Title:
Capturing biodiversity complexities while accounting for imperfect detection: the application of occupancy-based diversity profiles.

Running headline:
Occupancy-based diversity profiles

Authors:
Jesse F. Abrams [1,2]*, Rahel Sollmann [3], Simon L. Mitchell [4], Matthew J. Struebig [4], and Andreas Wilting [1]

[1] Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315, Berlin, Germany
[2] Global Systems Institute and Institute of Data Science and Artificial Intelligence, University of Exeter, Exeter EX44PY, United Kingdom
[3] Department of Wildlife, Fish, and Conservation Biology, University of California Davis, Davis, CA 95616, USA
[4] Durrell Institute of Conservation and Ecology (DICE), School of Anthropology and Conservation, University of Kent, Canterbury, Kent, UK

* corresponding author: abrams@izw-berlin.de

Abstract

1. 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 all the common diversity indices. However, diversity profiles require information on species abundance and, although it has been shown that detectability varies greatly between species and study sites, raw count data that do not account for imperfect and varying detection are often used to assess diversity. Hierarchical occupancy models explicitly account for variation in detectability, but they only provide information on occurrence, not abundance.
2. Here, we used simulated animal communities and empirical data from a megadiverse tropical bird community to develop a method that allows estimation of diversity profiles from occupancy models. Specifically, we altered the diversity profile formula to use occupancy probabilities instead of abundance. We further tested a novel occupancy thresholding approach to reduce potential biases in the estimated diversity profiles. Finally, we extended the framework to consider trait or phylogeny-based similarity when calculating diversity profiles for our empirical dataset.
3. The simulation study showed that occupancy-based diversity profiles produced among-community patterns in diversity similar to true abundance diversity profiles, although estimates of occupancy were positively biased for species with low detection probability. While this positive bias could be reduced with occupancy thresholding, the threshold resulted in negative biases in species richness estimates and reduced the ability to reproduce true differences among the simulated communities. Application of our approach to a large bird dataset revealed differential diversity patterns in

communities of different habitat types. Accounting for phylogenetic and ecological similarities between species, however, reduced the diversity within and among habitats.

4. Our framework allows for the investigation of the complexity of diversity for incidence 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. Therefore, our extension to the diversity profile framework will be a useful tool for studying and monitoring biodiversity.

Keywords
Biodiversity, Diversity index, Diversity profile, presence, occupancy, species distribution modeling, specificity, threshold.

1. Introduction

Biological diversity represents the variety of organisms or traits and plays a central role in ecological theory (Loreau et al., 2001; Tilman, Isbell, & Cowles, 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, Baetens, & De Baets, 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, Baetens, & De Baets, 2018).

The most commonly used diversity indices are species richness, Shannon's diversity H' and Simpson's diversity D1; the latter two combine measures of richness and abundance, whereas species richness solely presents the number of detected 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 & Hector, 2000; Daly, Baetens & De Baets, 2018).

To address this shortcoming, several researchers have suggested using parametric families of diversity indices (Hill, 1973; Patil & Taillie, 1982; Gattone & Battista, 2009; Leinster & 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). Plotting the effective number of species as a function of q allows us to view diversity from multiple vantage points (Hill, 1973; Leinster & 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 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 & 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 & 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 & 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 detection/non-detection data. Even with incidence data, however, we must consider that species may be detected imperfectly (MacKenzie et al., 2002; MacKenzie et al., 2006).

Occupancy modelling provides a framework to handle the problem of imperfect detection, producing unbiased estimates of species occurrence (MacKenzie et al., 2002, 2006). The development of hierarchical multi-species occupancy models (Dorazio & 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 (Guillera-Arroita, Kéry, & Lahoz-Monfort, 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, Hooten, & Fitzpatrick (2014) followed with a formulation to calculate occupancy-based Hill numbers. Here, we extend this framework to facilitate the calculation, visualization, and thus, interpretation of occupancy-based diversity profiles. We first explore the use of occupancy-based diversity profiles to compare communities across landscapes with varying levels of habitat disturbance using simulated data. 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).

2. Methods

Our study followed five steps: (1) simulation of a forest degradation gradient, animal communities, and detection data of those communities, (2) community occupancy analysis for simulated detection data, (3) utilization of community occupancy model output to construct occupancy-based diversity profiles, (4) application of thresholding to occupancy-based profiles, and (5) application of community occupancy diversity profiles to an empirical bird dataset from Malaysian Borneo. We evaluated the performance of the occupancy-based diversity profiles to true community abundance diversity profiles. We repeated the community simulation and analysis 100 times.

2.1 Forest degradation and community simulation

We simulated five virtual 10 x 10 km forest landscapes, with 200 x 200 m grid cells (50 x 50 cells). We simulated a habitat covariate representing “habitat disturbance” (where 0 represents undisturbed forest and 5 represents complete deforestation) for each landscape (Fig. 2A) 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 ϕ to specify how the relationship to other

cells changes with distance (Fig. S1). 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, (3) low disturbance across the entire area, (4) patchy high disturbance, and (5) high disturbance across the entire area.

We then simulated the cell-level abundance of forty virtual species, *i*, at grid cell *j* (*j* = 1, 2,...,*n*) in our 5 landscapes to generate "true" abundance-based diversity profiles. We set the average response of species to habitat disturbance ($\mu_1 = -2$) as negative since most forest-adapted species will respond to logging negatively (Fig. S2). 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 abundance per species per grid cell at habitat=0 (i.e. μ_0) was 1. This resulted in five communities with species richness ranging from 18 to 40. See Supplementary materials for full simulation details.

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. We repeated the observation process for 10 occasions, *k*. Species-specific detection on the logit scale, $\alpha 0$, was drawn from a normal distribution with $\mu_\alpha = 0$ and $\sigma_\alpha = 1$.

Finally, we reduced the resulting observed counts 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 constructed and saved abundance diversity profiles for later comparison to the occupancy-based diversity profiles. All calculations were carried out in R version 3.6.0 (R Core Team, 2019).

2.2 Community occupancy model

We adopted the hierarchical formulation of occupancy models by Royle & Dorazio (2008) extended to a community occupancy model (Dorazio & Royle, 2005; Dorazio et al., 2006). To analyze our simulated data, following common practice in analyzing field data, we 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 reserves and among species. We further modelled species-specific effects on occupancy of the simulated habitat "disturbance" covariate ($\beta 1$, Fig. 2). Detection probability included a species-specific random intercept with landscape specific hyperparameters, to allow for differences in baseline detection among reserves. In our case, the different landscapes had different abundances of animals, which leads to differences in species-level detection (Royle & Nichols, 2003). 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 model 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).

2.3 Diversity profiles

In order to investigate the performance of estimating diversity profiles with occupancy-based information, we first constructed diversity profiles for the simulated true abundance across the whole landscape, following Leinster & Cobbold (2012). Diversity profile values (${}^{q}D^{Z}$) for abundance data can be calculated according to (Leinster & Cobbold, 2012) as:

$${}^{q}D^{Z}|\lambda = \left(\sum_{i=1}^{M} \frac{\lambda_i}{\Sigma(\lambda)} (Z_\lambda)_i^{q-1}\right)^{\frac{1}{1-q}} \qquad \textbf{[Equation i]}$$

where

$$(Z_\lambda)_i = \sum_{h=1}^{M} Z_{ih} \left(\frac{\lambda_h}{\Sigma(\lambda)}\right), \qquad \textbf{[Equation ii]}$$

M is the number of species in the assemblage, and the *i*th 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, ${}^{q}D^{Z}$ 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 ${}^{q}D^{Z}$ for more uneven assemblages than for more even assemblages. The diversity profile framework from Leinster & Cobbold (2012) allows for the consideration of similarity between species through the inclusion of an *M* x *M* similarity matrix $\mathbf{Z}$ which represents the similarity between the *i*th and *h*th species. Values of 0 in $\mathbf{Z}$ indicate total dissimilarity, whereas values of 1 indicating 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 (see 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 & Cobbold (2012) as follows:

$${}^{q}D^{Z}|\bar{\psi} = \left(\sum_{i=1}^{M} \frac{\bar{\psi}_i}{\Sigma(\bar{\psi})} \left(Z_{\bar{\psi}}\right)_i^{q-1}\right)^{\frac{1}{1-q}} \qquad \textbf{[Equation 1]}$$

where

$$\left(Z_{\bar{\psi}}\right)_i = \sum_{h=1}^{M} Z_{ih} \left(\frac{\bar{\psi}_h}{\Sigma(\bar{\psi})}\right) \qquad \textbf{[Equation 2]}$$

where *M* is again the number of species in the assemblage, $\bar{\psi}_i$ is the average occupancy probability of species *i*, the *i*th species has relative occupancy $\frac{\bar{\psi}_i}{\Sigma(\bar{\psi})}$, and $\mathbf{Z}$ is the similarity matrix.

We used the parameter estimates from the occupancy model fit to our simulated data to predict species distributions across the five simulated landscapes (i.e., not just the sampling points) and then constructed diversity profiles using the mean landscape-wide occupancy

probability for each species using Equations 1 and 2 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 and 95% credible intervals to provide a sense of the uncertainty surrounding the estimated profiles. We then visually compared the resulting occupancy-based profiles against the true abundance profiles and compared estimates of species richness, Shannon diversity, and Simpson's index, the three most commonly used diversity metrics.

In some cases, it is not possible to predict occupancy across the entire landscape (missing covariates for unsurveyed cells). Therefore, we also calculated diversity profiles for both true abundance and occupancy data for just the sampling points and compared these to the landscape scale profiles.

2.4 Occupancy Threshold

To explore methods to account for overestimation of richness in occupancy models and improve the agreement between true abundance and estimated occupancy-based diversity profiles we tested the use of an occupancy threshold. 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.

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} (Fig. S3) threshold for each species for each landscape based on the mean occupancy estimate using the *optimal.thresholds* function from the R package *PresenceAbsence* (Freeman & 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.

2.5 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 the 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.

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 on the probability of detection (e.g., Ellis & Taylor, 2018).

We separated species communities according to the four habitat-types described above for diversity profile construction with and without occupancy thresholds. We did not do landscape scale predictions for the case study, but instead used the occupancy probabilities for each sampling station to construct the diversity profiles. Additionally, we constructed similarity matrices (see Supplementary Material) according to diet, taxonomy, and phylogeny to demonstrate how similarity can be incorporated into our occupancy-based diversity profile framework.

3. Results

3.1 Simulation results

The profiles generated for the 100 community simulations are shown in Figure 1. The occupancy-based diversity profiles show similar trends in diversity as the diversity profiles based on true abundance (Fig. 1).

Both the predicted and sampled occupancy-based diversity profiles showed similar trends in diversity when compared to the respective true abundance diversity profiles (example results from one simulation are shown Figs. 2 and 3). In the landscape-wide predicted occupancy profiles the occupancy-based estimates of species richness were generally similar to true richness. Only for the more disturbed landscapes where species abundance and predicted occupancy probabilities were low there was a positively bias in estimated species richness. In the most degraded landscape, estimated species richness was inflated by an average of 50% over the 100 simulations (8 equivalent species), although the 95% Bayesian Credible Intervals (BCIs) of the occupancy-based estimation still overlapped true richness. For all other landscapes the species richness was underestimated by an average of 3% over all landscapes over all simulations when compared to the true abundance estimates. For all five landscapes, at values of $q > 0$ there was a positive bias in the non-threshold occupancy-based profiles when compared to the true abundance profiles (Fig. 2). The occupancy-based estimates averaged over the 5 landscapes and 100 simulations for Shannon diversity (q=1, D) and Simpson's index (q=2, H') were an average of 57% and 80% higher, respectively, than their abundance-based counterparts. Positive bias in profiles was stronger for $q > 1$ than for $q < 1$ and was larger for the less disturbed landscapes (no disturbance and local low disturbance), with 95% BCIs not overlapping the true profiles (at $q > 1$) (Fig. 2B). The application of an occupancy threshold (Fig. 2C) resulted in negative bias in estimates of species richness for the three more disturbed landscapes, as the 95% BCI of the occupancy predictions did not overlap true richness. In these three more disturbed sites species had lower occupancies; communities were, therefore, more vulnerable to the exclusion of some species after application of the threshold. For the two less disturbed landscapes the threshold-based

estimates of richness were comparable to the true abundance estimates. The threshold improved the agreement between the occupancy and abundance-based profiles for values of $q > 0$, with the 95% BCIs of the threshold-corrected occupancy-based profiles largely overlapping the abundance-based true profiles, even for the less disturbed sites. For example, for Shannon diversity and Simpson's index the average positive bias in the occupancy-based estimates was reduced from 57% to 13% and 80% to 34%, respectively.

We obtained similar results for the occupancy-based diversity profiles using only the occupancy estimates for the sampling stations (Fig. 3) as we did for the landscape-wide occupancy-based diversity profiles, even though the former only covered about 4% of the study area. The positive bias in species richness estimates for the more disturbed landscapes, however, was more pronounced and for the three more disturbed landscapes the 95% BCIs of the occupancy-based estimate did not overlap the true value. Similar to the landscape-wide occupancy profiles, species richness was underestimated with threshold application (Fig. 3C). However, profile values at $q > 1$ for landscapes with no disturbance (i.e., high occupancy) were also underestimated. In fact, the true abundance profile for this landscape did not even fall within the occupancy profile's 95% BCI. With threshold application, the bias in occupancy-based profiles changed from 63% to 14% (averaged across the five landscapes) and 84% to 35% for Shannon diversity (q=1) and Simpson's index (q=2), respectively.

The comparison between the true abundance-based diversity profiles (Fig 4A and 4B) revealed clear differences between four out of five simulated landscapes. Only the landscapes with low disturbance and local high disturbance showed rather similar profiles. The occupancy-based diversity profiles maintained the same order of diversity amongst the landscapes (Fig. 4C and 4D), differentiating the distinct simulated landscapes well. In particular, the two less disturbed landscapes showed non-overlapping 95% BCIs. Although the most disturbed landscape showed a distinct profile, its 95% BCI partly overlapped with the two other more disturbed landscapes. The application of the occupancy threshold greatly reduced the ability to differentiate between our simulated landscapes (Fig 4E & 4F). Although the diversity profile in the no-disturbance landscape was still higher than in the low-disturbance landscape, the BCIs of the landscapes now overlapped. Similarly, the BCIs between the three disturbed landscapes greatly overlapped following the application of the threshold.

3.2 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 (Fig. 5A). Although the BCIs of the profiles for these two more diverse habitats overlapped, this indicates that the continuous forest community is the richest in species but that this community is less even than the riparian forest community. 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 BCIs) 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% BCIs 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. 5C) 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% BCIs was even greater. When we consider dietary (Fig. 5B) and phylogenetic (Fig. 5D) 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% BCIs.

With thresholding (see Fig. S4), 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.

4. 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). We aimed to provide a reliable inference framework for estimating biodiversity via diversity profiles based on output from occupancy models, which take into account imperfect and varying species detection. We found that occupancy-based diversity profiles generally mirrored the true community composition well, albeit with some shortcomings.

Using occupancy model estimates to construct diversity profiles, with or without thresholding, maintained the same inter-landscape diversity pattern as observed in the true abundance diversity profiles. 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 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 to compare diversity through time, an important aspect of biodiversity monitoring.

Within a landscape, the match between abundance-based and occupancy-based profiles was generally good, although occupancy-based profiles often overestimated diversity. This is similar to what Broms, Hooten, & Fitzpatrick (2014) found with their occupancy-based Hill number approach and is a result of 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 and, thus, their distribution may be overestimated (Guillera-Arroita, Kéry, & Lahoz-Monfort, 2019).

Some degree of mismatch between abundance and occupancy-based diversity profiles was expected, as occupancy is a coarser measure, containing less information about how relatively abundant or rare a given species is in its community. Acknowledging these difficulties, we do not suggest that occupancy probability can simply be used as an index or surrogate of abundance. Our results do, however, indicate that under conditions representative of the methods commonly used in wildlife research, 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 & 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 incidence data that could be combined with occupancy-based diversity profiles.

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 probably 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 improved the match between the estimated occupancy-based and true abundance-based profiles (Fig. 2), especially in more diverse and less disturbed communities. At the same time, however, the threshold often led to negative bias in the occupancy-based diversity profiles (Figs. 2 & 3). Species richness, in particular, was strongly and significantly underestimated when using a threshold. Interestingly, the use of a threshold did not improve, but rather reduced, our ability to correctly replicate true underlying differences in diversity among landscapes, particularly for $q > 1$ in the more disturbed landscapes. 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% BCIs largely overlapped for $q > 1$ when a threshold was applied. The application of the threshold therefore depends on the research question and the diversity within the community. If the goal is the comparison of profiles between sites or of the same site across time, we recommend to not use a threshold. In this case, researchers should be aware that species richness (especially in disturbed, less diverse landscapes) as well as the overall diversity profile at $q > 0.5$ (particularly in undisturbed more diverse landscapes) may be overestimated.

Occupancy-based diversity profiles derived from landscape scale predictions were comparable with occupancy-based profiles using only the sampling stations, which covered 4% of our simulated landscapes. This indicates that reliable profiles can be generated in situations where landscape-wide predictions are not possible due to missing covariate information at unsampled locations. However, this only holds when sampling is representative of the entire landscape. The risk of non-representative sampling is higher in more heterogeneous and more logistically challenging landscapes or if the number of stations is much lower (the simulated 100 stations is a high number for many field projects).

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, Carscadden & Mirotchnick, 2011; Cadotte, Dinnage & Tilman, 2012). Therefore, considering these additional dimensions of diversity provides a more complete picture of a community (Rodrigues & 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 (e.g., 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 Mitchel 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).

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 & 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.

Acknowledgements

We would like to 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.

Data accessibility

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.

Conflict of interest

The authors declare no conflict of interests.

Authors' contributions

JFA, RS, and AW conceived the idea of the study. JFA performed the simulation study analysis. SLM and MJS collected the data for the case study and generated covariate information. JFA performed the occupancy analysis for the case study. JFA lead the writing of the manuscript. All authors read, contributed to, and approved the final manuscript.

References

Boyer, A. G. (2010). Consistent ecological selectivity through time in Pacific island avian extinctions. *Conservation Biology*, *24*(2), 511-519.

Broms, K. M., Hooten, M. B., & Fitzpatrick, R. M. (2015). Accounting for imperfect detection in Hill numbers for biodiversity studies. *Methods in Ecology and Evolution, 6*(1), 99-108.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., ... & Dawson, T. P. (2017). Connecting Earth observation to high-throughput biodiversity data. *Nature Ecology & Evolution*, *1*(7), 0176.

Cadotte, M. W., Carscadden, K., & Mirotchnick, N. (2011). Beyond species: functional diversity and the maintenance of ecological processes and services. *Journal of Applied Ecology*, *48*(5), 1079-1087.

Cadotte, M. W., Dinnage, R. & Tilman, D. (2012). Phylogenetic diversity promotes ecosystem stability. *Ecology 93*: S223-S233.

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs, 84*(1), 45-67.

Daly, A., Baetens, J., & De Baets, B. (2018). Ecological Diversity: Measuring the Unmeasurable. *Mathematics, 6*(7), 119.

Dorazio, R. M. & Royle, J. A (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association* 100, 389–398.

Dorazio, R. M., Royle, J. A., Söderström, B. & Glimskär, A (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* 87, 842–854.

Ellis, M. V., & Taylor, J. E. (2018). Effects of weather, time of day, and survey effort on estimates of species richness in temperate woodlands. *Emu-Austral Ornithology, 118*(2), 183-192.

Freeman, E. A. and Moisen, G. (2008). PresenceAbsence: An R Package for Presence-Absence Model Analysis. *Journal of Statistical Software, 23*(11):1-31.

Frishkoff, L. O., Karp, D. S., M'Gonigle, L. K., Mendenhall, C. D., Zook, J., Kremen, C., ... & Daily, G. C. (2014). Loss of avian phylogenetic diversity in neotropical agricultural systems. *Science*, *345*(6202), 1343-1346.

Gattone, S. A., & Battista, T. D. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society: Series C, 58*(2), 267-284.

Guillera-Arroita, G., Kéry, M., & Lahoz-Monfort, J. J. (2019). Inferring species richness using multispecies occupancy modeling: Estimation performance and interpretation. *Ecology and Evolution, 9*(2), 780-792.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., ... & Martin, T. G. (2013). Predicting species distributions for conservation decisions. *Ecology Letters, 16*(12), 1424-1435.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology, 54*(2), 427-432.

Joseph, L. N., Field, S. A., Wilcox, C., & Possingham, H. P. (2006). Presence–absence versus abundance data for monitoring threatened species. *Conservation Biology*, *20*(6), 1679-1687.

Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363-375.

Kéry, M., & Schmidt, B. R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, *9*(2), 207-216.

Kéry, M., & Royle, J. A. (2009). Inference about species richness and community structure using species-specific occupancy models in the national Swiss breeding bird survey MHB. In Modeling demographic processes in marked populations (pp. 639-656). Springer, Boston, MA.

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology, 93*(3), 477-489.

Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution, 6*(1), 337-348.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., ... & Tilman, D. (2001). Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science, 294*(5543), 804-808.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology, 83*, 2248-2255.

MacKenzie, D. I., & Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance?. *Ecology*, *83*(9), 2387-2393.

MacKenzie, D. I. (2005). What are the issues with presence-absence data for wildlife managers?. *The Journal of Wildlife Management*, *69*(3), 849-860.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, Waltham, MD.

Mitchell, S. L., Edwards, D. P., Bernard, H., Coomes, D., Jucker, T., Davies, Z. G., & Struebig, M. J. (2018). Riparian reserves help protect forest bird communities in oil palm dominated landscapes. *Journal of Applied Ecology*, *55*(6), 2744-2755.

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., ... & Socher, S. A. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution, 4*(18), 3514-3524.

Naeem, S. (1998). Species redundancy and ecosystem reliability. *Conservation Biology, 12*, 39–45.

Nichols, J. D., Boulinier, T., Hines, J. E., Pollock, K. H., & Sauer, J. R. (1998). Inference methods for spatial variation in species richness and community composition when not all species are detected. *Conservation Biology*, *12*(6), 1390-1398.

Olea, P.P. & Mateo-Tomas, P. (2011) Spatially explicit estimation of occupancy, detection probability and survey effort needed to inform conservation planning. *Diversity and Distributions, 17*, 714–724.

Oldeland, J., Dreber, N., & Wesuls, D. (2010). Diversity measures in comparative rangeland studies: application and advantages of species abundance distributions and diversity profiles. *Dinteria, 31*, 50-66.

Olden, J. D., Poff, N. L., Douglas, M. R., Douglas, M. E., & Fausch, K. D. (2004). Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology & Evolution*, *19*(1), 18-24.

Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association, 77*(379), 548-561.

Patil, G. P. (2014). Diversity Profiles. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels).

Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 20–22.

Plummer, M (2019). rjags: Bayesian Graphical Models using MCMC. R package version 4-9.

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

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature, 405*(6783), 212.

R Core Team (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, A. S., & Gaston, K. J. (2002). Maximising phylogenetic diversity in the selection of networks of conservation areas. *Biological Conservation*, *105*(1), 103-111

Rosenfeld, J. S. (2002). Functional redundancy in ecology and conservation. *Oikos*, *98*(1), 156-162

Royle, J.A. & Dorazio, R.M. (2008) Hierarchical modeling and inference in ecology. Academic Press, London, UK.

Royle, J.A. & Nichols, J.D. (2003) Estimating abundance from repeated presence-absence data or point counts. *Ecology, 84*, 777–790.

Russell, G. J., Brooks, T. M., McKinney, M. M., & Anderson, C. G. (1998). Present and future taxonomic selectivity in bird and mammal extinctions. *Conservation Biology*, *12*(6), 1365-1376.

Sollmann, R., Mohamed, A., Samejima, H. & Wilting, A. (2013) Risky business or simple solution–Relative abundance indices from camera-trapping. *Biological Conservation, 159*, 405–412.

Sollmann, R., Mohamed, A., Niedballa, J., Bender, J., Ambu, L., Lagan, P., ... & Wilting, A. (2017). Quantifying mammal biodiversity co-benefits in certified tropical forests. *Diversity and Distributions, 23*(3), 317-328.

Sutherland, C., Brambilla, M., Pedrini, P., & Tenan, S. (2016). A multiregion community model for inference about geographic variation in species richness. *Methods in Ecology and Evolution*, *7*(7), 783-791.)

Tilker, A., Abrams, J. F., Nguyen, A., Hörig, L., Axtner, J., Louvrier, J., ... & Le, M. (2020). Identifying conservation priorities in a defaunated tropical biodiversity hotspot. *Diversity and Distributions.*

Tilman, D., Isbell, F., & Cowles, J. M. (2014). Biodiversity and ecosystem functioning. *Annual review of ecology, evolution, and systematics, 45*, 471-493.

Zipkin, E. F., Royle, J. A., Dawson, D. K., & Bates, S. (2010). Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation, 143*(2), 479-484.

Zipkin, E. F., Grant, E. H. C., & Fagan, W. F. (2012). Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. *Ecological Applications*, *22*(7), 1962-1972.

Figures

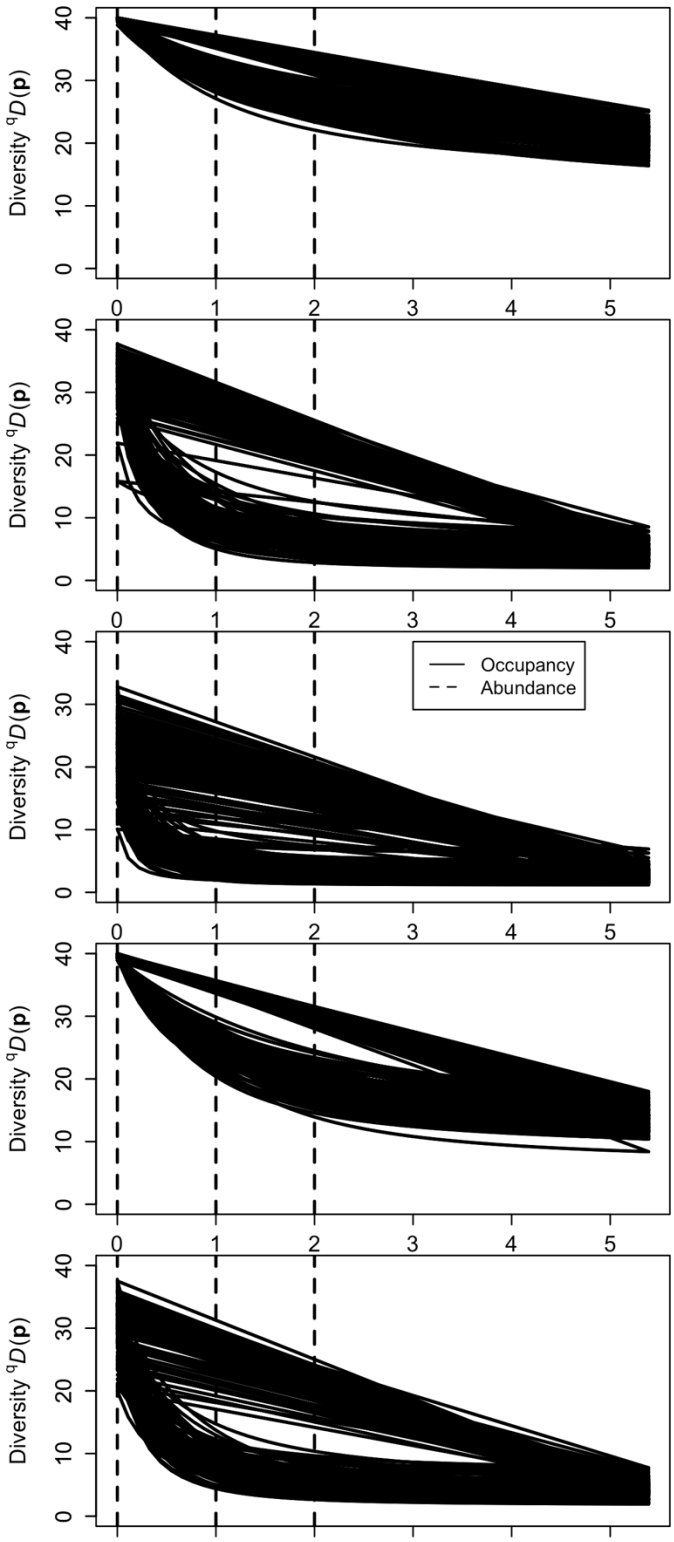

Figure 1. *Simulated occupancy-based profiles for 100 community simulations.*

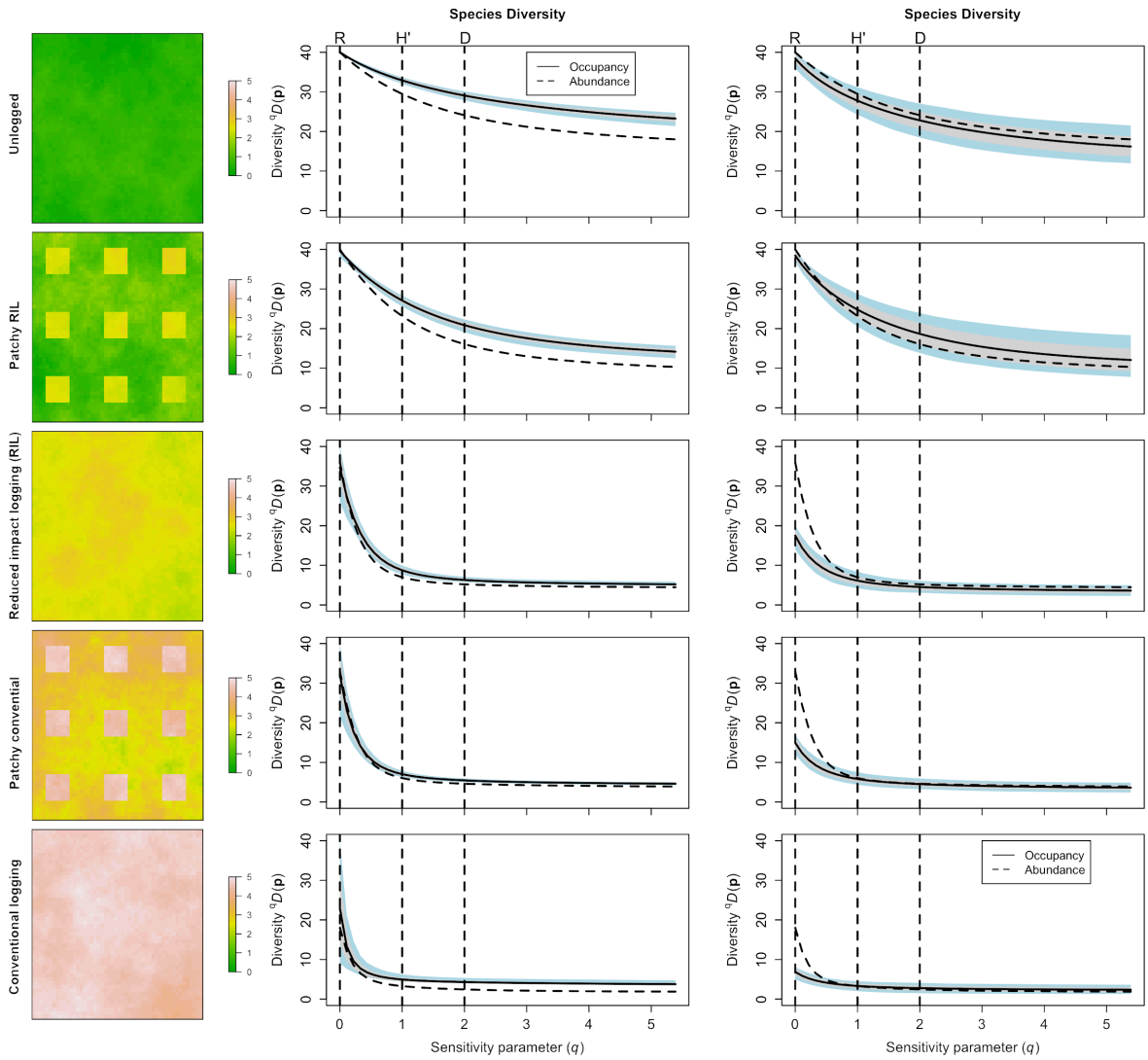

Figure 2. *Results from the simulation study.* ***(A)*** *Simulated "forest quality" habitat covariate for five virtual landscapes. Diversity profiles for simulated data generated using the community occupancy predictions for the entire landscape* ***(B)*** *without thresholding and* ***(C)*** *with thresholding using the* max_{SSS} *method for the entire study landscape (solid line, SD grey shading, 95% CI blue shading) compared to the diversity profiles generated using the true abundance across the whole landscape (dashed line).*

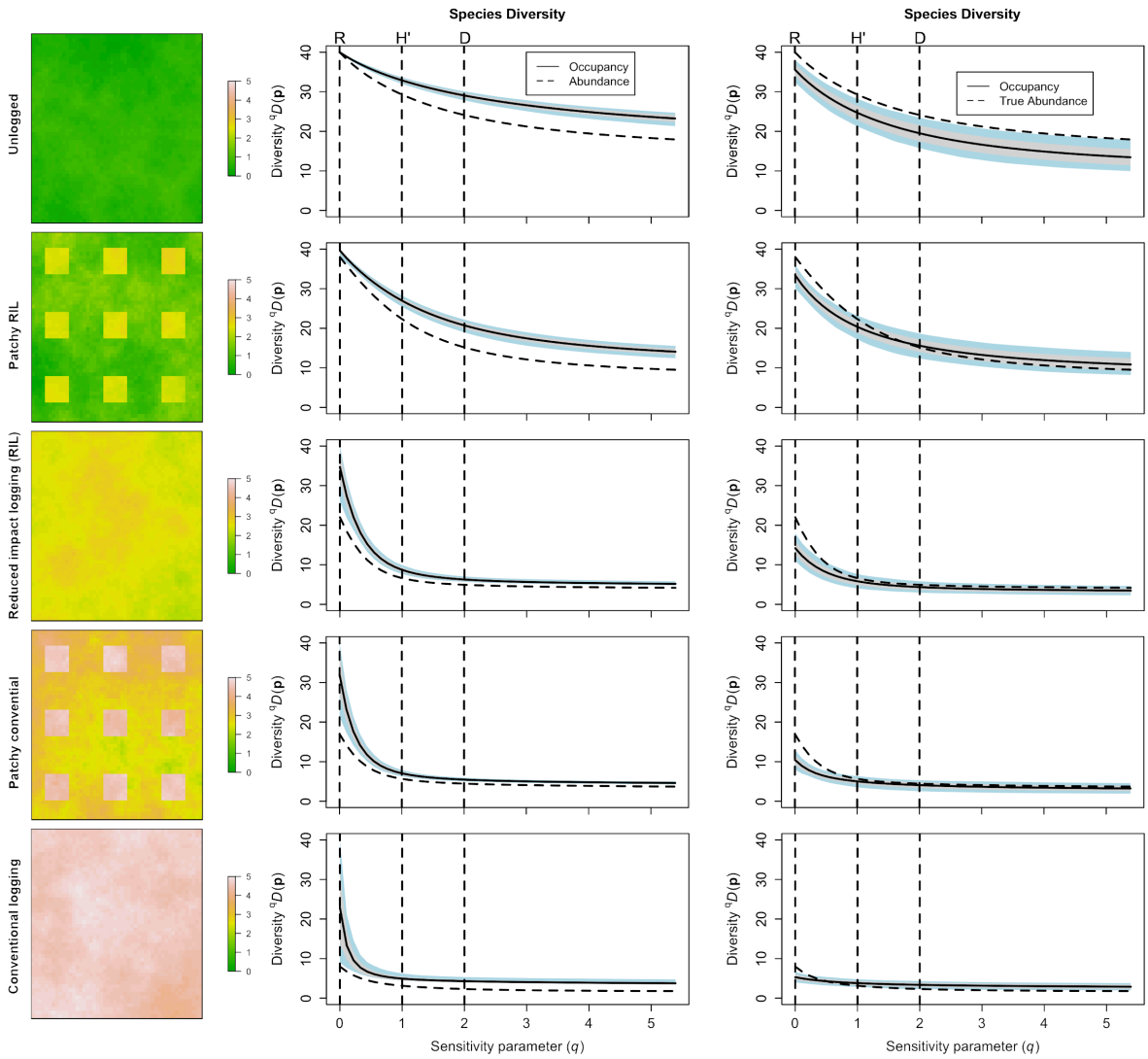

Figure 3. *Results from the simulation study.* ***(A)*** *Simulated "forest quality" habitat covariate for five virtual landscapes. Diversity profiles for simulated data generated using the community occupancy predictions for only the 100 surveyed stations* ***(B)*** *without thresholding and* ***(C)*** *with thresholding using the* max_{SSS} *method for the entire study landscape (solid line, SD grey shading, 95% CI blue shading) compared to the diversity profiles generated using the true abundance at the 100 sample stations in each landscape (dashed black line).*

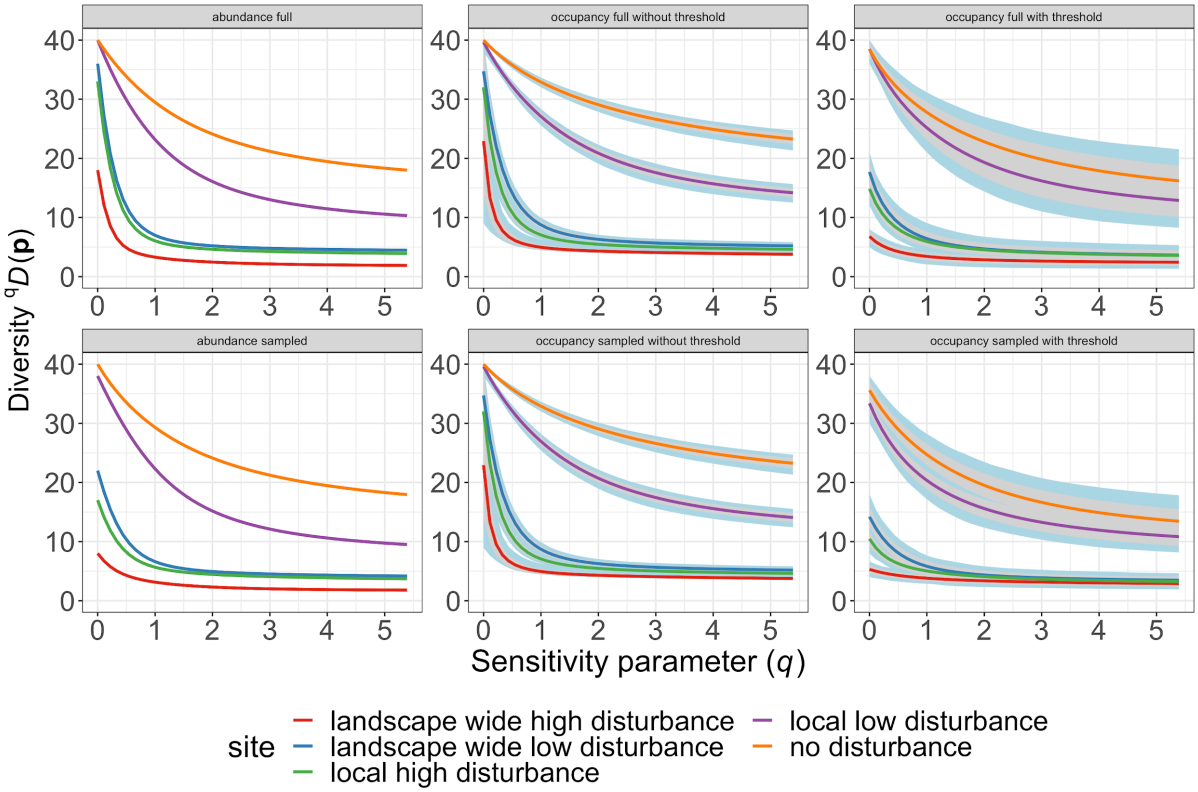

Figure 4. *Comparison of diversity profiles generated using* ***(A)*** *the true abundance across the whole landscape,* ***(B)*** *the true abundance at the 100 sample stations in each landscape,* ***(C)*** *occupancy based predictions across the whole landscape without thresholding,* ***(D)*** *occupancy based predictions at the 100 sample stations in each landscape without thresholding,* ***(E)*** *occupancy based predictions across the whole landscape with thresholding,* ***(F)*** *occupancy based predictions at the 100 sample stations in each landscape with thresholding. The occupancy-based diversity profiles are derived quantities from the Bayesian models allowing for the calculation of uncertainty as standard deviations (grey shading) and 95% credible intervals (blue shading).*

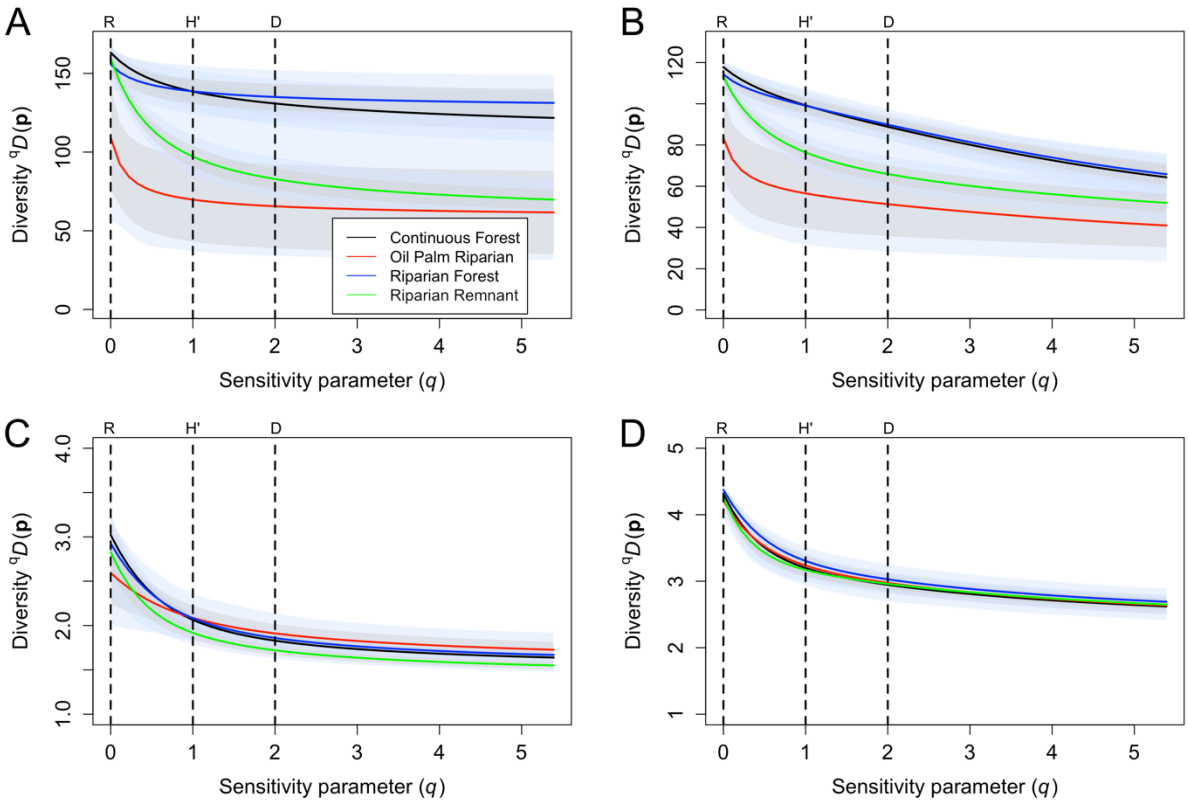

Figure 5. *Diversity profiles for the four habitat types in the case study calculated using the occupancy predictions without thresholding. (A) naive similarity matrix, (B) taxonomic similarity, (C) diet similarity, (D) phylogenetic similarity. The standard deviations (grey shading) and 95% credible intervals (blue shading) are included to quantify uncertainty.*