ECOGRAPHY

Research

Occupancy-based diversity profiles: capturing biodiversity complexities while accounting for imperfect detection.

Jesse F. Abrams, Rahel Sollmann, Simon L. Mitchell, Matthew J. Struebig and Andreas Wilting

J. F. Abrams (https://orcid.org/0000-0003-0411-8519) ✉ (j.abrams@exeter.ac.uk) and A. Wilting (https://orcid.org/0000-0001-5073-9186), Leibniz Inst. for Zoo and Wildlife Research, Berlin, Germany. JFA also at: Global Systems Inst., College of Life and Environmental Sciences, Univ. of Exeter, Exeter, UK. – R. Sollmann (https://orcid.org/0000-0002-1607-2039), Dept of Wildlife, Fish and Conservation Biology, Univ. of California Davis, Davis, CA, USA. – S. L. Mitchell (https://orcid.org/0000-0001-8826-4868) and M. J. Struebig (https://orcid.org/0000-0003-2058-8502), Durrell Inst. of Conservation and Ecology (DICE), School of Anthropology and Conservation, Univ. of Kent, Canterbury, Kent, UK.

Ecography
44: 1–12, 2021
doi: 10.1111/ecog.05577

Subject Editor: Tamara Munkemuller
Editor-in-Chief: Miguel Araújo
Accepted 5 March 2021

Measuring the multidimensional diversity properties of a community is of great importance for ecologists, conservationists and stakeholders. Diversity profiles, a plotted series of Hill numbers, simultaneously capture the common diversity indices. However, diversity metrics require information on species abundance, often relying on raw count data without accounting for imperfect and varying detection. Hierarchical occupancy models account for variation in detectability, and Hill numbers have been expanded to allow estimation based on occupancy probability. But the ability of occupancy-based diversity profiles to reproduce patterns in abundance-based diversity has not been investigated. Here, we fit community occupancy models to simulated animal communities to explore how well occupancy-based diversity profiles reflect patterns in true abundance-based diversity. Because we expect occupancy-based diversity to be overestimated, we further tested a occupancy thresholding approach to reduce potential biases in the estimated diversity profiles. Finally, we use empirical bird community data to present how the framework can be extended to consider species similarity. The simulation study showed that occupancy-based diversity profiles produced among-community patterns in diversity similar to true abundance diversity profiles, although within-community diversity was generally overestimated. Applying an occupancy threshold reduced positive bias, but resulted in negative bias in richness estimates and slightly reduced the ability to reproduce true differences among the simulated communities; thus, we do not recommend application of this threshold. Application of our approach to a large bird dataset indicated differential species diversity patterns in communities of different habitat types. Accounting for phylogenetic and ecological similarities between species reduced variability in diversity among habitats. Our framework allows investigating the complexity of diversity from species detection data, while accounting for imperfect and varying detection probabilities, as well as species similarities. Visualizing results in the form of diversity profiles facilitates comparison of diversity between sites or across time. The approach offers opportunities for further development, for example by using local abundances estimated using the Royle–Nichols or N-mixture models and further exploration of thresholding methods. In spite of some challenges, occupancy-based diversity profiles are useful for studying and monitoring patterns in biodiversity.

NORDIC SOCIETY OIKOS

www.ecography.org

Keywords: biodiversity, diversity index, diversity profile, occupancy, presence, species distribution modeling, specificity, threshold

Introduction

Biological diversity represents the variety of organisms or traits and plays a central role in ecological theory (Loreau et al. 2001, Tilman et al. 2014). Mathematical functions known as diversity indices aim to summarize properties of communities that allow comparison among different regions, taxa and trophic levels (Morris et al. 2014, Daly et al. 2018). They are often used in conservation as indicators of the integrity or stability of ecosystems, and are, therefore, of fundamental importance for environmental monitoring and conservation (Morris et al. 2014). Diversity is, however, a generic term describing the complex multidimensional properties of a community. Any diversity index reduces these multidimensional properties to a single number (Morris et al. 2014), which is problematic (Daly et al. 2018).

The most commonly used diversity indices are species richness, Shannon's diversity H′ and Simpson's diversity D; the latter two combine measures of richness and abundance, whereas species richness solely presents the number of species. It is not uncommon that diversity increases according to one index, but decreases according to another (Patil 2014), demonstrating the difficulties in quantifying biodiversity in a single number (Purvis and Hector 2000, Daly et al. 2018).

To address this shortcoming, several researchers have suggested using parametric families of diversity indices (Hill 1973, Patil and Taillie 1982, Gattone and Battista 2009, Leinster and Cobbold 2012). Jost (2006) proposed the use of Hill numbers (Hill 1973) that incorporate relative abundance and species richness to show the number of equally abundant species necessary to produce the observed value of diversity. Individual Hill numbers differ by the parameter q, which quantifies how much the measure discounts rare species when calculating diversity (the higher q, the less these rare species contribute to diversity). Leinster and Cobbold (2012) further developed this framework to incorporate similarity between species and present them along a gradient of q that includes Rao's quadratic entropy and species richness. The naive similarity collapses diversity to the Hill number of order q (Hill 1973). Hill numbers include typical diversity indices, where $q = 0$ reflects species richness, $q = 1$ is the exponential of the Shannon entropy and $q = 2$ represents the inverse of Simpson's concentration. Plotting the effective number of species as a function of q allows us to view diversity from multiple vantage points (Hill 1973, Leinster and Cobbold 2012) and the resulting curves have become known as diversity profiles. These curves display different properties of diversity and often drop sharply between $q = 0$ and $q = 1$ and level off soon after $q = 2$, indicating that many communities are dominated by a few highly abundant species (Preston 1948).

To calculate diversity indices (except for species richness), as well as diversity profiles, information about the relative abundances of species, or evenness, is required (Leinster and Cobbold 2012). Obtaining information on species abundance can, however, be challenging. Raw count data are typically fraught with detection bias (Nichols et al. 1998, MacKenzie and Kendall 2002, Sollmann et al. 2013), and inference about diversity from indices based on count-based relative abundance estimates that do not account for imperfect and varying detectability may therefore be biased. Estimating abundance of all species in a community while accounting for varying detection, for example using capture–recapture methods (Royle and Dorazio 2008), is extremely difficult as different organisms require different sampling methods to obtain sufficient data for reliable abundance estimation. Thus, community studies often resort to the collection of much cheaper and easier to obtain species detection/non-detection data. Even with these incidence data we must consider that species may be detected imperfectly (MacKenzie et al. 2002, 2006).

Occupancy modelling provides a framework to handle the problem of imperfect and varying detection, producing unbiased estimates of species occurrence (MacKenzie et al. 2002, 2006). The development of hierarchical multi-species occupancy models (Dorazio and Royle 2005, Dorazio et al. 2006) has enabled estimation of richness at the level of the study area and survey location (Sollmann et al. 2017), and to model variation in richness across areas as a function of covariates (Sutherland et al. 2016). However, it has been shown that multi-species occupancy models overestimate true species richness (Zipkin et al. 2012), and only a few applications for other diversity indices exist (Broms et al. 2016, Guillera-Arroita et al. 2019). Consequently, accounting for imperfect detection has often been neglected in calculating diversity metrics in the past.

Only recently, Chao et al. (2014) described a method to calculate Hill numbers from incidence data, and Broms et al. (2015) followed with a formulation to calculate detection-corrected occupancy-based Hill numbers. Here, we extend this framework to facilitate the calculation, visualization, and thus, interpretation of occupancy-based diversity profiles. We first explored how occupancy-based diversity profiles compare to true abundance-based diversity profiles using simulated data. Because of the asymptotic relationship between occupancy and abundance (i.e. occupancy approaches 1 as abundance increases, thus decreasing differences between species), we expect occupancy-based profiles to overestimate diversity for $q > 0$ (i.e. suggest a more even community). Moreover, based on previous work (Zipkin et al. 2012, Broms et al. 2016, Guillera-Arroita et al. 2019) we expect occupancy-based profiles to overestimate richness, due to characteristics of the underlying community occupancy models (see 'Occupancy threshold' section for details) and that likely inflates diversity at $q > 0$ as well. But because this affects all communities, we expect occupancy-based profiles to still be able to correctly order areas by their diversity. Therefore, we tested their ability to compare communities across landscapes

with varying levels of habitat disturbance, associated with varying levels of diversity. We then used an empirical dataset of diverse bird communities collected in Sabah, Malaysian Borneo, to demonstrate how the framework can be extended to a trait-based diversity analysis of occupancy-data by incorporating measures of similarity as proposed by Leinster and Cobbold (2012).

Methods

Our study was divided into a simulation study and an application of occupancy-based diversity profiles to an empirical dataset. The simulation study followed six steps: 1) simulation of a forest degradation gradient, 2) simulation of animal communities, and detection data of those communities, 3) community occupancy analysis of simulated detection data, 4) utilization of community occupancy model output to construct occupancy-based diversity profiles, 5) application of thresholding to occupancy-based profiles and 6) evaluation of the performance of the occupancy-based diversity profiles by comparing them to true community abundance diversity profiles. In the second part of our study, we applied the community occupancy diversity profiles to an empirical bird dataset from Malaysian Borneo and demonstrate how these profiles can account for phylogenetic or ecological trait similarities.

Forest degradation and community simulation

We simulated five virtual 10 × 10 km forest landscapes, with 200 × 200 m grid cells (50 × 50 cells). We simulated a habitat covariate representing 'habitat disturbance' (where 0 represents undisturbed forest and 5 represents complete deforestation) for each landscape (Fig. 1A) by drawing random samples from a multivariate normal distribution. To increase realism by simulating nonrandom habitat, we explicitly included spatial autocorrelation in the simulation of the habitat covariate by using the distance matrix as our variance-covariance matrix and a decay function with a decay constant

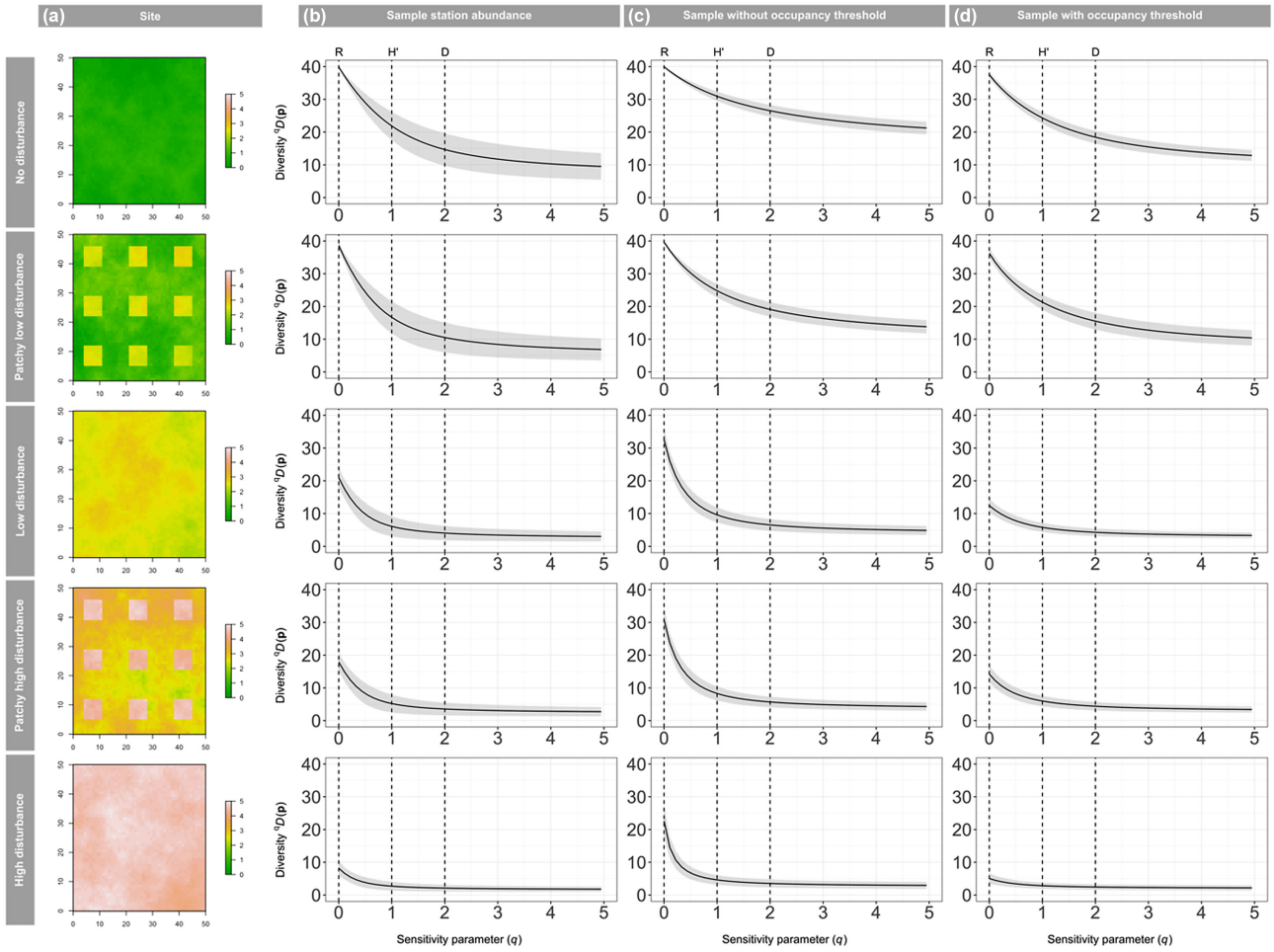

Figure 1. Results from the simulation study. (A) Simulated 'forest quality' habitat covariate for five virtual landscapes. Average diversity profiles (solid line) for simulated data generated using data from only the 100 surveyed stations for (B) the simulated true abundance, and the community occupancy predictions for the entire landscape (C) without thresholding and (D) with thresholding using the max_{SSS} method for the entire study landscape, showing the standard deviation (grey shading).

ϕ to specify how the relationship to other cells changes with distance (Supporting information). The five landscapes constitute a habitat degradation gradient representative of three different logging regimes and two 'patchy' landscapes that simulate activities such as compartmental logging: 1) no disturbance, 2) patchy low disturbance (i.e. low impact logging restricted to a few logging compartments), 3) low disturbance across the entire area (i.e. low impact logging conducted throughout the logging concession), 4) patchy high disturbance (i.e. conventionally logging restricted to a few logging compartments) and 5) high disturbance across the entire area (i.e. conventional logging conducted throughout the logging concession).

We then simulated the cell-level abundance of 40 virtual species distributed within our five landscapes. We simulated abundance to generate true abundance-based diversity profiles. Abundance was simulated for each of the 40 species, *i*, at grid cell *j* ($j = 1, 2, \ldots n_T$) in each of our five landscapes under the following model:

$$N_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right) \qquad \text{True abundance}$$

$$\log\left(\lambda_{ij}\right) = \beta 0_i + \beta 1_i \times \text{habitat}_j \qquad \text{Expected abundance}$$

where the species-specific intercepts ($\beta 0$) and dependence on the habitat covariate ($\beta 1$) are simulated as normally distributed with community hyperparameters (Supporting information):

$$\beta 0_i \sim \text{Normal}\left(\mu_0, \sigma_0^2\right)$$

$$\beta 1_i \sim \text{Normal}\left(\mu_1, \sigma_1^2\right)$$

We set the average response of species to habitat disturbance as negative ($\mu_1 = -2$) since most forest-adapted species will respond to logging negatively (Supporting information). We allowed this to vary, generating a community of species with mostly negative responses, but with few species that responded positively to habitat disturbance. Average expected abundance per species per grid cell in undisturbed forests was 0.37. This resulted in five communities with species richness ranging from 17.2 ($\pm$ 2.8) for the most disturbed site to 40 ($\pm$ 0) for the undisturbed site.

We then generated detection/non-detection data by simulating systematic repeated sampling of the community. For each landscape, we picked 100 sampling points in a grid spaced 2 km apart. At each point, we then simulated the observation process with imperfect detection:

$$y_{ijk} \sim \text{Binomial}\left(N_{ij}, r_{ijk}\right) \qquad \text{Observed count data}$$

$$\text{logit}\left(r_{ijk}\right) = \alpha 0_i \qquad \text{Detection probability}$$

where y_{ijk} is the observed count data and r_{ijk} is the probability of detection of an individual present at site *j*. We repeated the observation process for 10 occasions, *k*. The average detection probability for the community for all landscapes was set to 0.5. Species-specific intercepts for detection on the logit scale, $\alpha 0$, were drawn from a normal distribution describing the community ($\mu_\alpha = 0$, $\sigma_\alpha = 1$).

Finally, we reduced the observed count data *y* to detection/non-detection data. We chose to represent the detection process in two steps (generating counts, then reducing these to binary detection/non-detection data) to reflect how data from typical non-invasive survey methods, such as bird point counts or camera-trapping, are prepared for occupancy modeling. We compared the performance of the occupancy-based diversity profiles to true community abundance diversity profiles. We repeated the simulation process to generate 100 communities. All calculations were carried out in R ver. 3.6.0 (<www.r-project.org>).

Community occupancy model

We adopted the hierarchical formulation of occupancy models by Royle and Dorazio (2008) extended to a community occupancy model (Dorazio and Royle 2005, Dorazio et al. 2006). To analyze our simulated data, we combined data from all five landscapes and, following common practice in analyzing field data, modelled occupancy probability as having species-specific random intercepts, $\beta 0_{is}$, with landscape specific (indicated by *s* indexing) hyperparameters ($\mu_{\beta 0,s}$, $\sigma_{\beta 0,s}$), to allow for different baseline occupancy in the different landscapes and among species. We further modelled species-specific effects on occupancy of the simulated habitat 'disturbance' covariate ($\beta 1_i$). Detection probability included a species-specific random intercept with landscape specific hyperparameters, to allow for differences in baseline detection among landscapes. In our case, the different landscapes had different abundances of animals, which leads to differences in species-level detection (Royle and Nichols 2003) and likely also to differences in species-level baseline occupancy probability. When fewer than the 40 simulated species were observed across all landscapes, we augmented the observed data set with 40 minus the number of observed species with all-0 detection data (note that this implies more data augmentation, and thus a higher potential for positive bias in diversity estimates, in more disturbed landscapes with fewer species, compared to less disturbed landscapes where more species were present). The formal model description can be found in the Supporting information. We implemented the model in a Bayesian framework using JAGS (Plummer 2003) accessed via the R packages *rjags* (Plummer 2019). We ran three parallel Markov chains with 250 000 iterations, of which we discarded 50 000 as burn-in, and we thinned the remaining iterations by 20 to make the output more manageable. We assessed chain convergence using the Gelman–Rubin statistic (Gelman et al. 2004). Values under 1.1 indicate convergence,

and all parameters in our models had a Gelman–Rubin statistic < 1.1. We tested whether the model adequately fit the data by calculating a Bayesian p-value (Gelman et al. 1996). We observed no lack of fit for any of our models.

Diversity profiles

In order to investigate the performance of estimating diversity profiles with occupancy-based information, we first constructed diversity profiles for the simulated abundance across the sampling stations. Diversity profile values ($^qD^Z$) for abundance data can be calculated according to Leinster and Cobbold (2012) as:

$$^qD^Z(\lambda) = \left(\sum_{i=1}^{M} \frac{\lambda_i}{\Sigma(\lambda)} \left(Z_\lambda\right)_i^{q-1} \right)^{\frac{1}{1-q}} \quad (1)$$

where

$$\left(Z_\lambda\right)_i = \sum_{h=1}^{M} Z_{ih} \left(\frac{\lambda_h}{\Sigma(\lambda)} \right) \quad (2)$$

where λ represents abundance, M is the number of species in the assemblage and the ith species has relative abundance $\frac{\lambda_i}{\Sigma(\lambda)}$. The parameter q determines the sensitivity of the measure to the relative abundances of species. This allows us to calculate diversity along a continuum of values of q. At $q=0$, $^qD^Z$ equals species richness where all species are considered equally. As q becomes larger, more weight is placed on common species thereby incorporating evenness into the diversity measure and resulting in a lower value of $^qD^Z$ for more uneven assemblages than for more even assemblages. The diversity profile framework from Leinster and Cobbold (2012) allows for the consideration of similarity between species through the inclusion of an $M \times M$ similarity matrix $\mathbf{Z}$ which represents the similarity between the ith and hth species. Values of 0 in $\mathbf{Z}$ indicate total dissimilarity, whereas values of 1 indicate identical species. This matrix can be used to adjust the profiles by incorporating any measure of similarity (such as phylogenetic or trait) between different species or taxonomic groups. In our simulation study, we use a naive similarity matrix (an identity matrix with all cells on the diagonal equal to 1 and all other values = 0). In the empirical dataset (below) we adjusted the profiles using a diet, taxonomic and a phylogenetic similarity matrix. We refer to $q=0$ as 'richness' (R), which, depending on the nature of Z_{ih}, can represent species richness, or trait richness.

To use occupancy probabilities instead of abundances to construct diversity profiles we altered the diversity profile method of Leinster and Cobbold (2012) as follows:

$$^qD^Z(\bar{\psi}) = \left(\sum_{i=1}^{M} \frac{\bar{\psi}_i}{\Sigma(\bar{\psi})} \left(Z_{\bar{\psi}}\right)_i^{q-1} \right)^{\frac{1}{1-q}} \quad (3)$$

where

$$\left(Z_{\bar{\psi}}\right)_i = \sum_{h=1}^{M} Z_{ih} \left(\frac{\bar{\psi}_h}{\Sigma(\bar{\psi})} \right) \quad (4)$$

where M is again the number of species in the assemblage, $\bar{\psi}_i$ is the average occupancy probability of species i, the ith species has relative occupancy $\frac{\bar{\psi}_i}{\Sigma(\bar{\psi})}$ and $\mathbf{Z}$ is the similarity matrix. This adjustment is similar to that developed by Broms et al. (2015) but extends their approach by allowing occupancy probability to be modeled as a function of covariates.

We used the parameter estimates from the occupancy model fit to our simulated data to calculate species occupancy for the sample stations in the five simulated landscapes and then constructed diversity profiles using the mean occupancy probability across sampling stations for each species using Eq. 3 and 4 for all posterior samples of the community occupancy model. This effectively creates posterior distributions for the diversity profiles themselves and allowed us to determine their standard deviations (SDs) and 95% Bayesian credible intervals. We then compared estimates of R, H′ and D from the occupancy-based profiles against indices based on the true abundance profiles by evaluating the diversity profiles at $q=0$, $q=1$ and $q=2$, respectively. Specifically, for each landscape, we present the average relative bias (occupancy-based index minus true abundance index divided by true index) across all 100 communities and coverage, i.e. the proportion of communities for which the 95% CI of the occupancy-based index estimate included the true-abundance based index. We also generated occupancy-based diversity profiles for landscape wide predictions of occupancy generated using the parameters estimated in the community occupancy model. We did this to compare the results between the sampling station-based profiles and the landscape-wide profiles.

To evaluate how well occupancy-based diversity profiles were able to order landscapes by site diversity rank, we compared them to the diversity ranking in the true abundance-based profiles. Since not all communities were different in the true abundance-based profiles (e.g. no disturbance and patchy low disturbance landscapes both had an average species richness of 40 (Supporting information) and could, therefore, not be distinguished in the true-abundance based profiles), we first determined how many sites we could reliably distinguish. To do so we used the results of the true abundance-based simulations and checked which landscapes could be distinguished in 95% of the 100 simulations for R, H′ and D. We were able to distinguish between 3, 4 and 3 of the 5 sites in the true abundance profiles for R, H′ and D, respectively (Supporting information). We then calculated the proportion of communities for which occupancy-based diversity profiles resulted in the same rank order of these distinguishable landscapes as true abundance-based profiles.

Occupancy threshold

Community occupancy models have been shown to overestimate diversity through several processes: when data

Table 1. Percentage of the 100 simulated communities within each of five landscapes where occupancy-based diversity profiles were ordered the same as the true abundance-based diversity profiles. The maximum number of landscapes that could be reliably separated in the true abundance profiles was 3, 4 and 3 of the 5 landscapes for R, D and H′, respectively (Supporting information).

Index	Landscape-wide occupancy	Landscape-wide occupancy with threshold	Sample-station occupancy	Sample-station occupancy with threshold
Species richness (R, $q=0$)	97%	100%	97%	100%
Shannon diversity (H′, $q=1$)	96%	94%	96%	92%
Simpson's index (D, $q=2$)	94%	90%	94%	82%

augmentation is used to estimate richness including species never observed, it induces a non-0 probability of occurrence for augmented species, which can inflate richness (and likely other diversity) estimates. Further, because rare species borrow information from common species (through shared hyperparameters), their occupancy probability may be overestimated (Broms et al. 2016, Guillera-Arroita et al. 2019). Finally, for data sparse species, detection probability may be underestimated, and occurrence probability consequently overestimated, which has been shown to lead to positive bias in occupancy-based Hill numbers (Broms et al. 2015). To explore methods to account for the expected overestimation of diversity profiles in community occupancy models, we tested the use of an occupancy threshold. An occupancy threshold is traditionally used to transform occupancy probabilities into binary outputs. Here, we explore the use of a threshold to determine at which occupancy levels we can consider a species to truly occupy a landscape, thus reducing the overestimation in richness.

When using presence/absence data, the identities of both presence and absence data are (assumed to be) known (Liu et al. 2016). However, with detection/non-detection data we have no information about 'true absences', which presents challenges for threshold selection. Liu et al. (2016) identified the max_{SSS} method, which maximizes the sum of sensitivity and specificity, as the most suitable objective approach for determining thresholds with incidence data (presence-only data).

The max_{SSS} method described by Liu et al. (2016) requires the use of 'pseudo-absences', which are randomly picked from the sampling stations with no detections. Here, we use the estimates from the occupancy model to draw 'pseudo-absences' randomly for stations without detections but weighed by the probability of a station being unoccupied, $1-\psi$.

We calculated the max_{SSS} (Supporting information) threshold for each species for each landscape based on the mean occupancy estimate using the *optimal.thresholds* function from the R package *PresenceAbsence* (Freeman and Moisen 2008). We set occupancy probabilities for stations with estimates below the occupancy threshold to zero. We then averaged the threshold-adjusted occupancy for each species across landscapes for each model iteration and generated new diversity profiles using the adjusted dataset. We compared threshold occupancy-based profiles against true-abundance based profiles as described for non-threshold profiles and evaluate if using a threshold improves our ability to distinguish between the landscapes.

Case study

We sampled bird communities at 307 point-count localities in and around the Stability of altered forest ecosystems (SAFE) project (117°5'N, 4°6'E) in Sabah, Malaysian Borneo (Mitchell et al. 2018). Thirty-eight localities were in continuous logged forest (CF) of the Ulu Segama Forest Reserve, with an additional 156 in the neighboring SAFE landscape, in forest that had been logged several times and recently salvage logged. A further 113 localities were sampled alongside rivers in oil palm plantations, including 88 with riparian forest remnants (RR) on each riverbank and 15 with no natural vegetation (OPR). Localities are classified by habitat into four categories: non-riparian continuous forest (CF), riparian forest (RF), riparian remnant (RR) and oil palm river (OPR). Forest quality, based on aboveground carbon density measured via LiDAR, also varied substantially across the landscape.

Point counts were undertaken by a single experienced observer (SLM) for 15 min on mornings without rain, recording all birds heard or seen within a 50-m radius. Each point was sampled three times, typically within a few weeks, during field work undertaken 2015–2018 (for details, Mitchell et al. 2018).

We observed a total of 169 bird species. Two species, *Leptocoma brasiliana* and *Zanclostomus javanicus*, were excluded because there is no phylogenetic information available, which is necessary for the trait-based analysis. Further, three species of swift (*Aerodramus maximus*, *A. salangana* and *A. fuciphagus*) could not be reliably separated and are considered as *Aerodramus* spp.

To analyze the case study data, we used a similar community occupancy model structure as used for the simulation study. Following Mitchell et al. (2018), we modeled occupancy using above-ground carbon density, forest cover and riparian remnant width as predictors, with species-specific random intercepts with habitat-specific hyperparameters. Covariates were derived using remotely sensed data and calculated following Mitchell et al. (2018). Detection probability included a species-specific random intercept with habitat specific hyperparameters and accounted for the effect of time and date of a survey on the probability of detection (Ellis and Taylor 2018).

We separated species communities according to the four habitat-types described above for diversity profile construction with and without occupancy thresholds. We used the occupancy probabilities for each sampling station to construct the diversity profiles. Additionally, we constructed

similarity matrices according to diet, taxonomy and phylogeny (Supporting information) to demonstrate how similarity can be incorporated into the occupancy-based diversity profile framework.

Results

Simulation results

The number of species present in the simulated true abundance data in each landscape ranged between, on average, 17 and 40, whereas the number of species detected in each landscape ranged between, on average, 8 and 39, indicating that our simulation of the detection process resulted in the detection of between 47 and 99% of the species present in the five landscapes. On average across the five landscaped, six species were missed due to imperfect detection. We present the average and standard deviation (as a measure of variability among simulated communities) for the true abundance and occupancy-based profiles generated for the sample stations in 100 simulated communities in Fig. 1. The occupancy-based diversity profiles (Fig. 1C without thresholding and Fig. 1D with the threshold, for the application of this method to a single site see the Supporting information) showed similar trends in diversity as the diversity profiles based on true abundance (Fig. 1B). Occupancy-based species richness estimates without thresholding corresponded well to true richness for all but the most disturbed landscape (average bias of 38%). As expected, for $q > 0.5$ (including H′ and D), occupancy-based diversity profiles showed consistent positive bias (Fig. 1B, Supporting information). Patterns in coverage mirrored patterns in bias, with nominal (> 95%) coverage of richness, but poor coverage (between 0 and 27%) of the other two indices (Supporting information). Applying the threshold reduced positive bias and improved coverage of diversity for $q > 0$, but resulted in a negative bias and poorer coverage of species richness, particularly in more disturbed landscapes (Supporting information).

For both abundance and occupancy-based profiles, we did not find great discrepancies between the sampling station-based profiles and the landscape-wide profiles, even though the sampling stations only covered about 4% of the study area (for landscape wide profiles see the Supporting information).

For the comparison across landscapes the occupancy-based diversity profiles generally maintained the same rank order of diversity amongst the landscapes (Fig. 2A, for an example of results from a single site see the Supporting information) as the abundance-based diversity profiles. The three communities that showed significantly different species richness based on true abundance data were ordered the same by occupancy-based richness estimates in 97% of the simulated communities (Fig. 2B, Table 1). Similarly, the four and three landscapes that could be significantly distinguished for H′ and D were ordered the same as occupancy-based profiles in 96% and 94% of all simulations, respectively. The occupancy threshold slightly increased the ability to correctly rank simulated landscapes by species richness (to 100%), but slightly reduced it for H′ and D (Fig. 2C, Table 1).

Borneo bird community results

We detected 143, 118, 121 and 30 species in continuous forest, riparian forest, riparian remnant and oil palm river, respectively. In the diversity profiles, species richness was highest in continuous forest, followed by (in decreasing order) riparian remnant, riparian forest and oil palm river. At $q < 1$, continuous forest was the most diverse habitat type, while at $q > 1.5$ riparian forest was the most diverse

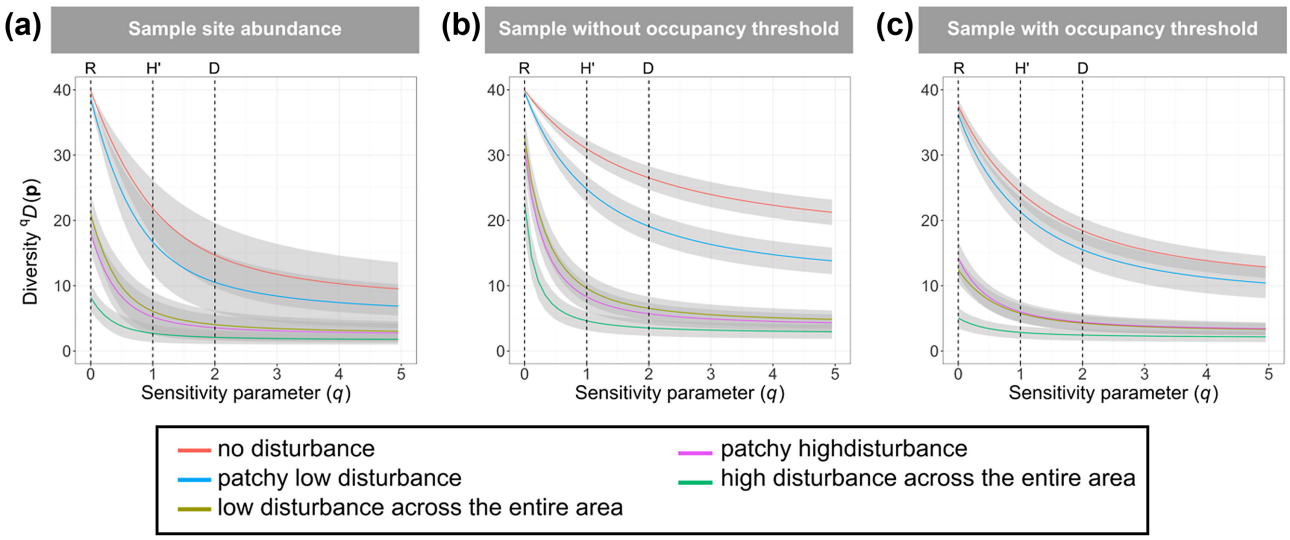

Figure 2. Comparison among landscapes of diversity profiles generated using (A) the true abundance at the 100 sample stations in each landscape, (B) occupancy based predictions at the 100 sample stations in each landscape without thresholding, (C) occupancy based predictions at the 100 sample stations in each landscape with thresholding. All results are shown as averages (solid line) of the 100 simulations with the uncertainty shown as standard deviations (grey shading).

(Fig. 3A), although the 95% CIs of the profiles for these two more diverse habitats overlapped. Although species richness in the riparian remnant was similar to continuous forests and riparian forest, the diversity decline of the profile was much greater, so that riparian remnants were significantly (non-overlapping 95% CIs) distinct from continuous forest and riparian forest at $q > 0.5$. Oil palm river and riparian remnant were always the habitats with the lowest and second lowest biodiversity, respectively, and at $q > 2$ the 95% CIs of these two profiles largely overlapped.

When we incorporated similarity (diet, taxonomic and phylogenetic), the overall community diversity was reduced for all habitat types. In line with species richness, the taxonomic richness (Fig. 3C) was similar for the riparian remnant, continuous forest and riparian forest. Continuous forest and riparian forest showed very similar overlapping profiles. The shape of the riparian remnant and oil palm river taxonomic profiles were very similar to the original profiles but overlap in the 95% CIs was even greater. When we considered dietary (Fig. 3B) and phylogenetic (Fig. 3D) similarity, we also saw a large reduction in the diversity of the communities. In both cases, all habitats showed very similar diversity profiles with widely overlapping 95% CIs.

With thresholding (Supporting information), the estimated species richness was lower for all habitat types. Profiles with the threshold showed similar patterns, but differences among habitats, particularly for the more depleted communities in the riparian remnant and the oil palm river, were less distinct.

Discussion and conclusion

Diversity profiles allow researchers to characterize and compare communities while considering the contributions of abundant and rare species, thus acknowledging the multidimensional nature of diversity (Morris et al. 2014). Occupancy-based diversity profiles provide a reliable inference framework for estimating biodiversity based on output from community occupancy models, which take into account imperfect and varying species detection. As expected, we found that occupancy-based diversity profiles generally mirrored true among-community patterns well, albeit with some shortcomings that are primarily related to overestimations of non-richness diversity measures.

Using occupancy model estimates of average occupancy probability to construct diversity profiles, with or without thresholding, generally maintained the same inter-landscape diversity pattern as observed in the true abundance diversity profiles (Fig. 2). Further, the two simulated landscapes that had very similar abundance-based profiles (landscape-wide low disturbance and local high disturbance) also showed very similar occupancy-based profiles. The general agreement of the occupancy and true abundance profiles suggests that

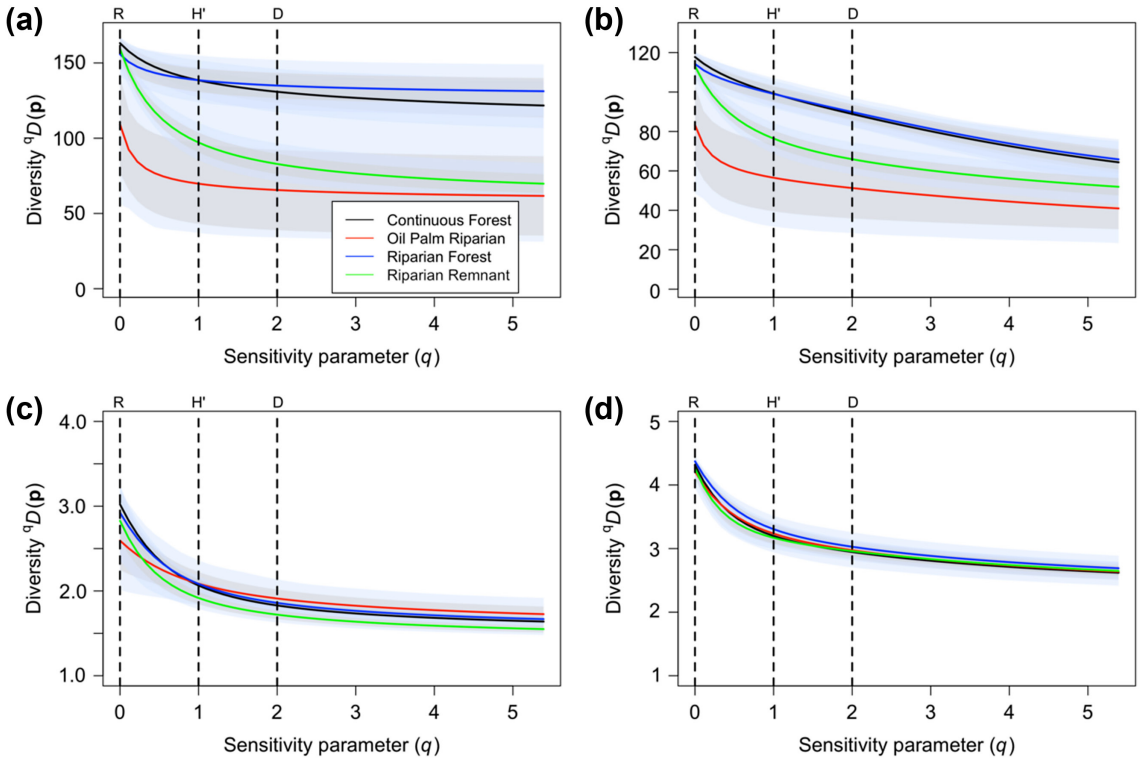

Figure 3. Diversity profiles for a bird community from Malaysian Borneo in four habitat types calculated using the occupancy predictions without thresholding. (A) Naive similarity matrix, (B) taxonomic similarity, (C) diet similarity, (D) phylogenetic similarity. The standard deviations (grey shading) and 95% credible intervals (blue shading) are included to quantify uncertainty.

detection/non-detection surveys may be sufficient to compare the multidimensional properties of diversity between landscapes. Similarly, the repeated collection of detection/non-detection data from one landscape will likely allow comparison of diversity through time, an important aspect of biodiversity monitoring.

Within a landscape, occupancy-based profiles generally overestimated diversity for $q > 0$. This is expected because occupancy approaches 1 as abundance increases; consequently, occupancy-based diversity should suggest more even communities (i.e, be higher) than abundance-based diversity. Further, Broms et al. (2015) also found positive bias when comparing their occupancy-based to true incidence-based Hill numbers and attributed that to positively biased estimates of occupancy in species with low detection rates. This bias is caused by the structure of the community occupancy framework, as rare species borrow information from common species (Broms et al. 2016, Guillera-Arroita et al. 2019), and may also contribute to the positive bias we observed. Similar to Broms et al. (2015), we saw no or low negative bias in estimates of species richness ($q = 0$), with exception of the most disturbed landscape, which had 33.8% positive bias in richness estimates, on average. Several factors likely contribute to this positive bias: First, the rarity of many of the species in this landscape translates into low species detection probabilities and thus, likely positive bias in occupancy, as explained above. Second, our approach to analyze data from all landscapes combined implies augmenting data for the most disturbed site with a large number of all-0 encounter histories (reflecting the implicit assumption that all species from the regional pool could potentially occur in the disturbed landscapes). Occupancy probability for these species, even if low, is non-0, and thus contributes to richness estimates. The potential for this source of bias is lower in landscapes where true richness is closer to the 40-species threshold. As an alternative, we could have analyzed data from each landscape separately, conditioning on the number of species observed in each landscape. That, however, would have induced negative bias in richness (and possibly other diversity) estimates, as not all species present were always observed.

We attempted to overcome the overestimation of diversity from community occupancy models by implementing an occupancy threshold. Liu et al. (2016) suggest the use of randomly selected points as pseudo-absences for threshold determination with incidence data. Here, we used the output of the occupancy model to generate pseudo-absences based on estimated occupancy probability. This approach likely leads to more realistic pseudo-absences than completely random generation as the use of modeled occupancy probabilities allows for a more informed selection. Incorporating a threshold into occupancy-based diversity profile calculation had mixed results. The threshold reduced the overestimation of diversity at $q > 0.5$ (Supporting information). At the same time, however, thresholding often resulted in underestimated species richness, particularly for more disturbed landscapes (Supporting information) for $q > 0$. Interestingly, the use of a threshold did not improve, but rather slightly reduced our ability to correctly replicate true underlying differences in diversity among landscapes (Table 1). In the empirical dataset, we saw a similar reduction in the distinctiveness of diversity profiles for the riparian remnants and the oil palm rivers, whose 95% CIs largely overlapped for $q > 1$ when a threshold was applied. As we expect that the main objective of most biodiversity research and monitoring projects is the comparison of profiles between sites or of the same site across time, we recommend to not use the threshold we applied. In this case, researchers should be aware that species richness and potentially entire diversity profiles may be overestimated when detection is low for many species. We acknowledge that we only explored one threshold and that effects may be different for other threshold methods; this is an avenue for future development of this approach.

Acknowledging the differences between abundance and occupancy, we do not suggest that occupancy probability can simply be used as an index or surrogate for abundance. Our results do, however, indicate that occupancy-based diversity measures and profiles can reflect patterns in diversity despite the loss of information entailed in using occupancy rather than abundance data. This is a promising finding for biodiversity research and monitoring, as community-wide species detection/non-detection data are generally much easier and cheaper to obtain than data for abundance estimation (Joseph et al. 2006, Kéry and Schmidt 2008). In addition to traditional methods used to detect wildlife, such as points counts and visual transects, a growing number of technologies are available for detecting and identifying biodiversity, such as automated acoustic recorders (Bush et al. 2017) or eDNA and metabarcoding (Bush et al. 2017). These new technologies present powerful methods to collect community level detection/non-detection data that could be combined with occupancy-based diversity profiles.

It is important to note that with an average detection probability of 0.5 and 10 sampling occasions, our simulated data represent a scenario where most species in a community are detected. These conditions are representative of typical camera-trap surveys in which the number of possible repeat visits is much less limited by logistics than methods requiring a human observer to return to each sampling location on each occasion (e.g. point counts). Based on our findings for the most disturbed landscape (e.g. strong positive bias in richness estimates from community occupancy models), we expect sparser data from surveys with lower species detectability and/or fewer repeated visits to increase the potential for positive bias in diversity profiles. Further studies are needed to explore the effects of varying p and k on occupancy-based diversity profiles.

The diversity profile framework presented here also allows for the incorporation of trait similarities between species by defining a similarity matrix. Incorporating species trait similarities can be an additional way to display diversity in a community as it puts a greater emphasis on more dissimilar species (Leinster and Cobbold 2012). From an ecological perspective, accounting for such similarities reduces the functional redundancies in the community, for example, species

having the same dietary niche could functionally replace each other (Rosenfeld 2002, Olden et al. 2004). Phylogenetically and functionally diverse communities are known to better maintain ecosystem stability (Cadotte et al. 2011, 2012). Therefore, considering these additional dimensions of diversity provides a more complete picture of a community (Rodrigues and Gaston 2002) and may improve predictions of ecosystem function and resilience.

Our empirical data showed that considering dietary, taxonomic or phylogenetic similarities among bird species led to very similar diversity profiles for all habitat types. In the case of this bird dataset, all taxonomic, phylogenetic and dietary groups were present in all habitats. As a result, even at considerably lower species diversity, disturbed habitats such as oil palm plantations maintained dietary and phylogenetic diversity of birds essentially identical to that of continuous forests. This is surprising, given that previous studies have found that dietary traits and taxonomy (among other characteristics) can affect response to habitat alterations and extinction risk in birds (Russell et al. 1998, Boyer 2010, Frishkoff et al. 2014). Despite this apparent maintenance of phylogenetic and functional diversity, the loss of overall species diversity in more disturbed habitats suggest a loss in redundancy, another measure that has been associated with ecosystem stability (Naeem 1998).

The diversity profiles of the bird communities reinforced the findings by Mitchell at el. (2018) that riparian remnants supported similar diversity value to continuous logged forest habitats (both riparian and non-riparian). However, when evenness of the community is given more weight (i.e. when $q > 1$), riparian remnants have reduced diversity compared to logged forests. This suggests that much of the diversity in the habitat remnants (such as when measured via species richness directly), manifests from a number of species occurring rarely. If a greater proportion of the community in remnant occurs only rarely, this suggests such remnants may not sustain certain species in the long-term (i.e. we may be observing an extinction debt) and effectively act as population sinks from continuous forest habitats, a finding which is not apparent from assessing only species richness and community integrity as undertaken by Mitchell et al. (2018).

Beyond exploring alternative threshold approaches and the effects of sample size on occupancy-based diversity profiles, there are several opportunities to further develop the application of detection-corrected diversity profiles to typical wildlife survey data. We focused on detection/non-detection data as these are the most easily collected for communities of difficult-to-study species (though 'easy' is a relative term). When repeated count data are available, as is often the case for bird surveys, a community N-mixture model (Royle 2004, Yamaura et al. 2016) could be used to construct detection-corrected abundance-based diversity profiles. Distance sampling data on a community of species (Sollmann et al. 2016) could be used for that same purpose. Even with species-level detection/non-detection data, researchers could employ the Royle–Nichols occupancy model (Royle and Nichols 2003), which exploits the relationship between abundance and species detection probability to estimate local abundance from detection/non-detection data. At the same time, our results highlights that even a basic community occupancy model will in most cases be sufficient to compare different sites in terms of diversity.

In practice, information on species occurrence is often used to help develop management decisions and conservation strategies (Guisan et al. 2013). For many species of conservation concern, the detection/non-detection surveys underlying estimates of occurrence are the main source of information on their population status, and therefore have a significant role in setting conservation priorities (MacKenzie 2005, Joseph et al. 2006). They are useful for a wide range of purposes from estimating changes in occurrence to identifying high conservation priority areas (Zipkin et al. 2010, Olea and Mateo-Tomas 2011, Tilker et al. 2020). Occupancy-based diversity profiles are an important contribution to the occupancy toolkit as they allow comparing biodiversity across space and time while accounting for imperfect and varying detection. Specifically, these profiles can be used to: 1) monitor the diversity of a community over time and to evaluate the effectiveness of management/conservation efforts, and 2) compare general patterns of diversity according to different habitat, disturbance or trait regimes, helping to set conservation priorities. Incorporating this approach into conservation should improve biodiversity assessments of species and communities.

Data availability statement

Scripts and model code are available on github (<https://github.com/jabrams23/occudiversity>). For access to the data used in this study please contact the corresponding author.

Acknowledgements – We thank the German Federal Ministry of Education and Research (BMBF FKZ: 01LN1301A; FKZ: 01LC1703A) and the Leibniz Institute for Zoo and Wildlife Research for funding this research. SLM and MJS were funded by the UK Natural Environment Research Council (NERC: NE/K016407/1). SLM was supported by a PhD scholarship jointly funded by University of Kent and NERC. We thank the Sabah Biodiversity Council, Sabah Forest Department, Yayasan Sabah, Sime Darby, Benta Wawasan, Sabah Softwoods and Innoprise Foundation for permitting site access. We are grateful to David Coombs and Tom Swinfield for providing LiDAR data used as covariates in the occupancy model. Open access funding enabled and organized by Projekt DEAL.
Conflict of interest – The authors declare no conflict of interests.

Author contributions

Jesse F. Abrams: Conceptualization (equal); Data curation (equal); Formal analysis (lead); Methodology (equal); Visualization (lead); Writing – original draft (lead); Writing – review and editing (equal). **Rahel Sollmann**: Conceptualization (equal); Formal analysis (supporting);

Methodology (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Simon L. Mitchell**: Data curation (equal); Writing – original draft (supporting); Writing – review and editing (equal). **Matthew J. Stuebig**: Funding acquisition (lead); Writing – original draft (supporting); Writing – review and editing (equal). **Andreas Wilting**: Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Writing – original draft (supporting); Writing – review and editing (equal).

References

Boyer, A. G. 2010. Consistent ecological selectivity through time in Pacific island avian extinctions. – Conserv. Biol. 24: 511–519.

Broms, K. M. et al. 2015. Accounting for imperfect detection in Hill numbers for biodiversity studies. – Methods Ecol. Evol. 6: 99–108.

Broms, K. M. et al. 2016. Model selection and assessment for multi-species occupancy models. – Ecology 97: 1759–1770.

Bush, A. et al. 2017. Connecting Earth observation to high-throughput biodiversity data. – Nat. Ecol. Evol. 1: 0176.

Cadotte, M. W. et al. 2011. Beyond species: functional diversity and the maintenance of ecological processes and services. – J. Appl. Ecol. 48: 1079–1087.

Cadotte, M. W. et al. 2012. Phylogenetic diversity promotes ecosystem stability. – Ecology 93: S223–S233.

Chao, A. et al. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. – Ecol. Monogr. 84: 45–67.

Daly, A. et al. 2018. Ecological diversity: measuring the unmeasurable. – Mathematics 6: 119.

Dorazio, R. M. and Royle, J. A. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. – J. Am. Stat. Assoc. 100: 389–398.

Dorazio, R. M. et al. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. – Ecology 87: 842–854.

Ellis, M. V. and Taylor, J. E. 2018. Effects of weather, time of day and survey effort on estimates of species richness in temperate woodlands. – Emu-Austral Ornithol. 118: 183–192.

Freeman, E. A. and Moisen, G. 2008. PresenceAbsence: an R package for presence–absence model analysis. – J. Stat. Softw. 23: 1–31.

Frishkoff, L. O. et al. 2014. Loss of avian phylogenetic diversity in neotropical agricultural systems. – Science 345: 1343–1346.

Gattone, S. A. and Battista, T. D. 2009. A functional approach to diversity profiles. – J. R. Stat. Soc. Ser. C 58: 267–284.

Gelman, A. et al. 1996. Posterior predictive assessment of model fitness via realized discrepancies. - Stat. Sin. 6: 733-760.

Gelman, A. et al. 2004. Bayesian data analysis. - Chapman and Hall.

Guillera-Arroita, G. et al. 2019. Inferring species richness using multispecies occupancy modeling: estimation performance and interpretation. – Ecol. Evol. 9: 780–792.

Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – Ecology 54: 427–432.

Joseph, L. N. et al. 2006. Presence–absence versus abundance data for monitoring threatened species. – Conserv. Biol. 20: 1679–1687.

Jost, L. 2006. Entropy and diversity. – Oikos 113: 363–375.

Kéry, M. and Schmidt, B. R. 2008. Imperfect detection and its consequences for monitoring for conservation. – Commun. Ecol. 9: 207–216.

Leinster, T. and Cobbold, C. A. 2012. Measuring diversity: the importance of species similarity. – Ecology 93: 477–489.

Liu, C. et al. 2016. On the selection of thresholds for predicting species occurrence with presence-only data. – Ecol. Evol. 6: 337–348.

Loreau, M. et al. 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. – Science 294: 804–808.

MacKenzie, D. I. 2005. What are the issues with presence–absence data for wildlife managers? – J. Wildl. Manage. 69: 849–860.

MacKenzie, D. I. and Kendall, W. L. 2002. How should detection probability be incorporated into estimates of relative abundance? – Ecology 83: 2387–2393.

MacKenzie, D. I. et al. 2002. Estimating site occupancy rates when detection probabilities are less than one. – Ecology 83: 2248–2255.

MacKenzie, D. I. et al. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. – Academic Press.

Mitchell, S. L. et al. 2018. Riparian reserves help protect forest bird communities in oil palm dominated landscapes. – J. Appl. Ecol. 55: 2744–2755.

Morris, E. K. et al. 2014. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. – Ecol. Evol. 4: 3514–3524.

Naeem, S. 1998. Species redundancy and ecosystem reliability. – Conserv. Biol. 12: 39–45.

Nichols, J. D. et al. 1998. Inference methods for spatial variation in species richness and community composition when not all species are detected. – Conserv. Biol. 12: 1390–1398.

Olden, J. D. et al. 2004. Ecological and evolutionary consequences of biotic homogenization. – Trends Ecol. Evol. 19: 18–24.

Olea, P. P. and Mateo-Tomas, P. 2011. Spatially explicit estimation of occupancy, detection probability and survey effort needed to inform conservation planning. – Divers. Distrib. 17: 714–724.

Patil, G. P. 2014. Diversity profiles. – In: Balakrishnan, N. et al. (eds), Wiley StatsRef: statistics reference online. Wiley.

Patil, G. P. and Taillie, C. 1982. Diversity as a concept and its measurement. – J. Am. Stat. Assoc. 77: 548–561.

Plummer, M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. – In: Proc 3rd Int. workshop on distributed statistical computing, pp. 20–22. <www.ci.tuwien.ac.at/conferences/DSC-2003>.

Plummer, M. 2019. rjags: Bayesian graphical models using MCMC. – R package ver. 4-9.

Preston, F. W. 1948. The commonness, and rarity, of species. – Ecology 28: 254–283.

Purvis, A. and Hector, A. 2000. Getting the measure of biodiversity. – Nature 405: 212.

Rodrigues, A. S. and Gaston, K. J. 2002. Maximising phylogenetic diversity in the selection of networks of conservation areas. – Biol. Conserv. 105: 103–111.

Rosenfeld, J. S. 2002. Functional redundancy in ecology and conservation. – Oikos 98: 156–162.

Royle, J. A. 2004. N-mixture models for estimating population size from spatially replicated counts. – Biometrics 60: 108–115.

Royle, J. A. and Dorazio, R. M. 2008. Hierarchical modeling and inference in ecology. – Academic Press.
Royle, J. A. and Nichols, J. D. 2003. Estimating abundance from repeated presence–absence data or point counts. – Ecology 84: 777–790.
Russell, G. J. et al. 1998. Present and future taxonomic selectivity in bird and mammal extinctions. – Conserv. Biol. 12: 1365–1376.
Sollmann, R. et al. 2013. Risky business or simple solution–relative abundance indices from camera-trapping. – Biol. Conserv. 159: 405–412.
Sollmann, R. et al. 2016. A hierarchical distance sampling model to estimate abundance and covariate associations of species and communities. – Methods Ecol. Evol. 7: 529–537.
Sollmann, R. et al. 2017. Quantifying mammal biodiversity co-benefits in certified tropical forests. – Divers. Distrib. 23: 317–328.
Sutherland, C. et al. 2016. A multiregion community model for inference about geographic variation in species richness. – Methods Ecol. Evol. 7: 783–791.
Tilker, A. et al. 2020. Identifying conservation priorities in a defaunated tropical biodiversity hotspot. – Divers. Distrib. 26: 426–440.
Tilman, D. et al. 2014. Biodiversity and ecosystem functioning. – Annu. Rev. Ecol. Evol. Syst. 45: 471–493.
Yamaura, Y. et al. 2016. Study of biological communities subject to imperfect detection: bias and precision of community N-mixture abundance models in small-sample situations. – Ecol. Res. 31: 289–305.
Zipkin, E. F. et al. 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. – Biol. Conserv. 143: 479–484.
Zipkin, E. F. et al. 2012. Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. – Ecol. Appl. 22: 1962–1972.