (Gotelli and McCabe, 2002; Ulrich et al., 2018). To calculate SES, we used a fixed–fixed randomisation algorithm. Let V_{raw} denote the raw metric value (C-score or NODF) calculated from a presence–absence matrix, which may originate from real data or synthetic data sampled from the niche–neutral model; let $\overline{V}_{\text{rand}}$ denote the mean value when the matrix is randomised according to the fixed–fixed algorithm; and let σ_{rand} denote its standard deviation. Then

$$\text{SES } V = \frac{V_{\text{raw}} - \overline{V}_{\text{rand}}}{\sigma_{\text{rand}}}.$$

Note that the randomisation used to calculate the SES scores here does not represent a "null" hypothesis, but rather it serves to scale the metrics so their values can be more meaningfully compared. It is recommended to use SES when comparing across studies (Gotelli and McCabe, 2002) particularly to account for differences in species richness (Ulrich et al., 2018). We chose the fixed–fixed randomisation based on the findings that it produces SES C-scores and SES NODFs that did not vary strongly with richness (Ulrich et al., 2018) and has favourable statistical properties when used with C-score (Gotelli, 2000) and other nestedness metrics (Ulrich and Gotelli,

2007) (but see Molina and Stone (2020) for a conceptual critique of those favourability criteria). The fixed–fixed randomisation was performed using the curveball algorithm (Strona et al., 2014) implemented in the `EcoSimR` package for `R` (Gotelli et al., 2015); however, we added two additional steps to account for autocorrelation between samples. First, we visually inspected the traceplots (i.e., plots of the metric value versus iteration number) to verify that the burn-in length was sufficient. Second, we calculated the effective sample size using the `CODA` package for `R` (Plummer et al., 2006) to ensure that metrics were calculated from an effective sample size of at least 1000 samples.

3. Results

3.1. The homogeneous niche–neutral model versus observed data

The homogeneous niche–neutral model provided a good fit to the species–area relationship. In particular, it captured the small-island effect ($R^2 = 0.81$, Fig. 2a). However, the model did not reproduce other aspects of the data well (Table 1 "Homogeneous"; see also SI B). The total archipelagic richness predicted by the model was too low (Fig. 2b), and when the sequential sampling scheme was used to explore the variation in the model, most island richness values fell outside of the 95th percentile of samples (16 of 23, Fig. 2a).

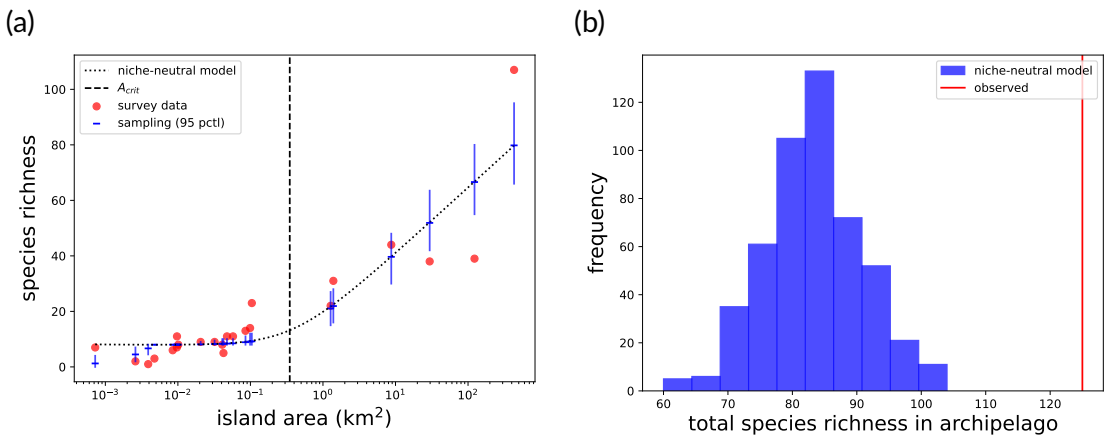

Figure 2 – A comparison between the survey data and the predictions of the niche–neutral model. (a) The SAR, best-fitting niche–neutral model predictions (solid line), and 95 percentiles drawn from the sequential sampling scheme. (b) The total archipelago richness (line) compared to the distribution of total richnesses from the sequential sampling scheme. The best-fitting niche–neutral model parameters were: $n^* = 8$, $\theta = 10.3$, $m = 3.6 \times 10^{-3}$, $R^2 = 0.81$, resulting in $A_{\text{crit}} = 0.35$ km^2.

The raw C-score for the data was within the range of values obtained from the model (data raw C-score 1.28; model raw C-score 1.19, 95% prediction interval (PI) $[0.835, 1.81]$, $p = 0.33$) (Fig. 3a). However, the SES C-score for the data was larger than the model values, indicating segregated patterns in the data (data SES C-score 6.60; model SES C-score -0.476, 95% PI $[-1.76, 1.40]$, $p < 0.01$) (Fig. 3b).

The niche–neutral model predicted higher NODF scores than observed in the survey data. The disagreement between model and data was stronger for the raw NODF scores (data raw NODF 57.2; model raw NODF 82.5, 95% PI $[80.2, 84.2]$, $p < 0.01$; Fig. 3c) than in the standardised scores (data SES NODF -2.67; model SES NODF -0.932, 95% PI $[-4.24, 0.977]$, $p = 0.13$; Fig. 3d).

Table 1 – Observed versus predicted species-occurrence metrics from the niche–neutral models. In the fitted model, both niche diversity and immigration rates were varied across islands as proof-of-principle that the data could be consistent with niche–neutral processes.

Metric	Observed	Niche–neutral model predictions [95-percentile](lower-tail prob.)					
		Homogeneous			Fitted		
spp. richness	125	83.3	[69, 99]	($p < 0.01$)	(tuned to match observed)		
raw C-score	1.28	1.19	[0.835, 1.81]	($p = 0.33$)	1.64	[1.28, 2.07]	($p = 0.03$)
SES C-score	6.60	−0.476	[-1.76, 1.40]	($p < 0.01$)	6.16	[4.02, 8.80]	($p = 0.37$)
raw NODF	57.2	82.5	[80.2, 84.2]	($p < 0.01$)	60.3	[55.4, 65.5]	($p = 0.13$)
SES NODF	−2.67	−0.932	[-4.24, 0.977]	($p = 0.13$)	−4.39	[-7.78, -1.49]	($p = 0.13$)

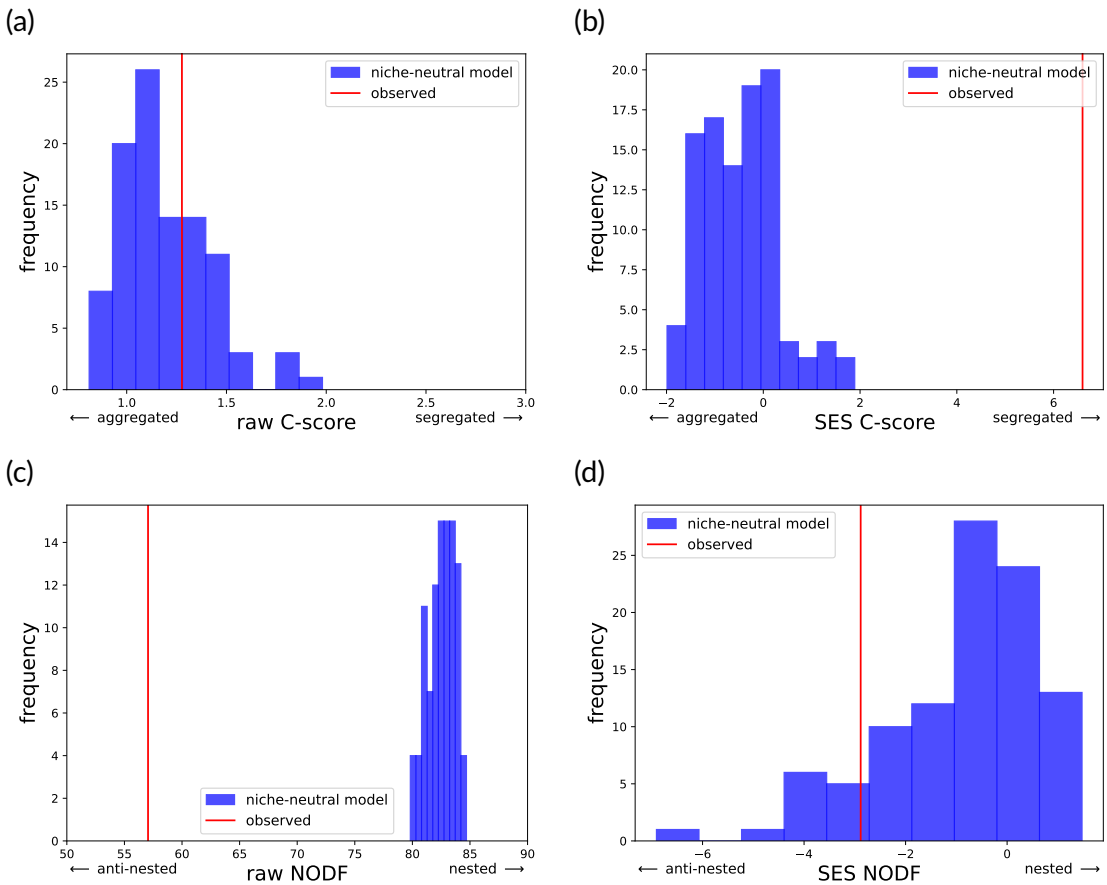

Figure 3 – Frequency histograms comparing the distribution of characteristics of the fitted niche–neutral model (100 draws, blue) to the observed survey data (red): (a) raw C-score, (b) SES C-score, (c) raw NODF, (d) SES NODF. The SES is calculated in comparison to fixed–fixed randomised matrices, generated using the sim9 algorithm with an effective sample size of at least 1000.

A traditional data-randomisation approach produces results that are broadly consistent with the comparison to the niche-neutral model above (Table 2). The traditional approach finds that the survey data was significantly segregated and anti-nested compared to the fixed-fixed randomisation.

Table 2 – Null-model analysis using randomised data for bird occurrence in Riau archipelago, including standardised effect sizes (SES) and associated lower-tail probabilities (p).

Metric	Observed	Mean randomised		Interpretation
C-score	1.28	1.05	(SES = 6.60; $p < 0.0003$)	Species are segregated across islands.
NODF	57.2	59.0	(SES = -2.67; $p = 0.01$)	Archipelago assemblage is anti-nested.

3.2. Identifying mechanisms that can explain observed species occurrence

The advantage of the niche–neutral model approach is that it provides a framework for incorporating additional mechanisms to improve data fitting. One limiting assumption of the niche–neutral model that may cause discrepancies with data is the homogeneity assumption, where all islands have the same number of equal-sized niches n^* and the same per-capita immigration parameter m. In subsection 3.2.1, we address this by relaxing the homogeneity assumption and allowing niche structure to vary across islands. In subsection 3.2.2, we modify the model by permitting per-capita immigration to vary across islands. Finally, in subsection 3.2.3, we combine these adjustments, allowing both niche structure and per-capita immigration rates to vary across islands.

3.2.1. Allowing niche structure to vary with island area. We created a synthetic archipelago of 16 islands with areas distributed evenly on a log scale (carrying capacity range $J = 10^2$ to 10^5). We generated biodiversity for these islands using the novel sequential sampling algorithm for the niche–neutral model, with parameter values chosen to capture the small-island effect (blue line, Fig. 4a). We initially kept m and n^* constant across islands, maintaining the homogeneity assumption. This homogeneous synthetic archipelago exhibited similar occurrence patterns to the homogeneous model fitted to the survey data above (Fig. 3), with a positive mean raw C-score (2.09, 95% PI $[1.34, 2.80]$) and a slightly negative mean SES C-score (-0.506, 95% PI $[-1.84, 1.88]$) and high nestednesss when measured by raw NODF (75.6, 95% PI $[69.5, 81.4]$) but slight anti-nestedness when measured by standardised NODF (-1.32, 95% PI $[-3.73, 0.730]$).

We then relaxed the homogeneity assumption by allowing niche diversity to increase with island size. We explored 10 scenarios; we focus on two here (Fig. 4b), and the others are available in the code repository.

In Scenario 1, the four largest islands in the archipelago have eight additional niches (all niches still equal in area) (orange line, Fig. 4a). This increased segregation compared to the baseline when using standardised scores (mean raw C-score 1.31, 95% PI $[0.797, 1.94]$; mean SES C-score 2.94, 95% PI $[1.28, 5.96]$) (Fig. 4b) and had small negative effects on raw nestedness (mean raw NODF 74.3, 95% PI $[69.9, 78.4]$) and anti-nestedness (mean SES NODF -3.40, 95% PI $[-5.98, -0.499]$). Qualitatively similar results were found in other scenarios.

To verify that the segregation effect in Scenario 1 was not simply because of the increased richness on larger islands, we modified the homogeneous model by fitting the immigration parameters (m) individually to larger islands to match the increase in richness in Scenario 1. This increased segregation, but not by as much (mean raw C-score 0.507, 95% PI $[0.263, 0.788]$, mean SES C-score 0.269, 95% PI $[-1.56, 2.14]$).

In Scenario 2, we split the ten niches into five coastal and five interior types, and assumed that the width of coastal niches around the perimeter of the island (assumed circular) would remain

constant (green line Fig. 4a). This reflects the notion that, as island size increases, the area covered by coastal habitat types will grow more slowly than inland types (e.g., forest, Davidar et al., 2001). This scenario also increased segregation compared to the homogeneous model (mean raw C-score 3.77, 95% PI [2.51, 5.19]; mean SES C-score 2.25, 95% PI [−0.150, 4.88]) (Fig. 4b). Nestedness decreased and thus the archipelago remained anti-nested according to the standardised scores (mean raw NODF 64.0, 95% PI [58.3, 69.7]; mean SES NODF −1.68, 95% PI [−3.84, 0.900]). Thus, the main effect of niche structure was to increase segregation with a smaller secondary effect of decreasing nestedness (additional nestedness results in SI C).

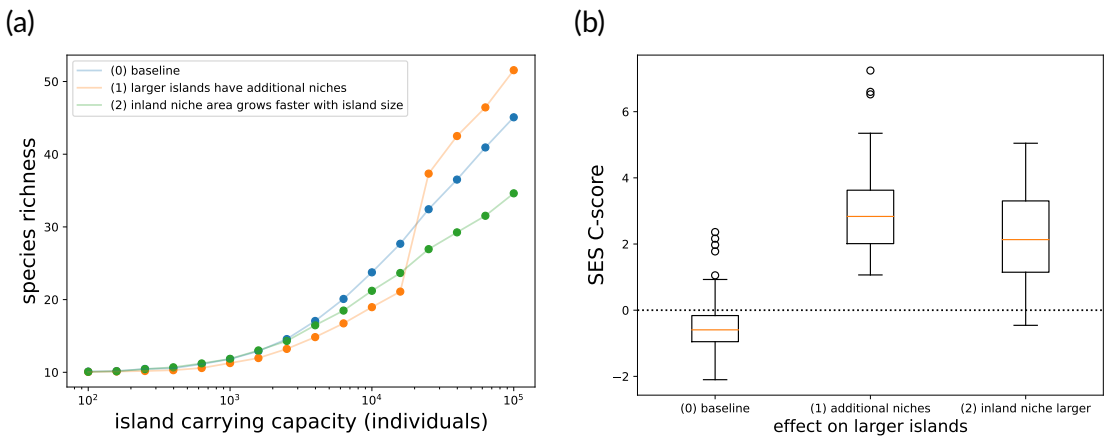

Figure 4 – Effects of changing niche diversity with island area under three scenarios. In the baseline scenario (0), the homogeneous scenario, each island has $n^* = 10$ equal-area niches and niche areas grow proportionally with island area. In scenario (1), the four largest islands have eight additional niches each. In scenario (2), the ten niches are split into coastal and inland types, and the area of coastal niches grows more slowly with island area. (a) The relationship between species richness and carrying capacity, which is a proxy for island area, under the three scenarios. (b) The SES C-scores under the three scenarios. Model parameters were $\theta = 10$, $m = 7.5 \times 10^{-4}$. The results for each scenario represent 100 random model realisations using the novel sequential sampling algorithm.

3.2.2. Allowing per-capita immigration rates to vary across islands. We explored a version of the niche–neutral model in which per-capita immigration rates can vary across islands. Instead of a constant rate m, each island i has rate m_i. When fitting only to archipelago species–area data, this allows the model to be overfitted because there are more parameters than data points, and so each island's richness can be captured exactly. We used the θ and n^* parameter values from the original model fit and tuned the value of m_i for each island i to match observed species richness. We found that this brought the nestedness results into closer agreement with the survey data (Fig. 5): the observed raw nestedness remained lower than the model (data raw NODF 57.2; model raw NODF 70.2, 95% PI [63.3, 76.2], $p < 0.01$) but the standardised nestedness was within the range of the model (data SES NODF -2.67, model SES NODF −1.57, 95% PI [−4.45, 0.653], $p = 0.19$). Meanwhile, the data remained segregated compared to the new model (model mean raw C-score 0.818, 95% PI [0.498, 1.18], $p < 0.025$; mean SES C-score: −0.753, 95% PI [−1.74, 0.660], $p < 0.025$).

3.2.3. Varying both niche diversity and immigration rates across islands. Our investigations in subsections 3.2.1 and 3.2.2 indicate that varying niche diversity across islands can lead to increased

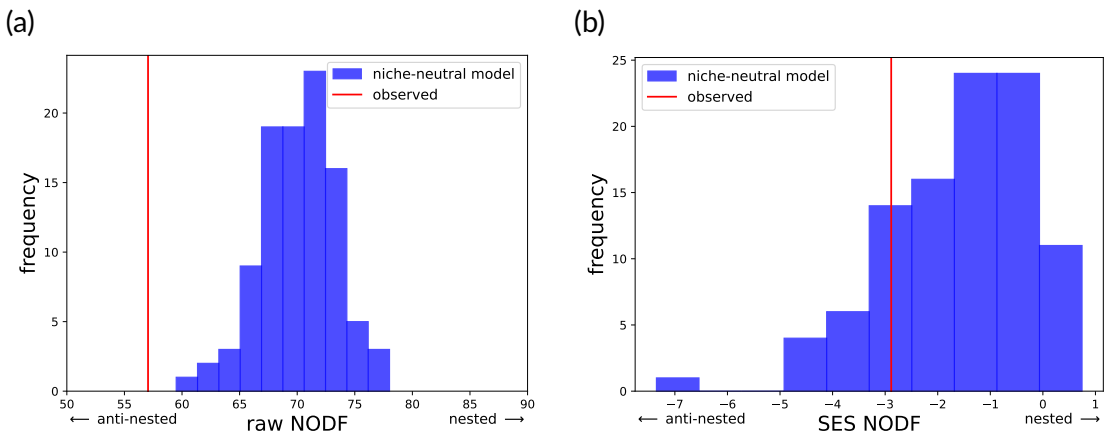

Figure 5 – Frequency histograms comparing nestedness metrics between the observed data (red) and fitted niche–neutral models (100 draws, blue): (a) Raw NODF; (b) SES NODF. The niche–neutral model was fitted by allowing immigration rate m to vary to match the observed species richness on each island. The SES is calculated in comparison to fixed–fixed randomised matrices, generated using the sim9 algorithm with an effective sample size of at least 1000.

segregation in co-occurrence patterns and that varying per-capita immigration rates across islands can lead to increased nestedness. We put these two observations together to test if a modified niche–neutral model with varying niche diversity and immigration rates across islands could plausibly reproduce the observed nestedness and segregation patterns. We added the assumption that the number of niches increased with island size, with some niches occurring exclusively on large islands, and we tuned island-specific parameters (n^* on small islands and m on large islands) to match the species richness pattern across islands. Through trial and error, we found parameter values (details in SI D) that brought the niche–neutral model's segregation and nestedness patterns into closer agreement with the survey data (Fig. 6; Table 1 "Fitted").

4. Discussion

We have introduced a novel sequential sampling algorithm designed to efficiently explore species occurrence patterns using simple mechanistic models, specifically simple neutral and niche–neutral models. This mechanistic approach can identify patterns of segregation and nestedness in community data, similar to traditional data-randomisation approaches. However, when predictions do not match observations, the mechanistic models have the advantage that they suggest mechanisms that might be missing from the model and provide a framework to incorporate them into refined models.

Our novel sequential sampling algorithm offers a significant advancement in the efficiency and application of neutral models for studying occurrence patterns. Neutral models have been valued in previous studies for their parsimony, simplicity, and mechanistic basis: they have minimal assumptions, are based on first principles, and can generate a large number of predictions per free parameter (Marquet et al., 2014). However, previous applications to co-occurrence analysis have relied on computationally heavy forwards-in-time simulations (Bell, 2005; Gotelli and McGill, 2006; Ulrich, 2004; Ulrich et al., 2017). Our algorithm facilitates more-efficient applications by massively accelerating simulations. It is based on the coalescence approach to neutral modelling (Rosindell et al., 2008), which involves running simulations backwards in time. This

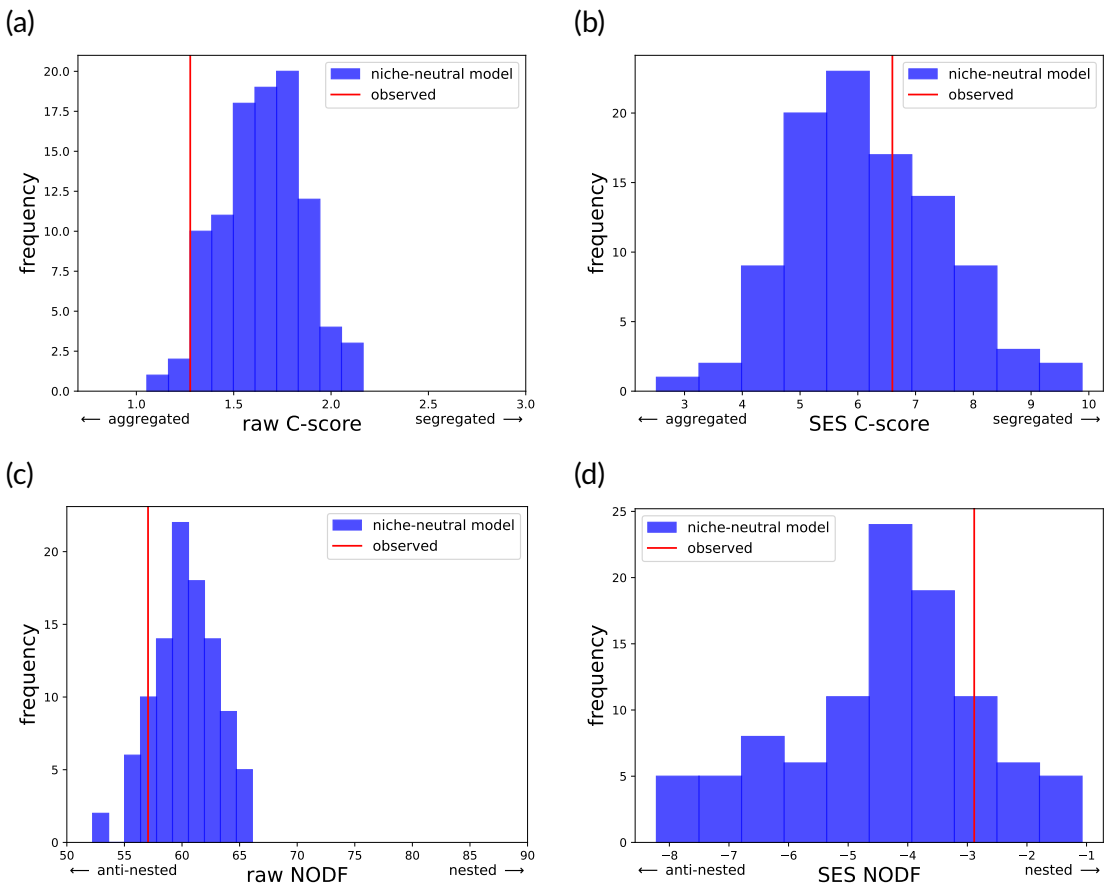

Figure 6 – Frequency histograms comparing the distribution of occurrence characteristics of the observed survey data (red) to a niche–neutral model with niche diversity and immigration parameters varying across islands and parameters tuned to match the data (100 draws, blue): (a) raw C-score, (b) SES C-score, (c) raw NODF, (d) SES NODF. Details are provided in SI D.

backward-in-time approach relies on the per-capita equivalence of species within niches, an assumption that allows for efficient simulation. While there are many mechanistic models outside the neutral framework that could be used as null models for occurrence analysis, most lack a similarly efficient method for simulating a large number of stochastic outcomes. Therefore, a prudent general approach to analysing species occurrence patterns may be to start with a computationally efficient neutral or niche–neutral model, and once key missing mechanisms have been identified, move on to more tailored, albeit likely more computationally expensive, models (e.g., see a series of papers disentangling species distribution patterns in ground beetles: Ulrich and Zalewski (2006, 2007) and Zalewskia and Ulrich (2009)).

In our application to birds in Riau, we demonstrated how starting with a simple niche–neutral model and then increasing its realism identified potential mechanisms behind species occurrence patterns. The simplest version of the niche–neutral model—the homogeneous model—could fit the species–area relationship, but it failed to accurately describe total species richness and occurrence patterns. Specifically, the data showed greater richness, greater segregation and anti-nestedness compared to the model. A traditional data-randomisation analysis would have led to qualitatively similar results (Table 2), but the advantage of the mechanistic approach is that it also provided a quantitative framework for testing hypotheses about missing mechanisms. In our case

study, we showed that by relaxing the homogeneity assumption and increasing the model's realism, the model better matched observations. Specifically, species segregation increased when we allowed habitat diversity to increase with island size, and nestedness decreased when we tuned model parameters to more accurately reproduce the richness pattern across islands. Thus, we conclude that differences in habitat diversity and per-capita immigration rates across islands, while not necessary for explaining coarse patterns like the island species–area relationship, may be crucial for explaining occurrence patterns.

Varying niche diversity across islands increased species segregation because islands with different niches (scenario (1) in Fig. 4) or different-sized niches (scenario (2)) tend to have fewer species in common than those with homogeneous niches (scenario (0)). Previous studies have noted that the positive effect of island area on species richness may be partially mediated by habitat diversity (Kohn and Walsh, 1994), particularly if large islands contain key habitats (e.g., rainforests, Davidar et al., 2001) and the species group under consideration includes habitat specialists (Ricklefs and Lovette, 1999). Scenario (2) is particularly appealing due to its parsimony: it simply adds the biologically reasonable assumption that coastal niches grow linearly with island perimeter, while other niches grow linearly with island area. Our exploration of these scenarios is intended as illustrative of this general idea and demonstrate our approach's ability to identify potential mechanisms driving occurrence patterns. Further work is needed to identify the niche processes actually increasing segregation in Riau, which may require going beyond the simple non-overlapping niche structure of the niche–neutral model. Previous authors have noted that it is unclear exactly what the "habitats" that influence island species diversity represent (Triantis et al., 2003, 2006). The n^* parameter in our model, which determines the number of niches, is perhaps better thought of as an effective parameter that describes how habitat diversity affects species diversity.

It is also retrospectively clear why varying per-capita immigration rates across islands decreased nestedness. In the homogeneous model, species on smaller islands are typically subsets of larger islands because smaller islands have both lower total immigration rates (per-capita immigration rate times area) and higher extinction rates. Allowing immigration rates to vary idiosyncratically across islands weakens the dependence of occurrence patterns on area and thus reduces nestedness. A limitation of our exploration of this phenomenon is that we allowed the per-capita immigration rate for each island to vary independently, leading to an over-fitted model. Therefore, our analysis should be viewed as simply proof-of-principle that inter-island variation in per-capita immigration rates reduces nestedness. A more detailed analysis would consider the mechanistic basis for this variation, which we conjecture is mainly due to differences in island isolation. For example, Durian and Ngenang, although smaller than Temiang and Sebangka (Fig. D.1), have higher species richness because they are closer to the large land masses of Sumatra and continental Asia, respectively, and larger nearby islands may also play a role. Quantifying island isolation and incorporating it when estimating per-capita immigration rates to each island would be a more parsimonious way to explore this phenomenon, although we acknowledge that parameterising neutral and similar models is not easy even in the homogeneous case (e.g., Gotelli and McGill, 2006).

Previous workers have suggested using neutral models as the null model for species occurrence analysis; however, we expect that, in most cases, neutral models will be easily ruled out as exact models at the species-richness pattern level. In our case study, although the niche–neutral

model could explain the overall trend in the island species–area relationship, it exhibited insufficient variability to explain the individual observed island richnesses (Fig. 2a). Therefore, its failure as an exact descriptor is clear even before considering higher-order metrics such as segregation and nestedness. We expect this situation will be common, and the real value of neutral models is as an exploratory framework as described above.

Previous work with various neutral-style models have observed a similar lack of variability. For example, Louca et al. (2016) fitted a neutral model designed for microbial systems. Their model reproduced the positive relationship between abundance and detection frequency, analogous to our positive relationship between area and species richness. However, they found that a large proportion of their taxonomic groups fell outside the probable range of model predictions, just as we found that 16 of 23 observed island richnesses fell outside the model's 95% prediction intervals (Fig. 2a). Similar results are also obtained from random-placement models, which simulate passive sampling by assuming that the probability of a species' presence at a site depends only on its relative regional abundance (Coleman, 1981; Coleman et al., 1982; Dolman and Blackburn, 2004). These models are known to produce nested patterns independent of other ecological mechanisms (Cutler, 1994; Wright et al., 1998) and are therefore often used as a preliminary test before more detailed nestedness analysis (Worthen et al., 1998). However, workers often find that the observed site species richness falls outside the random-placement model bounds and therefore rule out passive sampling as an explanation for the data (e.g., Calmé and Desrochers, 1999; González-Oreja et al., 2012; Tan et al., 2021; Wang et al., 2010).

Our niche–neutral model retains most of the assumptions of neutral models, and thus, these previous studies and our results tell a consistent story: while neutral models and their ilk may reproduce coarse biodiversity patterns, their predictions lack the variability of empirical data. Instead, we suggest that neutral models may be more useful for testing plausible 'what if?' scenarios

We suggest two general reasons why neutral models predict less variability than empirical data. Firstly, these models typically ignore spatial variability in model parameters (c.f. Dolman and Blackburn, 2004) as shown by the improved fit to occurrence patterns when we allowed niche diversity and per-capita immigration rates to vary with island area (Figs. 5–6). Secondly, these models ignore temporal variability in model parameters, which in nature is important because of temporal environmental stochasticity (Fung et al., 2016). In contrast, the only type of stochasticity in neutral models and our niche–neutral model is demographic stochasticity, which leads to much smaller fluctuations in species abundances over time than does environmental stochasticity. For these reasons, we expect that many future applications of our niche–neutral null model approach to analysing occurrence patterns will reveal higher-than-expected segregation and lower nestedness, as found here.

We conclude that simple models based on neutral models can be useful for exploring species occurrence patterns. Our novel sequential sampling scheme facilitates such applications by accelerating simulations of neutral and related models. To these ends, we hope that the code and tutorials we have provided for the algorithm (SI E) will be useful to future workers in this area.

5. Acknowledgements

We thank Tak Fung for helpful comments.

6. Funding

This research was funded by a Singapore Ministry of Education Tier 2 Grant: R-154-000-A59-112.

7. Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. NPK is a recommender of PCI Ecology.

References

Adler PB (2004). *Neutral models fail to reproduce observed species–area and species–time relationships in Kansas grasslands*. *Ecology* **85**, 1265–1272.

Almeida-Neto M, Guimaraes P, Guimaraes Jr PR, Loyola RD, Ulrich W (2008). *A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement*. *Oikos* **117**, 1227–1239.

Alonso D, Etienne RS, McKane AJ (2006). *The merits of neutral theory*. *Trends in Ecology & Evolution* **21**, 451–457.

Bausman WC (2018). *Modeling: neutral, null, and baseline*. *Philosophy of Science* **85**, 594–616.

Bell G (2001). *Neutral macroecology*. *Science* **293**, 2413–2418.

Bell G (2005). *The co-distribution of species in relation to the neutral theory of community ecology*. *Ecology* **86**, 1757–1770.

Branson D (1991). *An urn model and the coalescent in neutral infinite-alleles genetic processes*. *Lecture Notes – Monograph Series, Selected Proceedings of the Sheffield Symposium on Applied Probability* **18**, 174–192.

Calmé S, Desrochers A (1999). *Nested bird and micro-habitat assemblages in a peatland archipelago*. *Oecologia* **118**, 361–370.

Castelletta M, Thiollay JM, Sodhi NS (2005). *The effects of extreme forest fragmentation on the bird community of Singapore Island*. *Biological Conservation* **121**, 135–155.

Chisholm RA, Fung T (2022). *Examining the generality of the biphasic transition from niche-structured to immigration-structured communities*. *Theoretical Ecology* **15**, 1–16.

Chisholm RA, Fung T, Chimalakonda D, O'Dwyer JP (2016). *Maintenance of biodiversity on islands*. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20160102.

Chisholm RA, Pacala SW (2010). *Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities*. *Proceedings of the National Academy of Sciences* **107**, 15821–15825.

Chisholm RA, Pacala SW (2011). *Theory predicts a rapid transition from niche-structured to neutral biodiversity patterns across a speciation-rate gradient*. *Theoretical Ecology* **4**, 195–200.

Coleman BD (1981). *On random placement and species-area relations*. *Mathematical Biosciences* **54**, 191–215.

Coleman BD, Mares MA, Willig MR, Hsieh YH (1982). *Randomness, area, and species richness*. *Ecology* **63**, 1121–1133.

Connor EF, Simberloff D (1979). *The assembly of species communities: chance or competition?* *Ecology* **60**, 1132–1140.

Cutler AH (1994). *Nested biotas and biological conservation: metrics, mechanisms, and meaning of nestedness*. *Landscape and Urban Planning* **28**, 73–82.

Davidar P, Yoganand K, Ganesh T (2001). *Distribution of forest birds in the Andaman islands: importance of key habitats*. *Journal of Biogeography* **28**, 663–671.

Diamond JM (1975). *Assembly of species communities*. In: *Ecology and evolution of communities*. Ed. by M. L. Cody and J. M. Diamond. Cambridge, Massachusetts, USA: Harvard University Press, pp. 342–444.

Dolman AM, Blackburn TM (2004). *A comparison of random draw and locally neutral models for the avifauna of an English woodland*. *BMC Ecology* **4**, 1–12.

Donnelly P (1986). *Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles*. *Theoretical Population Biology* **30**, 271–288.

Etienne RS (2005). *A new sampling formula for neutral biodiversity*. *Ecology Letters* **8**, 253–260.

Etienne RS (2007). *A neutral sampling formula for multiple samples and an 'exact' test of neutrality*. *Ecology Letters* **10**, 608–618.

Etienne RS, Olff H (2004). *A novel genealogical approach to neutral biodiversity theory*. *Ecology Letters* **7**, 170–175.

Fung T, O'Dwyer JP, Rahman KA, Fletcher CD, Chisholm RA (2016). *Reproducing static and dynamic biodiversity patterns in tropical forests: the critical role of environmental variance*. *Ecology* **97**, 1207–1217.

Gibson-Hill C (1952). *A revised list of the birds known from the Rhio-Lingga Archipelago*. *Bulletin of the Raffles Museum* **24**, 344–380.

González-Oreja JA, De La Fuente-Díaz AA, Hernández-Santín L, Bonache-Regidor C, Buzo-Franco D (2012). *Can human disturbance promote nestedness? Songbirds and noise in urban parks as a case study*. *Landscape and Urban Planning* **104**, 9–18.

Gotelli, Hart, Ellison (2015). *EcoSimR: Null Model Analysis for Ecological Data*. https://doi.org/10.5281/zenodo.16636. URL: https://doi.org/10.5281/zenodo.16636.

Gotelli NJ (2000). *Null model analysis of species co-occurrence patterns*. *Ecology* **81**, 2606–2621.

Gotelli NJ, McCabe DJ (2002). *Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model*. *Ecology* **83**, 2091–2096.

Gotelli NJ, McGill BJ (2006). *Null versus neutral models: what's the difference? Ecography* **29**, 793–800.

Hubbell SP (2001). *The unified neutral theory of biodiversity and biogeography*. Vol. 32. Princeton, USA: Princeton University Press.

Kohli BA, Terry RC, Rowe RJ (2018). *A trait-based framework for discerning drivers of species co-occurrence across heterogeneous landscapes*. *Ecography* **41**, 1921–1933.

Kohn D, Walsh D (1994). *Plant species richness–the effect of island size and habitat diversity*. *Journal of Ecology*, 367–377.

Lomolino MV, Weiser M (2001). *Towards a more general species-area relationship: diversity on all islands, great and small*. *Journal of Biogeography*, 431–445.

Louca S, Jacques SM, Pires AP, Leal JS, Srivastava DS, Parfrey LW, Farjalla VF, Doebeli M (2016). *High taxonomic variability despite stable functional structure across microbial communities*. *Nature Ecology & Evolution* **1**, 1–12.

MacArthur R, Wilson E (1967). *The theory of island biogeography*. Princeton, NJ: Princeton University Press.

Marquet PA, Allen AP, Brown JH, Dunne JA, Enquist BJ, Gillooly JF, Gowaty PA, Green JL, Harte J, Hubbell SP, O'Dwyer J, Okie JG, Ostling A, Ritchie M, Storch D, West GB (2014). *On theory in ecology*. *BioScience* **64**, 701–710.

Matthews TJ, Whittaker RJ (2014). *Neutral theory and the species abundance distribution: recent developments and prospects for unifying niche and neutral perspectives*. *Ecology and Evolution* **4**, 2263–2277.

Meyer CF, Kalko EK (2008). *Bat assemblages on Neotropical land-bridge islands: nested subsets and null model analyses of species co-occurrence patterns*. *Diversity and Distributions* **14**, 644–654.

Missa O, Dytham C, Morlon H (2016). *Understanding how biodiversity unfolds through time under neutral theory*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150226.

Molina C, Stone L (2020). *Difficulties in benchmarking ecological null models: an assessment of current methods*. *Ecology* **101**, e02945.

Munoz F, Huneman P (2016). *From the neutral theory to a comprehensive and multiscale theory of ecological equivalence*. *The Quarterly Review of Biology* **91**, 321–342.

Odenbaugh J (in press). *Neutrality, Niche, and Nulls: Causal Relevance in Ecology*. In: *Biology and Causation*. Ed. by K Waters and J Woodward. Minnesota, USA: University of Minnesota Press.

Ofiţeru ID, Lunn M, Curtis TP, Wells GF, Criddle CS, Francis CA, Sloan WT (2010). *Combined niche and neutral effects in a microbial wastewater treatment community*. *Proceedings of the National Academy of Sciences* **107**, 15345–15350.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Peter Solymos M, Stevens HH, Szoecs E, Wagner H (2019). *vegan: Community ecology package (Version 2.5-6)*. URL: https://CRAN.xn--r-project-6ma.org/package=vegan.

Plummer M, Best N, Cowles K, Vines K (2006). *CODA: Convergence Diagnosis and Output Analysis for MCMC*. *R News* **6**, 7–11. URL: https://journal.r-project.org/archive/.

Ricklefs RE, Lovette IJ (1999). *The roles of island area* per se *and habitat diversity in the species–area relationships of four Lesser Antillean faunal groups*. *Journal of Animal Ecology* **68**, 1142–1160.

Rosindell J, Hubbell SP, Etienne RS (2011). *The unified neutral theory of biodiversity and biogeography at age ten*. *Trends in ecology & evolution* **26**, 340–348.

Rosindell J, Hubbell SP, He F, Harmon LJ, Etienne RS (2012). *The case for ecological neutral theory*. *Trends in Ecology & Evolution* **27**, 203–208.

Rosindell J, Wong Y, Etienne RS (2008). *A coalescence approach to spatial neutral ecology*. *Ecological Informatics* **3**, 259–271.

Sheldon FH, Styring AR (2011). *Bird diversity differs between industrial tree plantations on Borneo: implications for conservation planning*. *Raffles Bulletin of Zoology* **59**, 295–309.

Sieber H (1926). *List of Birds from Durian*. *Treubia* **8**, 306–310.

Sin YCK, Kristensen NP, Gwee CY, Chisholm RA, Rheindt FE (2021). *Bird diversity on shelf islands does not benefit from recent land-bridge connections*. *Journal of Biogeography*, In press.

Sodhi NS, Wilcove DS, Lee TM, Sekercioglu CH, Subaraj R, Bernard H, Yong DL, Lim SL, Prawiradilaga DM, Brook BW (2010). *Deforestation and avian extinction on tropical landbridge islands*. *Conservation Biology* **24**, 1290–1298.

Stone L, Roberts A (1990). *The checkerboard score and species distributions*. *Oecologia* **85**, 74–79.

Strona G, Nappo D, Boccacci F, Fattorini S, San-Miguel-Ayanz J (2014). *A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals*. *Nature communications* **5**, 1–9.

Strona G, Ulrich W, Gotelli NJ (2018). *Bi-dimensional null model analysis of presence-absence binary matrices*. *Ecology* **99**, 103–115.

Sutari S (2004). *Birding in Lingga Island, Indonesia*. *Singapore Avifauna* **17**, 27–34.

Tan X, Yang X, Chen C, Wang Y (2021). *Nestedness and underlying processes of bird assemblages in Nanjing urban parks*. *Current Zoology* **67**, 383–392.

Triantis K, Mylonas M, Lika K, Vardinoyannis K (2003). *A model for the species–area–habitat relationship*. *Journal of Biogeography* **30**, 19–27.

Triantis K, Vardinoyannis K, Tsolaki E, Botsaris I, Lika K, Mylonas M (2006). *Re-approaching the small island effect*. *Journal of Biogeography* **33**, 914–923.

Ulrich W (2004). *Species co-occurrences and neutral models: reassessing JM Diamond's assembly rules*. *Oikos* **107**, 603–609.

Ulrich W, Gotelli NJ (2007). *Null model analysis of species nestedness patterns*. *Ecology* **88**, 1824–1831.

Ulrich W, Jabot F, Gotelli NJ (2017). *Competitive interactions change the pattern of species co-occurrences under neutral dispersal*. *Oikos* **126**, 91–100.

Ulrich W, Kubota Y, Kusumoto B, Baselga A, Tuomisto H, Gotelli NJ (2018). *Species richness correlates of raw and standardized co-occurrence metrics*. *Global Ecology and Biogeography* **27**, 395–399.

Ulrich W, Zalewski M (2006). *Abundance and co-occurrence patterns of core and satellite species of ground beetles on small lake islands*. *Oikos* **114**, 338–348.

Ulrich W, Zalewski M (2007). *Are ground beetles neutral? Basic and Applied Ecology* **8**, 411–420.

Volkov I, Banavar JR, Hubbell SP, Maritan A (2003). *Neutral theory and relative species abundance in ecology*. *Nature* **424**, 1035–1037.

Wang Y, Bao Y, Yu M, Xu G, Ding P (2010). *Nestedness for different reasons: the distributions of birds, lizards and small mammals on islands of an inundated lake*. *Diversity and Distributions* **16**, 862–873.

Weigelt P, Kreft H (2013). *Quantifying island isolation–insights from global patterns of insular plant species richness*. *Ecography* **36**, 417–429.

Wootton JT (2005). *Field parameterization and experimental test of the neutral theory of biodiversity*. *Nature* **433**, 309–312.

Worthen WB, Jones MT, Jetton RM (1998). *Community structure and environmental stress: desiccation promotes nestedness in mycophagous fly communities*. *Oikos*, 45–54.

Wright DH, Patterson BD, Mikkelson GM, Cutler A, Atmar W (1998). *A comparative analysis of nested subset patterns of species composition*. *Oecologia* **113**, 1–20.

Zalewskia M, Ulrich W (2009). *Island colonization of carabidae: neutrality or habitat choice? Israel Journal of Ecology and Evolution* **55**, 253–262.

Zhang M (2020). *The use and limitations of null-model-based hypothesis testing*. *Biology & Philosophy* **35**, 1–22.

Supplementary Information

A novel sequential sampling algorithm for the application of mechanistic models to species occurrence patterns

N. P. Kristensen, Y. C. K. Sin, H. S. Lim, F. E. Rheindt, R. A. Chisholm

Contents

A Sequential sampling scheme algorithm for the niche–neutral model ii
B Presence-absence matrices iii
C Raw nestedness decreases with increasing niche diversity iv
D Details of the tuned niche–neutral model v
E Overview of the code repository vii

Appendix A. Sequential sampling scheme algorithm for the niche–neutral model

Algorithm 1 provides an overview of the sequential sampling scheme. Tutorials and the implementation are available in the Github repository (`nadiahpk/niche-neutral-riau-birds/`):

(1) `./tutorial/Sequential_Sampling.pdf`, `./tutorial/Sequential_Sampling.ipynb`:
Tutorials explaining how the algorithm works using a concrete example.

(2) `draw_sample_species_generator_general()` in `./functions/my_functions.py`:
A function implementing the algorithm, which was used to generate results for this study.

Algorithm 1 Sequential sampling scheme for the niche–neutral model

```
Input: K, H                                  ▷ number of niches (n* in main text); number of islands
Input: θ = [θ_1, ... , θ_K]                  ▷ vector of fundamental biodiversity number for each niche
Input: m = [m_1, ... , m_H]                  ▷ vector of immigration rates to each island
Input: J = [J_{k,h}]                         ▷ matrix of no. individuals in niche k on island h (proportional to area)

a = [ ]                                      ▷ Initialise empty immigrant ancestors list and local community matrix, where
n = [ ]                                      ▷ index is species identity and value is abundance

for k = 1 ... K                              ▷ For each niche
    for h = 1 ... H                          ▷ For each island

        I = m_h(J_{k,h} − 1)/(1 − m_h)       ▷ Etienne's immigration parameter

        for j = 0 ... J_{k,h} − 1            ▷ For each individual

            Draw α_1 ~ U(0, 1)
            if α_1 ≤ I/(I + j)               ▷ Draw an immigrant individual

                Draw α_2 ~ U(0, 1)
                if α_2 ≤ θ_k/(θ_k + Σ_l a_{k,l})   ▷ Immigrant individual is a new species

                    a_k ← [a_k, 1]
                    n_{k,h} ← [n_{k,h}, 1]   ▷ new species added by appending 1 to lists

                else                         ▷ Immigrant individual is previously sampled species

                    Choose species i* with probability P(i = i*) = a_{k,i*} / Σ_l a_{k,l}
                    a_{k,i*} ← a_{k,i*} + 1
                    n_{k,h,i*} ← n_{k,h,i*} + 1   ▷ species i* abundance increased by 1

                end if

            else                             ▷ Draw a local individual

                Choose species i* with probability P(i = i*) = n_{k,h,i*} / Σ_l n_{k,h,l}
                n_{k,h,i*} ← n_{k,h,i*} + 1  ▷ species i* abundance increased by 1

            end if
        end for
    end for
end for
```

Appendix B. Presence-absence matrices

The presence-absence matrices provide an intuitively visual way to compare the observed species distribution to the various model predictions (Fig. B.1).

Comparing the observed matrix (Fig. B.1a) to examples of predictions from the homogeneous niche–neutral model (Fig. B.1b, c), it is apparent that the model predicts lower total richness and higher nestedness than observed.

When both niche diversity and immigration rates were varied across islands (S 3.2.3), matrices were visually similar to the observed matrix (Fig. B.1d-e).

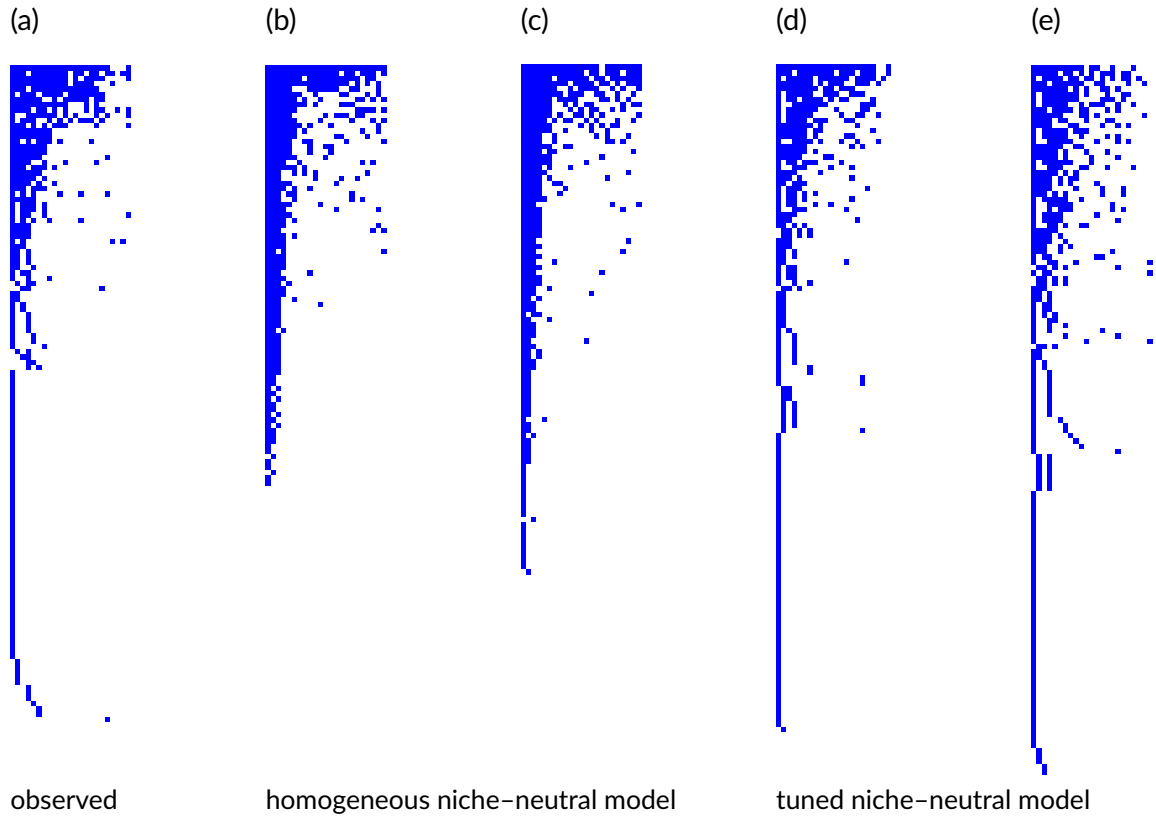

Figure B.1 – Comparing presence-absence matrices from (a) observations, (b-c) best-fitting homogeneous niche–neutral model, (d-e) scenario contrived to match segregation and nestedness. Rows are species, columns are islands, and a coloured square represents the presence of a species on an island.

Appendix C. Raw nestedness decreases with increasing niche diversity

To further explore the effect of niche diversity on species distribution patterns, the niche–neutral model was refitted to survey data at varying fixed n^* levels. Then, the fitted parameter values were used to create idealised archipelagos of 16 islands each (Fig. C.2a).

In general, increasing n^* decreased the raw nestedness score ($n^* = 1$ mean raw NODF 88.6, 95% PI [85.1, 91.5]; $n^* = 7$ mean raw NODF 83.8, 95% PI [78.6, 86.6]; Fig. C.2b) but increased the nestedness compared to the fixed–fixed randomisation ($n^* = 1$ mean SES NODF -1.49, 95% PI [-5.29, 0.91]; $n^* = 7$ mean SES NODF -1.08, 95% PI [-4.18, 0.994]).

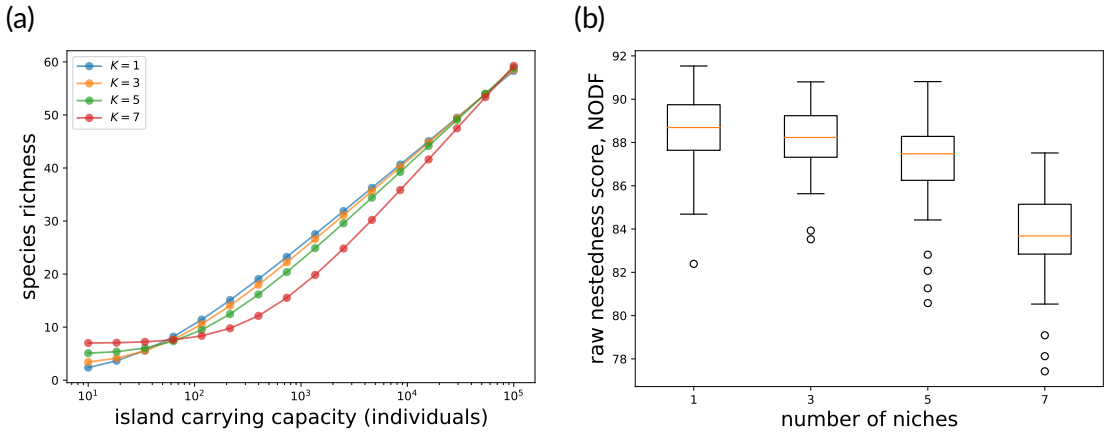

Figure C.2 – The effect of increasing the number of niches on every island on nestedness: (a) species richness versus carrying-capacity relationship; and (b) the effect of four n^* scenarios on raw nestedness.

Appendix D. Details of the tuned niche–neutral model

In this appendix, we detail the example scenario presented in main text. The scenario was generated with the script: `scripts/neutral_data_fitmK/fitmK_3.py`. Additional experiments can be found in the electronic repository, in subdirectories named `*data_fit*`.

The observed data has a lower raw NODF and higher SES C-score than is produced by the best-fitting homogeneous niche–neutral model (Fig. 3). However, the homogeneity assumption, i.e., that all islands have the same n^* and m, is unrealistic. In real archipelagos, habitat diversity typically increases with island size (Davidar et al., 2001; Kohn and Walsh, 1994; Ricklefs and Lovette, 1999), and the immigration rate varies in complicated ways depending on both the distance to the mainland and nearby islands as well as the species characteristics.

From the simulation experiments with hypothetical archipelagos (Sect. 3.2), we also learnt that (1) imposing a positive relationship between island size and number of niches n^* increases SES C-score, and (2) tuning the island-specific immigration rates m so the model correctly predicts each island's richness decreases raw nestedness. Therefore, we expected that combining these assumptions into a fitted niche–neutral model would bring the model predictions into closer agreement with observations.

The scenario presented in the main text was obtained through a combination of applying the 2 generalisations and a process of trial and error. The total number of niches in the archipelago was increased from $K = 8$ (best-fit value in the homogeneous model) to 11. The extra niches represent niches that are available on some islands but not others, which increases species segregation. Any (small) island with fewer than 8 species was set with an island-specific K_h equal to the number of species. For islands with more than 8 species, the island-specific immigration parameter m_h was fitted so that the expected number of species on the island was equal to the observed richness. For some large islands, fitting m_h in this way led to unrealistically high immigration rates, several orders of magnitude greater than best-fitting rate from the homogeneous model ($m = 0.0036$). Therefore, those islands were instead given a higher island-specific niche diversity ($K_h > 8$), which is consistent with the observation that large islands can sustain key habitats like rainforests (Davidar et al., 2001), and which allowed a more realistic m_h to be fitted.

Table D.1 summarises the example scenario presented in the main text. Niches numbered 0 to 10 were distributed somewhat arbitrarily across islands, but low-numbered niches represent habitats that occur on most islands, and high-number niches (e.g., niches 9 and 10) represent habitats that typically occur only on larger islands (e.g., intact rainforest). Fig D.3 shows how the fitted m varied with island size.

Table D.1 – Summary of the tuned model presented in the main text.

island name	area (km^2)	Niches present 0	1	2	3	4	5	6	7	8	9	10	fitted m
Lingga	431	■	■	■	■		■	■	■		■		0.02742
Sebangka	123					■	■	■	■	■		■	0.00047
Temiang	30.0		■	■	■	■	■	■	■	■			0.00069
Durian	8.74	■	■	■	■	■		■	■			■	0.00604
Ngenang	1.39	■	■	■	■	■	■	■	■	■			0.00849
Momoi	1.28	■	■	■	■	■	■	■	■				0.00372
Tinjul	0.104	■	■	■	■	■	■	■	■	■	■		0.07626
Akar	0.100	■	■	■	■	■	■	■	■				0.02312
Remang Besar	0.0855	■	■	■	■	■	■	■	■				0.02178
Sambu Kecil	0.0575	■	■	■	■	■	■	■	■				0.01965
Remang Kecil	0.0475	■	■	■	■	■	■	■	■				0.02616
Matang Cilik	0.0425		■	■	■			■	■				0.00361
Panggal Besar	0.0409	■	■	■	■	■	■	■	■				0.00361
Malang	0.0320	■	■	■	■	■	■	■	■				0.01369
Paloi Besar	0.0206	■	■	■	■	■	■	■	■				0.02968
Sepatu	0.0101	■	■	■	■	■	■	■	■				0.00361
Panggal Sedang	0.00972	■	■	■	■	■	■	■	■	■	■		0.27939
Nipong	0.00969	■	■	■	■		■	■	■				0.00361
Sayap	0.00848	■	■		■	■	■	■					0.00361
Telor Besar	0.00475	■		■					■				0.00361
Paloi Kecil	0.00390				■								0.00361
Telor Kecil	0.00260						■	■					0.00361
Panggal Kecil	0.000724	■	■		■	■	■	■	■				0.00361

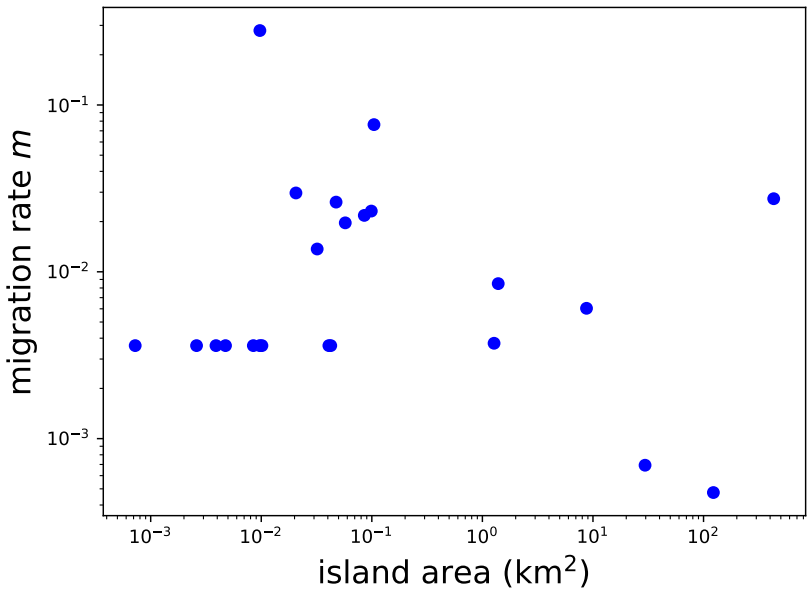

Figure D.3 – Fitted m versus island area.

Appendix E. Overview of the code repository

Functions and scripts used to generate the results are archived in the Github repository: `https://github.com/nadiahpk/niche-neutral-riau-birds` (Fig. E.4). C-score calculations are in scripts in directories `/scripts/cooccurrence*/`, and NODF calculations are in scripts in directories `/scripts/nestedness*/`.

Future workers who wish to use the niche–neutral model and sequential sampling algorithm may find the following resources useful:

(1) Tutorial for the sequential sampling algorithm: `/tutorial/`.

(2) Function implementing the algorithm is in `/functions/my_functions.py`: `draw_sample_species_generator_general()` and `draw_sample_species_generator()`.

(3) Function to fit the niche–neutral parameters to data is in: `scripts/area_richness_curve/fit_all_island_subsets.py`. It uses the function `fit_area_richness()` in `/functions/my_functions.py`.

```
repository
├── data
├── functions
│   ├── NODF_functions.R (calculates Almeida-Neto et al. (2008) NODF, includes example)
│   ├── helpers.py (string processing)
│   └── my_functions.py (sequential sampling algorithm and niche–neutral fitting)
├── scripts
│   ├── area_richness_curve (fit niche–neutral model)
│   ├── cooccurrence_data (data-randomised co-occurrence analysis, Table 2)
│   ├── cooccurrence_neutral_data (co-occurrence analysis of homogeneous niche–neutral archipelagos, Fig. 3)
│   ├── cooccurrence_neutral_data_fitm (non-homogenous, vary immigration rate, S 3.2.2, Fig. 5)
│   ├── cooccurrence_neutral_data_fitmK (non-homogenous, vary both immigration and niche number, Fig. 6)
│   ├── cooccurrence_neutral_jump_migration (verify Scenario 1 segregation cause, S 3.2.1)
│   ├── cooccurrence_neutral_vary_JK (Scenario 2 S 3.2.1,Fig. 4)
│   ├── cooccurrence_neutral_vary_K (baseline and Scenario 1, S 3.2.1, Fig. 4)
│   ├── nestedness_rowpack_data (data-randomised nestedness analysis, Table 2)
│   ├── nestedness_rowpack_neutral_data (nestedness analysis of homog. niche–neutral archipelagos, Fig. 3)
│   ├── nestedness_rowpack_neutral_data_fit* (non-homogenous, corresponding to scenarios above)
│   ├── nestedness_rowpack_neutral_vary_* (Scenarios in S 3.2.1)
│   ├── nestedness_rowpack_neutral_survey_vary_K (further expore effect of niche diversity, SI C)
│   ├── neutral_data (synthetic archipelagos, results Fig. 2)
│   ├── neutral_data_fit* (synthetic archipelagos for non-homogenous scenarios)
│   ├── neutral_* (other synthetic archipelagos with names corresponding to those above)
│   ├── process_data (convert data/raw/ to data/processed/)
│   ├── summary_plots_stats (manuscript figures)
│   └── verify (various checks of the algorithms)
├── results (corresponds to directory structure of scripts)
└── tutorial (tutorial for sequential sampling algorithm)
```

Figure E.4 – Overview of the contents of the Github repository.