Abstract

Integrated distribution models (IDMs), in which datasets with different properties are analysed together, are becoming widely used to model species distributions and abundance in space and time. To date, the IDM literature has focused on technical and statistical issues, such as the precision of parameter estimates and mitigation of biases arising from unstructured data sources. However, IDMs have an unrealised potential to estimate ecological properties that could not be derived from the source datasets if analysed separately. We present a model that estimates community alpha diversity metrics by integrating one species-level dataset of presence-absence records with a co-located dataset of group-level counts (i.e. lacking information about species identity). We illustrate the ability of IDMs to capture the true community alpha diversity through simulation studies and apply the model to data from the UK Pollinator Monitoring Scheme, to describe spatial variation in the diversity of solitary bees, bumblebees and hoverflies. The simulation and case studies showed that the proposed IDM produced more precise estimates of the community diversity than the single models, and the analysis of the real dataset further showed that the alpha diversity estimates from the IDM were averages of the single models. Our findings also revealed that IDMs had a higher prediction accuracy for all the insect groups in most cases, with this performance linked to the information provided by a data source into the IDM.

Key words: Alpha diversity, Bayesian models, Markov Chain Monte Carlo methods, Multispecies distribution models, UK Pollinator Monitoring scheme

1 Introduction

Biodiversity monitoring programs generate disparate data types that are used to infer and make predictions about species distributions, dynamics and diversity (Kéry and Royle, 2015, 2020; Isaac et al., 2020; Bird et al., 2014). From the various datasets available, there is now a plethora of modelling approaches to deal with various aspects of the ecological and observational processes in response to the availability of large and varied data from different sources and survey and sampling protocols. The vast majority of these modelling approaches were developed with one particular data type in mind, such as count data or presence-only records. In recent years, the growing heterogeneity of data types has made integrated distribution models (IDMs) an emerging development in ecological statistics and species distribution modelling (Pacifici et al., 2017; Koshkina et al., 2017; Miller et al., 2019; Isaac et al., 2020). IDMs involve integrating datasets of different

types into one model that explicitly captures the features of each.

There are different approaches to developing an integrated model. One data set can be used as a fixed effect to model the other (covariate structure), or datasets can share information through their correlation in space and possibly time (correlation structure; Pacifici et al., 2017; Miller et al., 2019). The most common implementation of IDMs uses a joint likelihood in a hierarchical Bayesian framework (Miller et al., 2019). In the joint likelihood framework, each dataset is conceptualised as an independent realisation of the same underlying ecological state variables (e.g. abundance or occupancy). The strength of the joint likelihood approach comes from sharing information between datasets through common parameters and/or by sampling the same locations in multiple datasets.

Most studies on IDMs are either case studies of particular applications (Doser et al., 2022) or explorations of statistical challenges that data integration brings (Simmonds et al., 2020; Ahmad Suhaimi et al., 2021). Typically, these have addressed the degree to which spatial biases in unstructured data can be overcome, and the precision of the parameters being estimated (Simmonds et al., 2020; Ahmad Suhaimi et al., 2021; Koshkina et al., 2017). These studies have shown, either by using the model predictive accuracy and/or the accuracy and precision of estimated parameters, that IDMs can, in some circumstances, perform better than single models or models developed from a subset of all the datasets (Koshkina et al., 2017; Pacifici et al., 2017; Miller et al., 2019; Isaac et al., 2020; Simmonds et al., 2020; Zulian et al., 2021).

An unrealised benefit of IDMs is the potential to estimate parameters that would not be estimable from either of the data sets if analysed separately. Usually, community alpha diversity measures such as Shannon and Simpson indices are estimated using abundance-based diversity metrics and these indices need species-level abundance information (Hill, 1973; Gatti et al., 2020). However, it is not always possible to identify individuals to their species level. This is often true for insect monitoring, where counts may be resolved to a coarser taxonomic level. This can arise for a number of reasons, such as: the cryptic nature of some species (requiring microscopic examination to separate similar species), the need for specialised taxonomy skills and organisms being observed only briefly (e.g. on the wing).

The vast majority of biodiversity data available, such as presence-absence, capture-recapture, and presence-only data, do not contain information on abundance but may have information on the species identity. Alpha diversity indices can be estimated from the presence-absence data when imperfect detection has been accounted for in a multi-species occupancy model, as has been done in some studies (Gotelli and Chao, 2013; Broms

et al., 2015; Guillera-Arroita et al., 2019). These species-level presence-absence data, however, are less informative than count data (Broms et al., 2015), and the diversity indices estimated can be strongly affected by the model structure such as parametric assumptions, prior specifications and prior choices (Guillera-Arroita et al., 2019).

In this study, we combine these two data types in an IDM to estimate community alpha diversity parameters that could not be "properly" estimated from the datasets when analysed separately. To date, no studies we are aware of have attempted to demonstrate this potential from IDMs, but it is something that integrated population models (IPMs) have been used for for a long time (Besbeas et al., 2002; Abadi et al., 2010; Schaub et al., 2007). For example, Besbeas et al. (2002) integrated census data (providing information about the total number of organisms) and ring recovery data (providing information on individual organisms) to estimate birth, death and fecundity at the population level.

Our model is parameterised using data from the UK Pollinator Monitoring Scheme (PoMS) (O'Connor et al., 2019; Breeze et al., 2021), which has been generating monitoring data on pollinating insects in the UK for the last five years and is now informing an EU-wide pollinator monitoring scheme (Potts et al., 2020). PoMS collects two types of data: one dataset contains presence-absence data on individual species (using pan traps), and the other contains counts that are not resolved to the species level (so-called Flower-Insect Timed Counts or "FIT Counts"). Our analyses of PoMS data are supported by simulations. We demonstrate that between them, these datasets can provide inferences about site-level alpha diversity that would not be possible using either dataset in isolation. The model developed here will be useful in situations where professional and mass participation schemes collect data on the same organisms and where the species are difficult to identify (e.g. most insect groups).

2 Methods

We first provide a motivation for the methods of this study by exploring the PoMS data. We then describe the overall structure of the data. We define the state models representing each species' unknown site-specific abundance and occupancy. The state variables are defined in terms of spatial point processes, which provide a flexible way to integrate datasets in different ecological currencies (Miller et al., 2019; Isaac et al., 2020). We then define sub-models for each of the two data types. The final two sections of the Methods deal with

inference and the estimation of community parameters.

2.1 UK Pollinator Monitoring Scheme Data

The data used is a subset of the PoMS data (Breeze et al., 2021; O'Connor et al., 2019; Scheme, 2022a,b). PoMS implements a systematic survey with 95 1 km square sites selected following a stratified sampling design across Great Britain (GB) and Northern Ireland (NI). The sites are surveyed up to four times per year from May to September, with a minimum of two weeks between each consecutive survey at a site. On the same visit, the observer implements two survey protocols: a pan trap survey and a FIT count survey (Breeze et al., 2021; O'Connor et al., 2019). On each visit, five pan trap stations (each hosting three coloured bowls painted UV-bright yellow, blue and white, mounted at vegetation height and filled with water) are set out along a diagonal of each 1 km square site and left for six hours. During this time, the surveyor undertakes at least two ten-minute FIT counts, which involves counting all insects landing on a target flower in a 50x50 cm patch. Pollinators are identified at the level of a broad taxonomic group, e.g. bumblebees, solitary bees, and hoverflies. After six hours, the samples from the pan traps are collected and sent to a lab for professional identification. Therefore, each visit to the 1km site produces a list of bee and hoverfly species found in the pan traps and group-level count data from the FIT counts. The data used in this study were from the first two years (2017 - 2018) of PoMS, during which 74 of the 75 survey sites across GB returned suitable data (PoMS was not active in NI in the first two years). The summary of the group-level count data and species occupancy data are presented in Table 1 and the distribution of each dataset at each study site is provided in Supplementary information 1 Figures S1-1 to S1-6.

	FIT counts	Pantrap occupancy	
Insect group	Average (SD)	$N_{species}$	Naive occupancy (SD)
Bumblebees	1.11 (5.18)	17	0.086 (0.20)
Hoverflies	2.74 (10.92)	79	0.055 (0.01)
Solitary bees	0.29 (1.62)	70	0.027 (0.111)

Table 1: Summaries of the group-level FIT count and species-level pan trap occupancy data. Both datasets were collected from 74 survey sites (N_{sites}) with 8 survey visits (N_{visit}) (four visits in each year 2017 and 2018). The average FIT counts and their standard deviation (in brackets) were calculated from the group-level FIT count data across all sites and visits and the average naive occupancy from the pantrap occupancy data across all species and sites. The number of species ($N_{species}$) in the species list for each insect group is also provided in the summary.

The above monitoring protocols generate two types of data, each collected at the same set of R indexed locations during replicated T number of visits at each site. One dataset comprises detection-nondetection data at the species level (henceforth 'species occupancy data'); the other is a count across all S species in the taxonomic group ('group count data').

2.2 State variables

We model species abundance as a spatial point process, in which the intensity of that point process determines the expected number of organisms per unit area. Let λ_{ij} be a latent variable describing the intensity of species j at location i and ψ_{ij} be the probability that species j occupies location i. We consider two ways by which the two latent variables can be linked in the IDM using the joint likelihood approach (Pacifici et al., 2017): the 'shared' and the 'covariate' formulation.

2.2.1 Shared formulation

The intensity of each species is linked to the occupancy probability using the complementary log-log link function, which defines the probability that at least one organism is present (Kéry and Royle, 2015):

$$\begin{aligned} log(\lambda_{ij}) = cloglog(\psi_{ij}) = \beta_{0j} + \beta_{1j} \times latitude; \\ \beta_{0j} \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}); \quad \beta_{1j} \sim N(\mu_{\beta_1}, \sigma^2_{\beta_1}), \end{aligned} \tag{1}$$

where β_{0j} the intercept for the species occupancy was normally distributed with mean μ_{β_0} and variance $\sigma^2_{\beta_0}$, β_{1j} the latitudinal gradient slope for species j was normally distributed with mean μ_{β_1} and variance $\sigma^2_{\beta_1}$. The hyperparameters of the intercept (μ_{β_0} and $\sigma^2_{\beta_0}$) and latitude effect (μ_{β_1} and $\sigma^2_{\beta_1}$) represent the community-level mean and variance parameters respectively in the IDM.

In this model structure, all the parameters are shared by the two latent states. Hence each dataset directly informs the latent state and both datasets provide equal weights to the joint likelihood of the IDM (Miller et al., 2019).

2.2.2 Covariate formulation

Although both FIT counts and pantrap surveys were performed by the same observer on the same survey visit, it makes sense for both latent variables to share covariates but allow each dataset to have separate intercepts because they have different survey protocols. The separate intercepts help model the average abundance and occupancy observation difference. Our use of the term "covariate formulation" differs slightly from previous studies (Pacifici et al., 2017; Miller et al., 2019), in which the term refers to using one dataset as a fixed effect in modelling the second dataset. In the classification of Miller et al. (2019), our covariate model is a form of joint-likelihood structure, but the term is useful to distinguish it from the shared formulation above.

The link function for the latent variables in this IDM framework becomes:

$$\begin{aligned} cloglog(\psi_{ij}) &= \beta_{0j} + \beta_{1j} \times latitude; \\ log(\lambda_{ij}) &= \omega_0 + \beta_{1j} \times latitude; \\ \beta_{0j} &\sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}); \\ \beta_{1j} &\sim N(\mu_{\beta_1}, \sigma^2_{\beta_1}); \\ \omega_0 &\sim N(\mu_{\omega_0}, \sigma^2_{\omega_0}), \end{aligned} \tag{2}$$

where β_{0j} the intercept for the species occupancy is normally distributed with mean μ_{β_0} and variance $\sigma^2_{\beta_0}$, ω_0 the intercept of the group counts is normally distributed with mean μ_{ω_0} and variance $\sigma^2_{\omega_0}$, β_{1j} the slope of the latitudinal gradient for species j is normally distributed with mean μ_{β_1} and variance $\sigma^2_{\beta_1}$. As described for the shared model (section 2.2.1), the intercept and covariate effect hyperparameters represent their respective community-level mean and variance parameters in the models.

The covariate structure allows both state variable models in the IDM to share important parameters but preserve their average abundance when no covariate effects exist. Moreover, the quality of both datasets determines how well the parameters are estimated (Pacifici et al., 2017) since some parameters will be estimated using each dataset, and the parameters that are not shared serve as (unequal) weights for the contribution from each dataset.

2.3 Sub-models for each dataset

Having defined the latent state variables λ_{ij} and ψ_{ij} and the possible ways they can be linked together, the sub-models for species occupancy data (section 2.3.1) and group-level count data (section 2.3.2) can be defined. In addition, the model can estimate various community ecology metrics (see section 2.5) as derived parameters. Based on preliminary analysis (Supplementary information 1), we used a negative binomial model with an intercept and covariate to fit the group count data. We also used logistic regression with intercept, covariate effect, site and visit random effect to fit the species occupancy data.

2.3.1 Sub-models for species occupancy data

We model the species occupancy data with an occupancy-detection model (MacKenzie et al., 2002). The true ecological state (true presence or absence denoted as z in this study) for species j at site i is modelled with a Bernoulli distribution with probability ψ_{ij}, where ψ_{ij} was the probability of species j occupying site i as defined by equations (1) and (2).

The detection probability (p_{ijk}) for species j at site i during the survey visit k is modelled with a site, species and visit random effect logistic regression using the logit link. That is:

$$\begin{aligned} logit(p_{ijk}) &= \zeta_i + \nu_j + \rho_k; \\ \zeta_i \sim N(0, \sigma_\zeta^2) \quad \text{and} \quad \nu_j \sim N(0, \sigma_\nu^2) \quad \rho_k \sim N(0, \sigma_\rho^2), \end{aligned} \tag{3}$$

where ζ_i, the effect of site i, is normally distributed with zero mean and variance σ_ζ^2; ν_j, the effect of species j, is normally distributed with zero mean and variance σ_ν^2; and ρ_k, the effect of survey visit k, is normally distributed with mean 0 and variance σ_ρ^2. We model the visit effect in the detection process to account for the significant visit effect found during the exploration phase for the species occupancy data (Supplementary information 1). By the definition of our model for the detection probability in equation (3), the average detection probability for a species in any given site is 0.5, which allows all species in the taxonomic group to have an equal chance of being detected or not detected on average.

Let the observation for species j during the k^{th} visit to location i be represented by X_{ijk}, for $i = 1, 2, \ldots, R$ indexed sites and $j = 1, 2, \ldots, S$ species. This observation, over the five pantrap replicates at each

site, is Binomially distributed with probability $z_{ij} \times p_{ijk}$, where p_{ijk} is the detection probability and z_{ij} is the true state of species j at site i (that is, $X_{ijk} \sim \text{Binomial}(5, z_{ij} \times p_{ijk})$).

2.3.2 Sub-model for group count data

Having defined the intensity of species j at location i (equations (1) and (2)), the intensity for the group counts will be a sum of all the intensities of the species that make up that taxonomic level. This is because we assume the group counts are made up of all the species in the pantrap data, and the sum of realisations from Poisson point processes is also a Poisson point process with an intensity equal to the sum of the intensities of the individual components (Harremoës, 2001; Jacod, 1975).

Let Y_{ik} be the observed count of individuals on the kth survey (across all species in the group). We modelled the counts with a negative binomial distribution (to allow for extra variation in the count data) with parameters θ and $\gamma = \frac{\theta}{\theta+\lambda_i^g}$, where $\lambda_i^g = \sum_j^S \lambda_{ij}$ is the intensity of the group counts at site i (that is, $Y_{ik} \sim \text{NB}(\theta, \gamma)$). The parameter θ is the overdispersion parameter, which allows us to model the extra variation in the group count data. Note that as $\theta \to \infty$, the negative binomial distribution converges upon the Poisson distribution.

We present the various joint likelihood structures defined in section (2.2) and the sub-models for each PoMS dataset defined in section (2.3) in Figure 1.

2.4 Community Diversity Indices

The community alpha diversity was estimated using the Shannon-Wiener diversity index. This is the most commonly used index from the Hills indices (Hill, 1973), and it places equal weights on rare and dominant species. We acknowledge that the Shannon-Wiener diversity index may have some limitations (Lande, 1996; Morishita, 1996; Itô, 2007; Gatti et al., 2020; Chao and Jost, 2015; O'hara, 2005), in which case other indices such as Simpson index may be preferable. However, we use the Shannon-Wiener diversity index to show how alpha diversity can be estimated using our proposed IDM and how all the models capture the true alpha diversity. For real-world applications, we urge caution about the choice of the index. Moreover, the Hills indices are all functions of the relative abundance proportion, so the method developed for one index can be

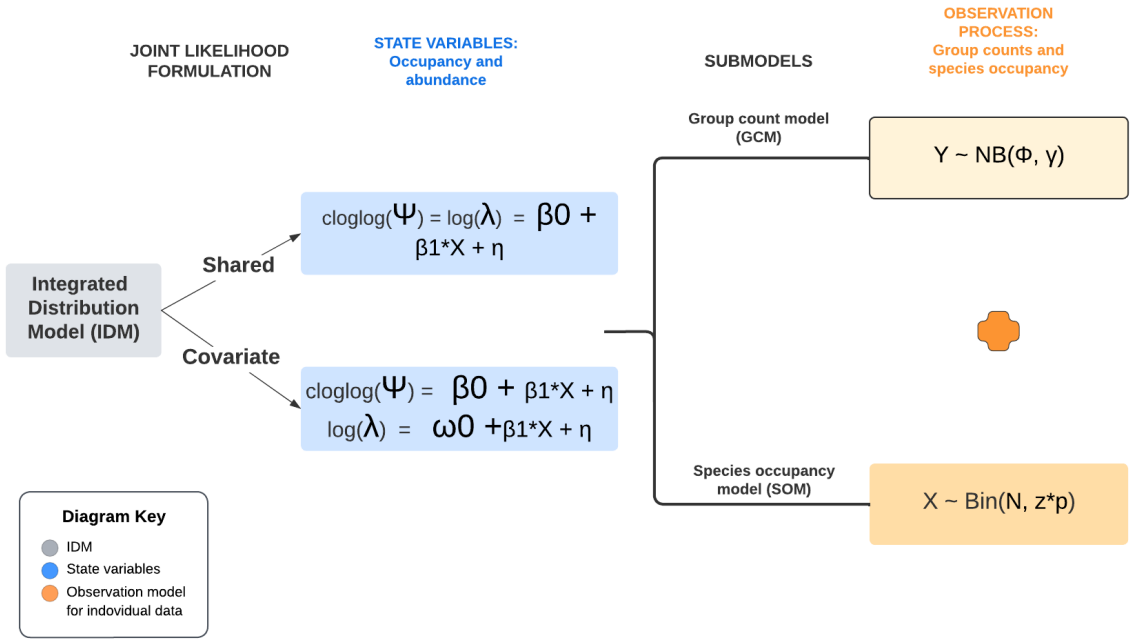

Figure 1: Flowchart showing the integrated distribution model process for the PoMS survey data. To fit the IDM, we used two joint likelihood links in this study: shared formulation (section 2.2.1) and covariate formulation (section 2.2.2). The state variables models are defined by equations (1) and (2). The sub-models for each dataset: the group count model (GCM) and species occupancy model (SOM) have been defined as the observation process of abundance and occupancy respectively (sections 2.3.2 and 2.3.1 respectively). The IDM combines the SOM and GCM for each joint likelihood link used. All the parameters in this flowchart are defined in sections 2.2 and 2.3. In addition to the intercept and covariate effect, we add a species interaction effect η in this flowchart.

easily extended to the others. The Shannon-Wiener diversity index was calculated as:

$$H^1 = -\sum_{j=1}^{S} r_{ij} log(r_{ij}), \tag{4}$$

where $r_{ij} = \frac{\lambda_{ij}}{\sum_{j=1}^{S} \lambda_{ij}}$ is the relative abundance of a species j at location i.

2.5 Evaluating model performance

We fitted five models to the PoMS survey data in this study: IDM and group count model (GCM) with the shared formulation of the joint likelihood defined in section 2.2.1 which we will refer to as **IDMSH** and **GCMSH** respectively, the IDM and GCM with covariate formulation of the joint likelihood defined in section

2.2.2 which we will refer to as **IDMCO** and **GCMCO** respectively, and the species occupancy model (**SOM**). The two IDMs (IDMSH and IDMCO) fitted a joint likelihood model with both the occupancy and group count data; the two GCMs (GCMSH and GCMCO) fitted the negative binomial regression model to the group count data with the two joint likelihood formulations, and SOM fitted the sub-model for the species occupancy data (section 2.4.1). We note that the alpha diversity estimates from the GCMs are strongly driven by the priors assigned to the parameters of λ_{ij}. In the absence of information in the data to contradict this prior, we anticipate that GCMs will estimate the local alpha diversity very poorly. We recognise that this model is not something that community ecologists would choose to fit, but the comparison with other models is informative. These models are summarised in Table 2.

Model	Model description	Type	Data used	Predictor
IDMSH	IDM with shared structure defined in equation (1)	Integrated	GC and SO	$\beta_{0j} + \beta_{1j} \times lat_i$
IDMCO	IDM with shared structure defined in equation (2)			$\beta_{0j} + \beta_{1j} \times lat_i$ $\omega_0 + \beta_{1j} \times lat_i$
GCMSH	GCM with shared structure defined in equation (1)	Single	GC	$\beta_{0j} + \beta_{1j} \times lat_i$
GCMCO	GCM with shared structure defined in equation (2)			$\omega_0 + \beta_{1j} \times lat_i$
SOM	species occupancy model	Single	SO	$\beta_{0j} + \beta_{1j} \times lat_i$

Table 2: Models fitted in this study, their descriptions, predictors, type and data used to fit them. Two integrated models: IDMSH and IDMCO, and three single models: GCMSH, GCMCO and SOM, are fitted. The data used are from the UK PoMS survey: FIT counts (GC) and Pantrap species occupancy (SO) data. The cloglog link was used for the occupancy model and the log link was used for the group count model. The definitions of the parameters used in the predictor column are described in section (2.3), with lat referring to the latitudinal gradient slope.

2.5.1 Fitting the models

We fitted the models in a Bayesian framework. We obtained samples of the parameters using the Markov chain Monte Carlo (MCMC) approach and estimated posterior summaries of model parameters using the NIMBLE package (de Valpine et al., 2017) in R (R Core Team, 2022). We chose a normal distribution with zero mean and variance of 100 as the prior for the mean hyperparameters of the state variables and an inverse gamma distribution with scale parameter 2 and shape parameter 1 as the prior distribution for the variance

hyperparameters. We ran three chains with 300000 iterations for all the models, and 200000 were discarded as burn-in samples. We keep a twentieth of the left-over samples to reduce the hard disk space used by our analysis. The convergence of the fitted model was checked by estimating Gelman-Rubin R-hat statistic (Brooks and Gelman, 1998) using the ggmcmc package (Fernández-i Marín, 2013) and rejected the models with R-hat greater than 1.1.

2.5.2 Simulation study

We performed simulation studies to assess which of the five models (described in Table 2) better estimated the true alpha diversity. We simulated 30 data replicates using the IDM framework for each latent variable formulation used in this study (i.e. using both the covariate and the shared formulation). We used the same number of sites and visits from the PoMS surveys but used 20 species for the simulations due to computational expensiveness in running the models for more species.

The true values for the hyperparameters defined in sections 2.2 were chosen as follows: $\mu_{\beta_0} = 0$, $\sigma_{\beta_0} = 0.2$, $\mu_{\beta_1} = -2$, $\sigma_{\beta_1} = 1$, $\sigma_{\omega_0} = 0.2$, $\sigma_\zeta = 0.3$, $\sigma_\nu = 1$ and $\sigma_\rho = 2$. We also randomly selected 25 sites for each visit in the group count and occupancy model and assigned them NAs to reflect missing species identifications and group counts in the PoMS data.

We fitted the five study models defined in Table 2 to the 30 simulated datasets for each joint likelihood formulation. By this, we employed a cross-design to ascertain the effect of fitting a wrong model in this study. For example, when the IDMCO or GCMCO is fitted to the dataset simulated under the shared formulation, we can infer the effect of fitting a covariate-formulated model ("wrong model") to the dataset. We assessed this effect by estimating the mean bias and precision of Shannon estimates at each site across the replicated datasets. That is, for each site i, we obtain the metrics:

$$\begin{aligned} \text{Mean bias} &= \frac{1}{30}\sum_{k=1}^{30}(\hat{H'}_i^{(k)} - {H'}_i^{(k)}), \\ \text{Mean precision} &= \frac{1}{30}\sum_{k=1}^{30}\frac{1}{SD\left(\hat{H'}_i^{(k)}\right)^2}, \end{aligned} \tag{5}$$

where $\hat{H'}_i^{(k)}$ is the posterior mean of the Shannon index for dataset k, ${H'}_i^{(k)}$ is the true Shannon index obtained from simulating dataset k and $SD(\hat{H'}_i^{(k)})$ is the posterior standard deviation of the Shannon index for dataset

k. The fitted models with mean bias closer to 0 and the highest mean precision are indicated to perform best.

2.5.3 Model Validation and Assessment

Model predictive performance

We performed two-fold cross-validation to ascertain the model's ability to predict new data. The same folds are used for all the models under study, and the log pointwise predictive density (Nicenboim et al., 2021; Gelman et al., 2014) was used to measure the cross-validation's predictive accuracy. The log pointwise predictive density (lppd) is defined as:

$$lppd = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N} log(P(y_{n,k}|y_{-n,k}, \hat{\theta}, M)) \tag{6}$$

where $log(P(y_{n,k}|y_{-n,k}, \hat{\theta}, M))$ is the log predictive density of the withheld data samples $y_{n,k}$ in fold k under model M, which was trained with data samples $y_{-n,k}$ in fold k to obtain estimated model parameters $\hat{\theta}$ with n being the number of samples in each fold and N is the number of samples of each fold. Larger values of the metric indicate better performance.

It must be pointed out that the withheld samples used in IDMSH and IDMCO have both species occupancy (X) and group count (Y) samples in their training and validation sets. Therefore, we estimated the lppd for the validation samples from the pantrap and group count data separately after we had estimated the model parameters $\hat{\theta}$ with both datasets in the training samples. Since the log predictive density is additive (Gelman et al., 2014), the log predictive density of the integrated model was obtained by summing the lppd of each dataset.

Information provided by each dataset

We also ascertained the information contributed to the IDM by each data type. This was done by comparing the log-likelihoods of the single models (SOM and GCMs) to that of the IDMs (where both single and integrated models being compared share the same joint likelihood structure), following Zulian et al. (2021). Since there are two data types in this study: pantrap data and FIT count data, including a data type that informs the IDMs should lead to better predictions (higher prediction accuracy) of the other data types. For example, by comparing the predictive log-likelihoods of the GCMCO to IDMCO for group count data, one can assess whether the pantrap occupancy data improves the predictive performance of IDMCO on the group count data.

We shall refer to this approach as the 'marginal contribution' of a data type in the rest of this paper. Negative values of the marginal contribution indicate that the data type did not contribute to the IDM.

It must be stated that the marginal contribution can only indicate whether a data type provides information. Due to differences in the data types (occupancy and count data) and sample sizes, it is unfeasible to compare the marginal contribution of the data types to ascertain which one provides the most information.

3 Results

The estimates of the mean bias and precision of Shannon diversity estimates over all 30 replicated simulated datasets are presented in Figure 2. The log-predictive density from the two-fold cross-validation and average Shannon index across all the study sites are summarised in Table 3. The site-specific precision estimates of the Shannon indices for each insect group and the model used to fit the data are presented in Figure 3. All other figures and tables referred to in this section are presented in Supplementary Information 2.

3.1 Simulation study

Figure 2 shows the distribution of the mean bias and precision of the Shannon indices at the 74 sites estimated from the five study models fitted to the simulated data described in section 2.5.2. Whether the data were simulated under the shared or covariate formulations, the mean bias of the Shannon index from the integrated models (irrespective of their joint likelihood structure) was similar to that from SOM, with the median bias around 0 with small variation across sites. This suggested that the integrated models and SOM well captured the true Shannon index across all the study sites.

As expected, the Shannon indices were poorly estimated the GCMs. Alpha diversity was consistently overestimated by the GCMCO model and underestimated by GCMSH (Figure 2), and both models have much lower precision than other models. As explained in the Methods, this is expected because the only information about species identities in these models derives from the priors.

Although the mean bias of the Shannon indices from the integrated models and SOM are similar, the Shannon indices were estimated with higher precision in the integrated models than in SOM (blue bars in Figure 2). When the data was simulated under the covariate formulation (red bars), the precision of Shannon diversity estimates from the integrated models was similar to that of SOM.

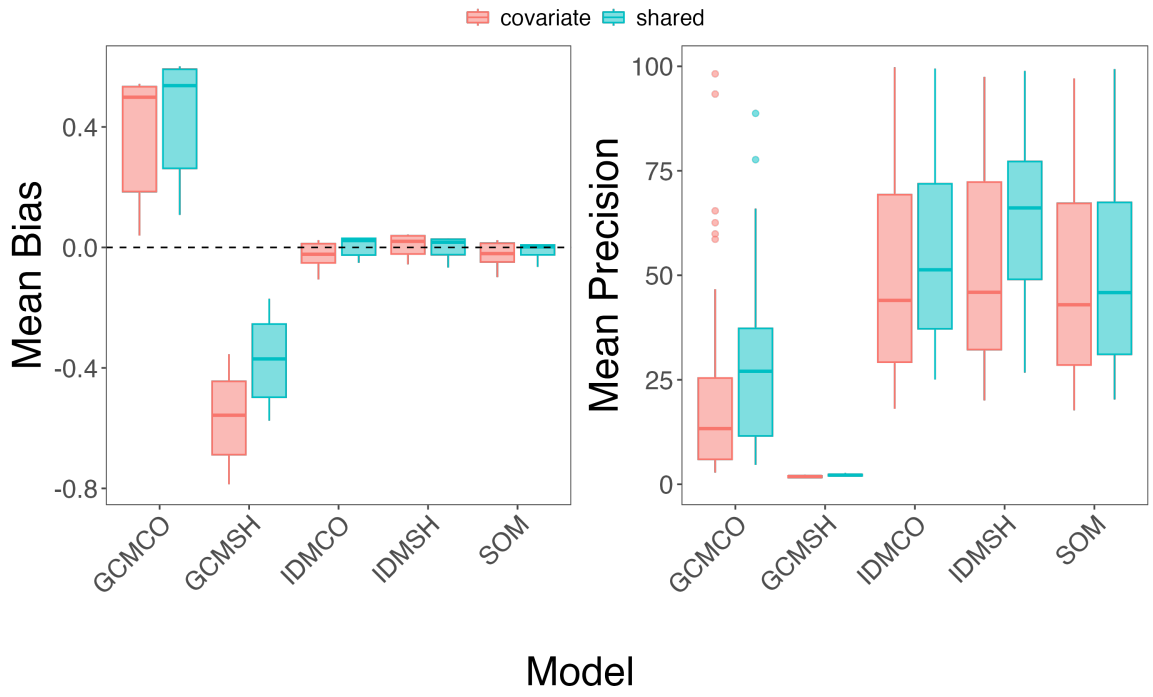

Figure 2: Mean bias and precision of Shannon index from the five study models fitted to the simulated data. The boxplot shows the distribution of the mean bias and precision for the 74 sites. The five models: two integrated models (IDMSH and IDMCO), two group count models (GCMSH and GCMCO) and a species occupancy model (SOM) were fitted to data simulated under the shared structure (coloured in blue) and those simulated under the covariate structure (red).

3.2 Analysis of PoMS dataset

3.2.1 Estimation of Shannon index (H')

Shannon diversity is expected to be higher for communities with a comparatively higher number of species (Roswell et al., 2021) and/or evenness (Nagendra, 2002). It is, therefore, not surprising that the Shannon indices were highest for hoverflies (n = 79 species) and lowest for bumblebees (n = 17 species; Tables 1 and 3).

The estimates of H' from the five models showed consistencies in the estimated diversity pattern for each insect group, as we observe in Table 3 and Supplementary information 2 Figure S2-1. Firstly, there was a negative latitudinal effect on the estimates of the Shannon indices (that is, the Shannon index decreased with the latitudinal gradient; Figure S2-2). The community intercept (μ_{ω_0}) estimated from GCMCO and IDMCO were relatively the same, but the estimated latitudinal and species effect (μ_{β_1} and μ_{β_0} respectively) from the integrated models lies between those estimated from the group count and species occupancy models (Figure

		Shannon index		log predictive density			Marginal Contribution	
Insect group	Model	Mean	SD	All dataset	Pantrap	FIT count	Pantrap	FIT count
Bumblebees	GCMSH	2.79	0.02	-	-	−239.90	-	-
	GCMCO	2.78	0.06	-	-	−290.15	-	-
	SOM	1.08	0.17	-	−3649.25	-	-	-
	IDMSH	2.29	0.07	−3561.26	- 3322.39	−238.90	326.86	1.0
	IDMCO	1.86	0.08	**−3489.85**	**−3250.97**	**− 238.87**	398.28	1.03
Hoverflies	GCMSH	4.34	0.014	-	-	−517.49	-	-
	GCMCO	4.32	0.05	-	-	−513.62	-	-
	SOM	3.98	0.13	-	−35118.55	-	-	-
	IDMSH	4.01	0.08	−36195.10	−35715.74	**−479.36**	38.14	−597
	IDMCO	4.03	0.09	**-32275.90**	**−31716.42**	−559.48	−45.86	**3402**
Solitary bees	GCMSH	4.10	0.08	-	-	**−150.51**	-	-
	GCMCO	4.08	0.18	-	-	−446.05	-	-
	SOM	3.04	0.40	-	- 14372.42	-	-	-
	IDMSH	3.25	0.33	- **11341**	**−11165.97**	- 175.39	**3206.45**	-24.88
	IDMCO	3.27	0.31	−11464.48	−11285.78	−178.70	3086.64	−28.19

Table 3: Log predictive density from the two-fold cross-validation, the marginal contribution of each dataset and the Bayesian p-values from the posterior predictive checks of all the models fitted to the PoMS dataset. For each insect group, the marginal contribution of pantrap data was estimated as $lppd_{SOM} - lppd_{IDMSH}$ and $lppd_{SOM} - lppd_{IDMCO}$; and the marginal contribution of FIT count data was estimated as the $lppd_{GCMSH} - lppd_{IDMSH}$ and $lppd_{GCMCO} - lppd_{IDMCO}$.

S2-2). Additionally, the group count models (GCMSH and GCMCO) had the highest average H' estimates (Table 3), followed by the integrated models (IDMSH and IDMCO) and finally, the species occupancy model (SOM). These observations suggested that the integrated models serve as the average model for the species occupancy and group count models.

We have already established from the simulation study that the priors strongly affect the Shannon indices estimated from the GCM models. Narrowing our observations to the site-specific precision estimates of the Shannon indices from the integrated models and SOM, we observed the estimates from the integrated models were more precise than those from SOM (Figure 3). This precision was also higher for the sites with higher Shannon diversity estimates (compare Figure 3 to Supplementary information 2 Figure S2-1).

3.2.2 Predictive Performance and information provided

Table 3 shows the two-fold cross-validation log-predictive density estimated from the five study models for the three insect groups: bumblebees, solitary bees and hoverflies. For bumblebees, we find that both IDMs outperform the single-dataset models in predicting both the pantrap and FIT count data. For hoverflies, the

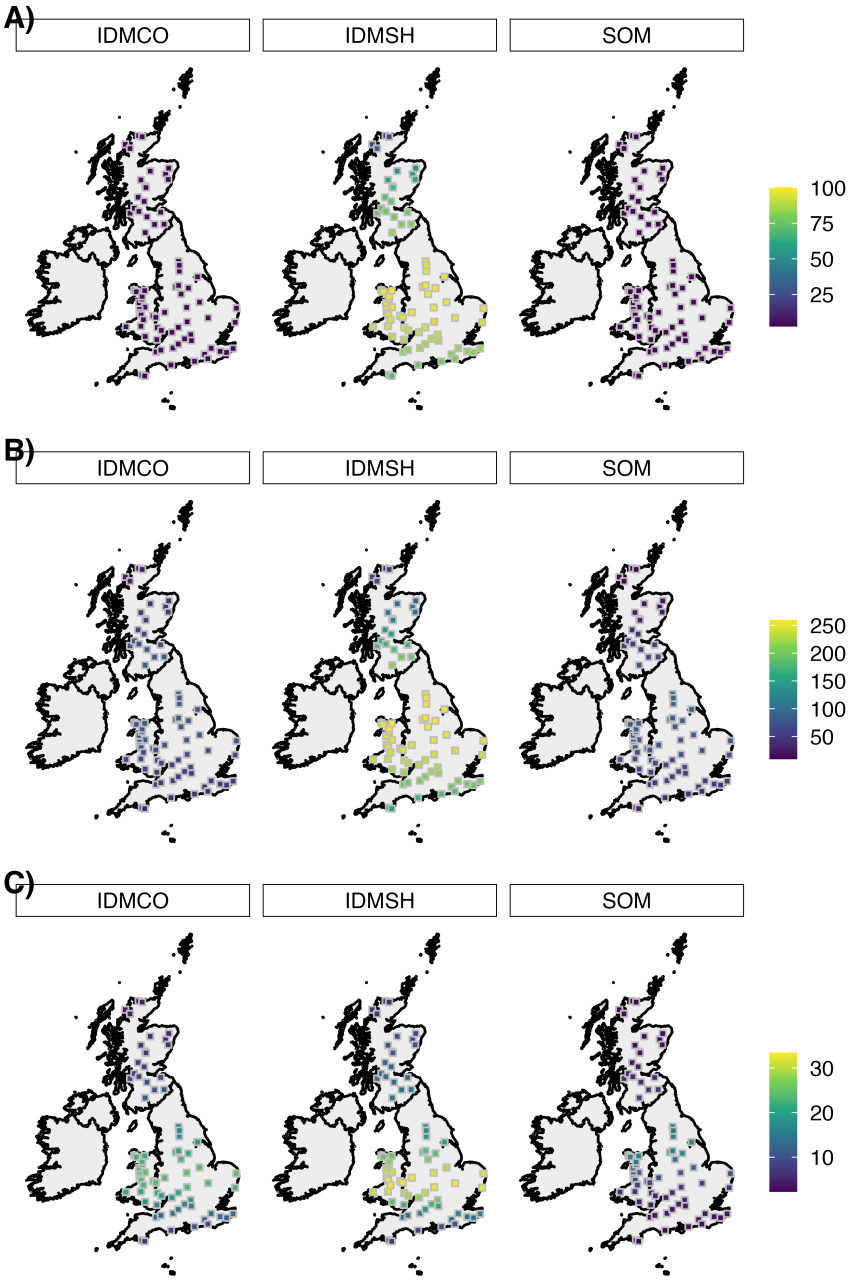

Figure 3: Precision of the Shannon diversity (H') estimates for each of the 74 PoMS sites from the five models in this study summarised in Table (2) for each of the insect groups: A) bumblebees, B) hoverflies and C) solitary bees.

best fitting model in each case is an IDM, with the covariate formulated model (IDMCO) performing best for pantrap data and the shared formulated model performing best for the FIT counts. For solitary bees, we find that both IDMs outperform the SOM for predicting the pantrap data, but that one of the group count models is the best predictor of the FIT count data. These results show that the IDMs outperformed the single models in the prediction accuracy of new data for bumblebees and hoverflies and at least as well for the solitary bees. In other words, the inclusion of FIT Count data from has added information to the models.

Table 3 also shows the marginal contribution of each dataset to the integrated models. For the models fitted to the bumblebees and hoverflies, both pantrap and FIT count data contributed to the IDM (with a positive marginal contribution), but the contribution was higher in IDMCO than IDMSH. For solitary bees, the marginal contribution of the FIT count was negative for both IDMCO and IDMSH (indicating they do not provide information into the IDM) and the marginal contribution of the pantrap data was positive, indicating that the FIT count data did not inform the IDMs. This was expected since the average FIT counts of bumblebees and hoverflies (1.11 ± 5.18 and 2.74 ± 10.92 respectively), were significantly higher than that of the solitary bees (0.29 ± 1.62; Table 1). There was much information from this count data to inform the IDMs to predict the pantrap data better for bumblebees and hoverflies.

4 Discussion

Many data types are available in community ecology, and many modelling techniques are available to analyse data types separately. In some cases, multiple datasets are available that differ in taxonomic resolution. We developed an integrated distribution model that combined data from different taxonomic levels to estimate alpha diversity in a community. Using a combination of simulations and analysis of empirical data, we showed that integrated models can produce useful estimates of community ecology parameters from datasets that lack the information to do so if analysed separately. In addition, the IDMs performed better than the single models in most cases.

Previous studies have shown that IDMs perform better than single models in estimating state variable parameters and prediction accuracy of new datasets (Strebel et al., 2022; Pacifici et al., 2017; Doser et al., 2022; Miller et al., 2019). Although Miller et al. (2019) noted that IDMs present opportunities to model community dynamics and diversity from multiple datasets, our study is the first to implement this by combining data

from different taxonomic levels (species and group level). Our simulation and case studies showed that the IDMs outperform the single models in producing precise alpha diversity estimates in a community if both datasets share information between them (Figure 2, Supplementary information 2 Figure S2-1 and Table 3). The information from each dataset was shared through the joint likelihood framework, and the information sharing process has been noted in the literature to be the benefit of using IDMs (Isaac et al., 2020; Miller et al., 2019).

Furthermore, the proposed IDMs outperform the single models' prediction accuracy of new datasets for some insect groups. From our model assessment of the PoMS data using two-fold cross-validation, IDMs outperformed the single models in predicting new data for all insect groups, except the solitary bees FIT count data (Table 3). The out-performance is evident from the information provided by each dataset into the IDM to inform the estimation of the model parameters directly. This observation is well noted in literature (Zulian et al., 2021). For instance, when modelling the solitary bees dataset, pantrap data did not inform the IDM to predict the FIT count data better, and as such the group count model outperforms the IDMs (Table 3).

In this study, we explored two IDM variants with different joint likelihood formulations. The covariate and shared formulations had very similar performance in terms of predictive performance and alpha diversity estimation but differed in how well they fit the two datasets. The shared joint likelihood structure ensured that all state variables were shared between both datasets. The covariate structure allowed some flexibility in sharing the state model definition by allowing each dataset to have a unique intercept. Previous studies on IDMs, using either covariate and shared joint likelihood structures, have all shown that IDMs have higher prediction accuracy than single models (Koshkina et al., 2017; Zulian et al., 2021; Fletcher Jr et al., 2019; Fletcher et al., 2016; Pacifici et al., 2017; Simmonds et al., 2020; Adde et al., 2021; Ahmad Suhaimi et al., 2021). These methods have been used to model species distributions and turnover using multiple data types from the same taxonomic levels. Just a few of these studies (such as Chevalier et al., 2021) exist that explore various joint likelihood structures for their IDMs. Our study showed that the choice of structure has little effect on the IDM's predictive performance over the single models since all the IDMs outperform the single models, except for solitary bees FIT count data (Table 3). Additionally, the pattern of estimated Shannon diversity and the precision of the estimates was invariant to the choice of the joint likelihood structure (Figures 2 and 3 and Table 3). This indicates the choice of the joint likelihood formulation is inconsequential to the performance of IDMs, and any alternative can be chosen to model alpha diversity.

The UK PoMS protocols are specifically designed to produce datasets with different taxonomic resolutions. New monitoring technologies create many situations in which analysts might encounter datasets that differ in taxonomic resolution. For example, data on the abundance of aquatic macroinvertebrates, such as those collected by kick-sampling for Water Framework Directive reporting, are typically reported at the genus level or higher (Haase et al., 2023). Modern DNA (meta)barcoding makes it possible to identify specimens in these samples to species level, but typically only as presence-absence data (Bohan et al., 2017). Another promising use case is the combination of traditional field surveys with data identified from images using computer vision: algorithms often have low confidence in the species identity, but high confidence in the Genus. Our model provides a ready-made solution for estimating community parameters in such situations.

Other situations might arise in which mixed taxonomic resolution is an unwanted byproduct of the data generation process. A good example would be a citizen science projects where participants differ in their taxonomic skill levels, such that some report counts at species level but others report at a coarser level. Our approach provides a way to use all the data at the resolution at which it was captured. Thus, our proposed model further extends the range of applications for IDMs in ecology and conservation to help researchers and conservationists make the most of available data, in order to provide better evidence and understanding about biodiversity.

Data availability

The code and PoMS dataset used is available on GitHub repository: https://github.com/Peprah94/Integrated-distribution-models-for-different-taxonomic-levels with DOI:10.5281/zenodo.8424494.

Conflict of Interest

The authors declare no conflict of interest.

Author Contribution

All the authors were involved in the idea conception and manuscript writing. KPA, RBO and NI were involved with the methodology development, KPA run the entire analysis and led the writing of the manuscript.

References

Abadi, F., Gimenez, O., Arlettaz, R. and Schaub, M. (2010) An assessment of integrated population models: bias, accuracy, and violation of the assumption of independence. *Ecology*, **91**, 7–14.

Adde, A., Casabona i Amat, C., Mazerolle, M. J., Darveau, M., Cumming, S. G. and O'Hara, R. B. (2021) Integrated modeling of waterfowl distribution in western canada using aerial survey and citizen science (ebird) data. *Ecosphere*, **12**, e03790.

Ahmad Suhaimi, S. S., Blair, G. S. and Jarvis, S. G. (2021) Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions*, **27**, 1066–1075.

Alexander, N., Moyeed, R. and Stander, J. (2000) Spatial modelling of individual-level parasite counts using the negative binomial distribution. *Biostatistics*, **1**, 453–463.

Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global ecology and biogeography*, **19**, 134–143.

Besbeas, P., Freeman, S. N., Morgan, B. J. and Catchpole, E. A. (2002) Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, **58**, 540–547.

Bhattacharya, A. and Dunson, D. B. (2011) Sparse bayesian infinite factor models. *Biometrika*, 291–306.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F. et al. (2014) Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, **173**, 144–154.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J. and Woodward, G. (2017) Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in ecology & evolution*, **32**, 477–487.

Breeze, T. D., Bailey, A. P., Balcombe, K. G., Brereton, T., Comont, R., Edwards, M., Garratt, M. P., Harvey, M., Hawes, C., Isaac, N. et al. (2021) Pollinator monitoring more than pays for itself. *Journal of Applied Ecology*, **58**, 44–57.

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

Brooks, S. P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, **7**, 434–455.

Chambert, T., Rotella, J. J. and Higgs, M. D. (2014) Use of posterior predictive checks as an inferential tool for investigating individual heterogeneity in animal population vital rates. *Ecology and Evolution*, **4**, 1389–1397.

Chao, A. and Jost, L. (2015) Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*, **6**, 873–882.

Chevalier, M., Broennimann, O., Cornuault, J. and Guisan, A. (2021) Data integration methods to account for spatial niche truncation effects in regional projections of species distribution. *Ecological Applications*, **31**, e02427.

Del Toro, I., Ribbons, R. R., Hayward, J. and Andersen, A. N. (2019) Are stacked species distribution models accurate at predicting multiple levels of diversity along a rainfall gradient? *Austral Ecology*, **44**, 105–113.

Doser, J. W., Leuenberger, W., Sillett, T. S., Hallworth, M. T. and Zipkin, E. F. (2022) Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods in Ecology and Evolution*, **13**, 919–932.

Durante, D. (2017) A note on the multiplicative gamma process. *Statistics & Probability Letters*, **122**, 198–204.

Fletcher, R. J., McCleery, R. A., Greene, D. U. and Tye, C. A. (2016) Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, **31**, 1369–1382.

Fletcher Jr, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A. and Dorazio, R. M. (2019) A practical guide for combining data to model species distributions. *Ecology*, **100**, e02710.

Gatti, R. C., Amoroso, N. and Monaco, A. (2020) Estimating and comparing biodiversity with a single universal metric. *Ecological Modelling*, **424**, 109020.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Statistics and computing*, **24**, 997–1016.

Gotelli, N. J. and Chao, A. (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data.

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

Haase, P., Bowler, D. E., Baker, N. J., Bonada, N., Domisch, S., Garcia Marquez, J. R., Heino, J., Hering, D., Jähnig, S. C., Schmidt-Kloiber, A. et al. (2023) The recovery of european freshwater biodiversity has come to a halt. *Nature*, 1–7.

Harremoës, P. (2001) Binomial and poisson distributions as maximum entropy distributions. *IEEE Transactions on Information Theory*, **47**, 2039–2041.

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

Hooten, M. B. and Hobbs, N. T. (2015) A guide to bayesian model selection for ecologists. *Ecological monographs*, **85**, 3–28.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A. et al. (2020) Data integration for large-scale models of species distributions. *Trends in ecology & evolution*, **35**, 56–67.

Itô, Y. (2007) Recommendations for the use of species diversity indices with reference to a recently published article as an example. *Ecological research*, **22**, 703–705.

Jacod, J. (1975) Two dependent poisson processes whose sum is still a poisson process. *Journal of Applied Probability*, **12**, 170–172.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.

Jost, L., DeVries, P., Walla, T., Greeney, H., Chao, A. and Ricotta, C. (2010) Partitioning diversity for conservation analyses. *Diversity and Distributions*, **16**, 65–76.

Kéry, M. and Royle, J. A. (2015) Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in r and bugs.

— (2020) *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and advanced models*. Academic Press.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M. and Stone, L. (2017) Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, **8**, 420–430.

Lahti, L., Shetty, S. and Ernst, F. (2021) Orchestrating microbiome analysis.

Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, 5–13.

Loeys, T., Moerkerke, B., De Smet, O. and Buysse, A. (2012) The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, **65**, 163–180.

MacKenzie, D. I. and Bailey, L. L. (2004) Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 300–318.

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

Fernández-i Marín, X. (2013) Using the ggmcmc package.

Miller, D. A., Pacifici, K., Sanderlin, J. S. and Reich, B. J. (2019) The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, **10**, 22–37.

Morishita, M. (1996) On the influence of the sample size upon the values of species diversity. *Jpn. J. Ecol.*, **46**, 269–289.

Nagendra, H. (2002) Opposite trends in response for the shannon and simpson indices of landscape diversity. *Applied geography*, **22**, 175–186.

Nicenboim, B., Schad, D. and Vasishth, S. (2021) An introduction to bayesian data analysis for cognitive science. *Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series*.

O'Connor, R. S., Kunin, W. E., Garratt, M. P., Potts, S. G., Roy, H. E., Andrews, C., Jones, C. M., Peyton, J. M., Savage, J., Harvey, M. C. et al. (2019) Monitoring insect pollinators and flower visitation: The effectiveness and feasibility of different survey methods. *Methods in Ecology and Evolution*, **10**, 2129–2140.

O'hara, R. (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, 375–386.

Ovaskainen, O. and Abrego, N. (2020) *Joint Species Distribution Modelling: Biotic Interactions*, 142–183. Ecology, Biodiversity and Conservation. Cambridge University Press.

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. and Collazo, J. A. (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, **98**, 840–850.

Pacifici, K., Reich, B. J., Miller, D. A. and Pease, B. S. (2019) Resolving misaligned spatial data with integrated species distribution models. *Ecology*, **100**, e02709.

Potts, S., Dauber, J., Hochkirch, A., Oteman, B., Roy, D., Ahrne, K., Biesmeijer, K., Breeze, T., Carvell, C., Ferreira, C. et al. (2020) Proposal for an eu pollinator monitoring scheme. *Publications Office of the European Union: Ispra, Italy*.

Pouteau, R., Bayle, É., Blanchard, É., Birnbaum, P., Cassan, J.-J., Hequet, V., Ibanez, T. and Vandrot, H. (2015) Accounting for the indirect area effect in stacked species distribution models to map species richness in a montane biodiversity hotspot. *Diversity and Distributions*, **21**, 1329–1338.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Redko, I., Morvant, E., Habrard, A., Sebban, M. and Bennani, Y. (2019) *Advances in domain adaptation theory*. Elsevier.

Roswell, M., Dushoff, J. and Winfree, R. (2021) A conceptual guide to measuring species diversity. *Oikos*, **130**, 321–338.

Schaub, M., Gimenez, O., Sierro, A. and Arlettaz, R. (2007) Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. *Conservation Biology*, **21**, 945–955.

Scheme, U. P. M. (2022a) Flower-insect timed count data from the uk pollinator monitoring scheme, 2017-2020 version 2. URL: https://doi.org/10.5285/13aed7ac-334f-4bb7-b476-4f1c3da45a13.

— (2022b) Pan-trap survey data from the uk pollinator monitoring scheme, 2017-2020. URL: https://doi.org/10.5285/06cc6b8f-9bd4-4ae4-af5d-65bcbd319e9f.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. and O'Hara, R. B. (2020) Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography*, **43**, 1413–1422.

Song, Q., Wang, B., Wang, J. and Niu, X. (2016) Endangered and endemic species increase forest conservation values of species diversity based on the shannon-wiener index. *iForest-Biogeosciences and Forestry*, **9**, 469.

Strebel, N., Kéry, M., Guélat, J. and Sattler, T. (2022) Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *Journal of Biogeography*.

Team, R. C., Team, M. R. C., Suggests, M. and Matrix, S. (2013) Package "stats.". *RA Lang. Environment Stat. Comput. Vienna, Austria: R Foundation for Statistical Computing*.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. and Ovaskainen, O. (2020) Joint species distribution modelling with the r-package hmsc. *Methods in ecology and evolution*, **11**, 442–447.

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P. and Sattler, T. (2019) Joint species distribution models with species correlations and imperfect detection. *Ecology*, **100**, e02754.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. and Bodik, R. (2017) Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, **26**, 403–413.

536 Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical bayesian model evaluation using leave-one-out cross-
537 validation and waic. *Statistics and computing*, **27**, 1413–1432.

538 Wright, W. J., Irvine, K. M. and Rodhouse, T. J. (2016) A goodness-of-fit test for occupancy models with
539 correlated within-season revisits. *Ecology and Evolution*, **6**, 5404–5415.

540 Zeleny, D. (2020) Analysis of community ecology data in r. davidzeleny. net.

541 Zulian, V., Miller, D. A. and Ferraz, G. (2021) Integrating citizen-science and planned-survey data improves
542 species distribution estimates. *Diversity and Distributions*, **27**, 2498–2509.