ECOGRAPHY

Research article

Integrating data from different taxonomic resolutions to better estimate community alpha diversity

Kwaku Peprah Adjei[1,3], **Claire Carvell**[2], **Nick J. B. Isaac**[2], **Francesca Mancini**[2] and **Robert B. O'Hara**[1,3]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[2]UK Centre for Ecology & Hydrology, Wallingford, UK
[3]Center for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence: Kwaku Peprah Adjei (kadjeipeprah94@gmail.com)

Ecography
2024: e07182
doi: 10.1111/ecog.07182
Subject Editor:
John-Arvid Grytnes
Editor-in-Chief: Miguel Araújo
Accepted 16 December 2023

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 properly 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 community IDMs to capture the true 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.

Keywords: alpha diversity, Bayesian models, Markov chain Monte Carlo methods, multispecies distribution models, UK Pollinator Monitoring Scheme

Introduction

Biodiversity monitoring programs generate disparate data types that are used to infer and make predictions about species distributions, dynamics and diversity (Bird et al. 2014, Kéry and Royle 2015, 2020, Isaac et al. 2020). 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 or leverages the information in one dataset to improve the inferences and predictions from models fitted to other datasets.

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 (Koshkina et al. 2017, Simmonds et al. 2020, Ahmad Suhaimi et al. 2021). These studies have shown that IDMs have improved predictive performance and higher precision and accuracy of estimated parameters than single-dataset 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).

Initial developments of IDMs have been developed for single species (Koshkina et al. 2017, Pacifici et al. 2017, Miller et al. 2019). In recent years, however, much focus has been on integrated community models (Doser et al. 2022, Lauret et al. 2023, Zipkin et al. 2023), that take advantage of data integration and hierarchical community modelling framework to combine multi-species datasets. These integrated community models allow parameter estimation for all species (irrespective of their sample sizes; Zipkin et al. 2009) by producing estimates of community-level parameters such as variance among species (Zipkin et al. 2023). The community-level parameters are then used to make inferences about the effects of covariates and other environmental stressors in a community (Zipkin et al. 2023).

An unrealised benefit of such integrated community models is the potential to estimate parameters that would be challenging to estimate (unless strong model assumptions are made; Royle and Nichols 2003) 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). 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).

Additionally, 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). These broader taxonomically resolved count data do not contain species-level information and are not used in estimating metrics that require species-level information such as alpha diversity.

In this study, we combine two data types (count data resolved to a broader taxonomical level and species-level presence–absence data) in an IDM – specifically an integrated community model – 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 used for a long time (Besbeas et al. 2002, Schaub et al. 2007, Abadi et al. 2010). 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, UK Pollinator Monitoring Scheme 2023), 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 standardised 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).

Material and 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 Material and methods deal with inference and the estimation of community parameters.

UK Pollinator Monitoring Scheme data

The data used is a subset of the PoMS data (UK Pollinator Monitoring Scheme 2022a, b). PoMS implements a systematic survey with 95 1 km^2 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 (O'Connor et al. 2019, Breeze et al. 2021). 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^2 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 50 × 50 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 1 km 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 the Supporting information.

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').

Table 1. Summaries of the group-level flower-insect timed (FIT) Count and species-level pan trap occupancy data. Both datasets were collected from 74 survey sites (N_{sites}) with eight survey visits (N_{visit}) (up to four visits in each year 2017 and 2018). The average FIT Counts and their SD (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 naive occupancy is defined as the proportion of species in an insect group that were detected across the 74 sites and its SD is its variation over the eight survey visits. The number of species ($N_{species}$) in the species list for each insect group is also provided in the summary.

Insect group	FIT Counts	Pantrap occupancy	
	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)

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 model intensity as a linear function of latitude, reflecting the strong latitudinal gradient in pollinator diversity across GB (Powney et al. 2019), although we do not claim that latitude is the only correlate of pollinator diversity, or even the most important one. 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 'complete-parameter-sharing' and the 'independent-intercept' formulation. Both variants are hierarchical community model (sensu Dorazio and Royle 2005), in that information is shared across species to yield more precise estimates of the species-specific and community-level parameters.

Complete-parameter-sharing formulation

The intensity of each species in an insect group is linked to the occupancy probability using the complementary log–log link function (using the equivalence relationship between the clog–log transformed occupancy probability and the log-transformed intensity; Kéry and Royle 2015, Bowler et al. 2019), which defines the probability that at least one organism is present (Kéry and Royle 2015)(Eq. 1):

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

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 for an insect group.

In this model structure, all the parameters are shared by the two state variables. 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).

Independent-intercept 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.

Here, we also describe a hierarchical community model and define the link function for the latent variables for an insect group in this IDM framework as (Eq. 2):

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

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 complete-parameter-sharing model, the intercept and covariate effect hyperparameters represent their respective community-level mean and variance parameters in the models for an insect group.

This independent-intercept formulation 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.

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 and group-level count data (can be defined.

In addition, the model can estimate various community ecology metrics as derived parameters. Based on preliminary analysis (Supporting information), 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.

Sub-models for species occupancy data

We model the species occupancy data with an occupancy-detection model (MacKenzie et al. 2002). We assume in this system that the occupancy models across the visits are closed (i.e. there will be no immigration and emigration in the system), since we treat the years as visits. This assumption can be relaxed by using a multiseason or dynamic occupancy model (MacKenzie et al. 2003, Altwegg and Nichols 2019). 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 Eq. 1–2.

The detection probability (p_{ijk}) for species j at site i during the survey visit k is modelled with a site and species and visit random effect logistic regression using the logit link. That is (Eq. 3):

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

ζ_i, the effect of site i, is normally distributed with zero mean and variance σ^2_ζ; ν_j, the effect of species j, is normally distributed with zero mean and variance σ^2_ν; and ρ_k, the effect of survey visit k, is normally distributed with mean 0 and variance σ^2_ρ. 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 (Supporting information). By the definition of our model for the detection probability in Eq. 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. This average detection probability can be allowed to deviate from 0.5 by adding an intercept term in Eq. 3.

Let the observation for species j during the kth 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, $X_{ijk} \sim$ Binomial $(5, z_{ij} \times p_{ijk})$).

Sub-model for group count data

Having defined the intensity of species j at location i (Eq. 1–2), 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 (Jacod 1975, Harremoës 2001).

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 $\gamma_i = \frac{\theta}{\theta + \lambda_i^g}$, where $\lambda_i^g = \Sigma_j^S \lambda_{ij}$ is the intensity of the group counts at site i (that is, $Y_{ik} \sim \text{NB}(\theta, \gamma_i)$ with mean and variance $\frac{\theta(1-\gamma_i)}{\gamma_i}$ and $\frac{\theta}{\gamma_i^2}(1-\gamma_i)$ respectively). The parameter θ is the overdispersion parameter, which allows us to model the extra variation in the group count data. Note that as $\theta \longrightarrow \infty$, the negative binomial distribution converges upon the Poisson distribution.

We present the various joint likelihood structures defined in section 'State variables' and the sub-models for each PoMS dataset defined in section ('Sub-models for each dataset') in Fig. 1.

Single-dataset models

We fit three single-dataset models and compare their results with those from the IDM. The three models are:

Species occupancy model (SOM)

We model the species occupancy data with the occupancy-detection model described in section 'Sub-models for species occupancy data'. The occupancy probability (ψ_{ij}) is modelled using the cloglog link as defined in the independent-intercept formulation described in Eq. 2. The occupancy probability (ψ_{ij}) can be converted into an estimate of the mean intensity (λ_{ij}) to be used in estimating the alpha diversity by using the relation (Eq. 4):

$$\lambda_{ij} = -\log(1 - \psi_{ij}). \quad (4)$$

Group count models (GCM)

Our model for FIT Count data has a negative binomial distribution as described in section 'Sub-model for group count data'. We define the mean intensity (λ_{ij}) using both the complete-parameter-sharing formulation described in Eq. 1 (which we will refer to as GCMSH in this study) and the independent-intercept formulation described in Eq. 2 (which we will refer to as GCMCO in this study).

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

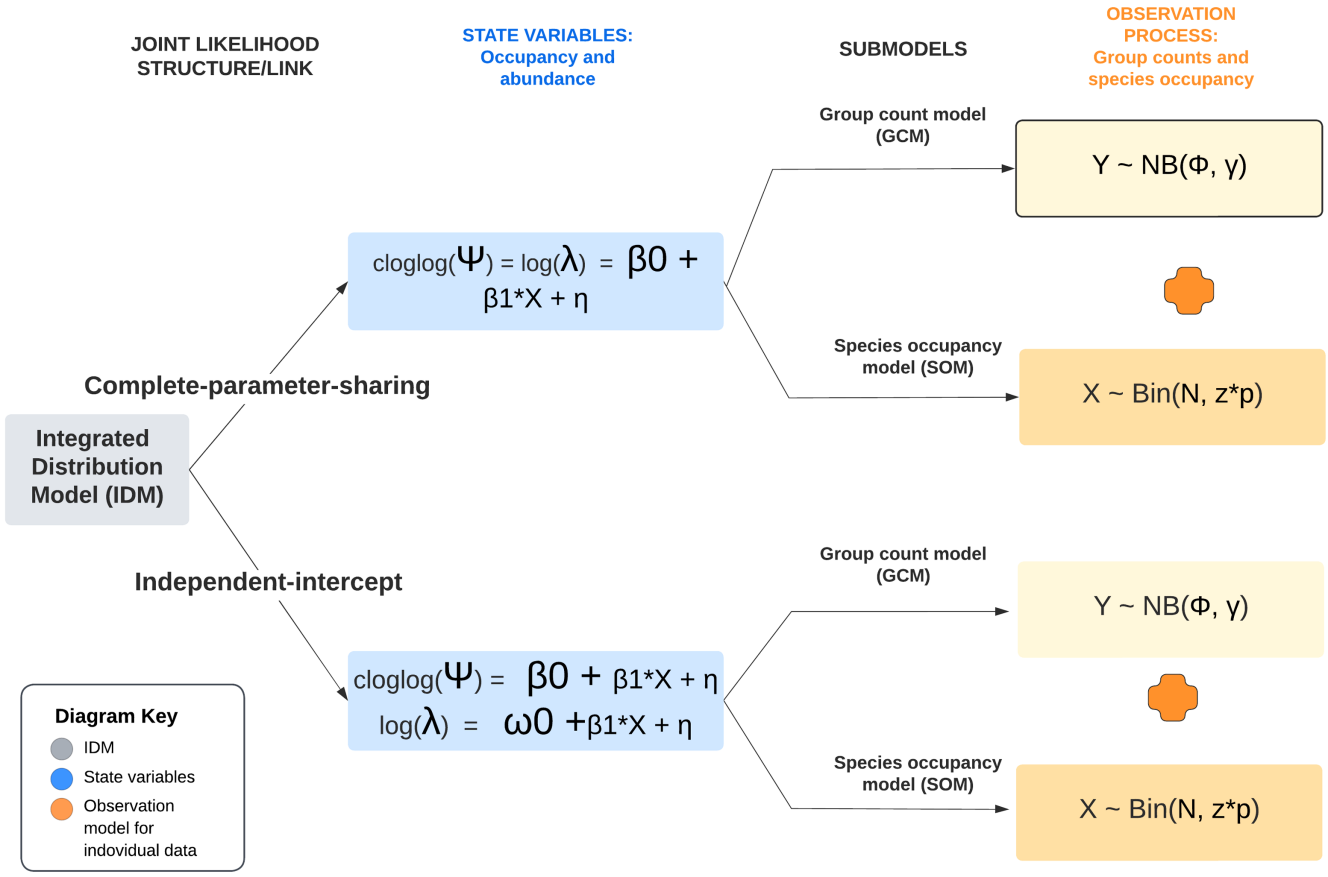

Figure 1. Flowchart showing the integrated distribution model process for the Pollinator Monitoring Scheme (PoMS) survey data. To fit the integrated distribution model (IDM), we used two joint likelihood links in this study: complete-parameter-sharing formulation and independent-intercept formulation. The state variables models are defined by Eq. 1–2. 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. The IDM combines the SOM and GCM for each joint likelihood link used. All the parameters in this flowchart are defined in sections 'State variables' and 'Sub-models for each dataset'. In addition to the intercept and covariate effect, we add a species interaction effect η in this flowchart.

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.

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, O'Hara 2005, Itô 2007, Chao and Jost 2015, Gatti et al. 2020), 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 easily extended to the others. The Shannon–Wiener diversity index was calculated as (Eq. 5):

$$H^1 = -\sum_{j=1}^{S} r_{ij} \log(r_{ij}), \tag{5}$$

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

Evaluating model performance

We fitted five models to the PoMS survey data in this study: IDM with the complete-parameter-sharing formulation defined in section 'Complete-parameter-sharing' which we will refer to as IDMSH, the IDM with independent-intercepts formulation defined in section 'Independent-intercept formulation' which we will refer to as IDMCO, and the three single-dataset models described in section 'Single-dataset models'. These models are summarised in Table 2.

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 (www.r-project.org). 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 300 000 iterations for all the models, and 200 000 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 2016) and rejected the models with R-hat greater than 1.1.

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 100 data replicates using the IDM framework for each latent variable formulation used in this study (i.e. using both the independent-intercept and complete-parameter-sharing 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 section 'State variables' 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 100 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 complete-parameter-sharing 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 (Eq. 6):

Table 2. Models fitted in this study, their descriptions, predictors, type and data used to fit them. Two integrated models: IDM with independent-intercepts formulation (IDMSH) and IDM with the complete-parameter-sharing formulation (IDMCO), and three single-dataset 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 'Sub-models for each dataset', with lat referring to the latitudinal gradient slope.

Model	Model description	Type	Data used	Predictor
IDMSH	IDM with complete-parameter-sharing structure defined in Eq. 1	Integrated	GC and SO	$\beta_{0j} + \beta_{1j} \times \text{lat}_i$
IDMCO	IDM with independent-intercept structure defined in Eq. 2			$\beta_{0j} + \beta_{1j} \times \text{lat}_i$ $\omega_0 + \beta_{1j} \times \text{lat}_i$
GCMSH	GCM with complete-parameter-sharing structure defined in Eq. (1)	Single	GC	$\beta_{0j} + \beta_{1j} \times \text{lat}_i$
GCMCO	GCM with independent-intercept structure defined in Eq. 2			$\omega_0 + \beta_{1j} \times \text{lat}_i$
SOM	Species occupancy model	Single	SO	$\beta_{0j} + \beta_{1j} \times \text{lat}_i$

Table 3. Log predictive density from the twofold cross-validation, the marginal contribution of each 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 $lppd_{GCMCO} - lppd_{IDMCO}$. Negative values of the marginal contribution indicate that a data source did not contribute any information into the IDM and larger positive values indicate that a data source contributed information into the IDM (numbers in bold are the largest log predictive density values which indicate the best model, and the dataset with the largest marginal contribution.).

		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	−3261.26	−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	**−11 341**	**−11165.97**	−175.39	**3206.45**	−24.88
	IDMCO	3.27	0.31	−11464.48	−11285.78	−178.70	3086.64	−28.19

$$\text{Mean bias} = \frac{1}{100}\sum_{k=1}^{100}(\hat{H}_i^{'(k)} - H_i^{'(k)}),$$
$$\text{Mean precision} = \frac{1}{100}\sum_{k=1}^{100}\frac{1}{\text{SD}(\hat{H}_i^{'(k)})^2}, \quad (6)$$

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 $\text{SD}(\hat{H}_i^{'(k)})$ is the posterior SD 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.

Model validation and assessment

Model predictive performance

We performed twofold 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 (Gelman et al. 2014, Nicenboim et al. 2021) was used to measure the cross-validation's predictive accuracy. The log pointwise predictive density (lppd) is defined as (Eq. 7):

$$\text{lppd} = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}\log(P(y_{n,k} \mid y_{-n,k}, \hat{\theta}, M)) \quad (7)$$

where $\log(P(y_{n,k} \mid y_{-n,k}, \hat{\theta}, M))$ 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-dataset 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 comparison of predictive log-likelihoods 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.

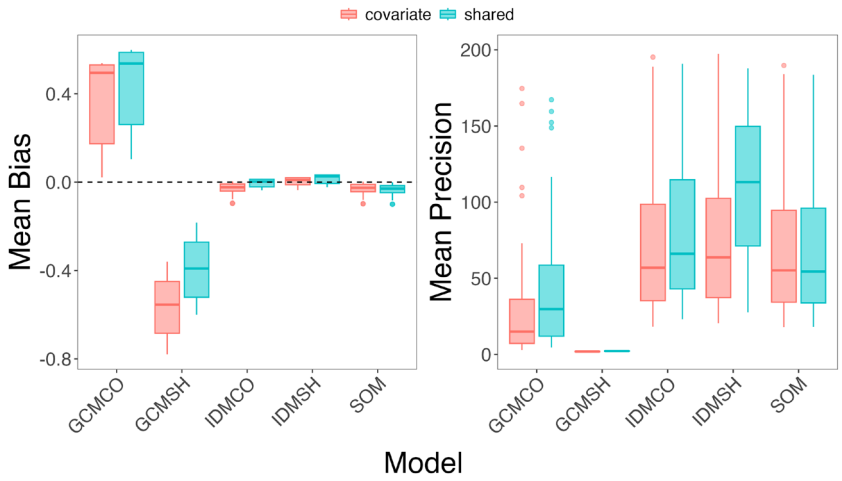

Figure 2. Mean bias and precision of Shannon index from the five study models fitted to the simulated data. The boxplot shows the distribution of the mean bias and precision for the 74 sites. The five models: two integrated models (IDMSH and IDMCO), two group count models (GCMSH and GCMCO) and a species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

Results

The estimates of the mean bias and precision of Shannon diversity estimates over all 30 replicated simulated datasets are presented in Fig. 2. The log-predictive density from the twofold 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 Fig. 2. All other figures and tables referred to in this section are presented in the Supporting information.

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 'Simulation study'. Whether the data were simulated under the complete-parameter-sharing and independent-intercept 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 (Fig. 2), and both models have much lower precision than other models. As explained in the Material and 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 Fig. 2). When the data was simulated under the independent-intercept formulation (red bars), the precision of Shannon diversity estimates from the integrated models was similar to that of SOM.

Analysis of PoMS dataset

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; Table 1, 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 the Supporting information. Firstly, there was a negative latitudinal effect on the estimates of the Shannon indices (that is, the Shannon index decreased with the latitudinal gradient; Supporting information). 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 (Supporting information). 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 (Fig. 3). This precision was also higher for the sites with higher Shannon diversity estimates (compare Fig. 3, Supporting information).

Predictive performance and information provided

Table 3 shows the twofold 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 best fitting model in each case is an IDM, with the independent-intercept formulated model (IDMCO) performing best for pantrap data and the complete-parameter-sharing formulated model performing best for the FIT Counts. For solitary bees, we find that both IDMs outperform the SOM for predicting the pantrap data,

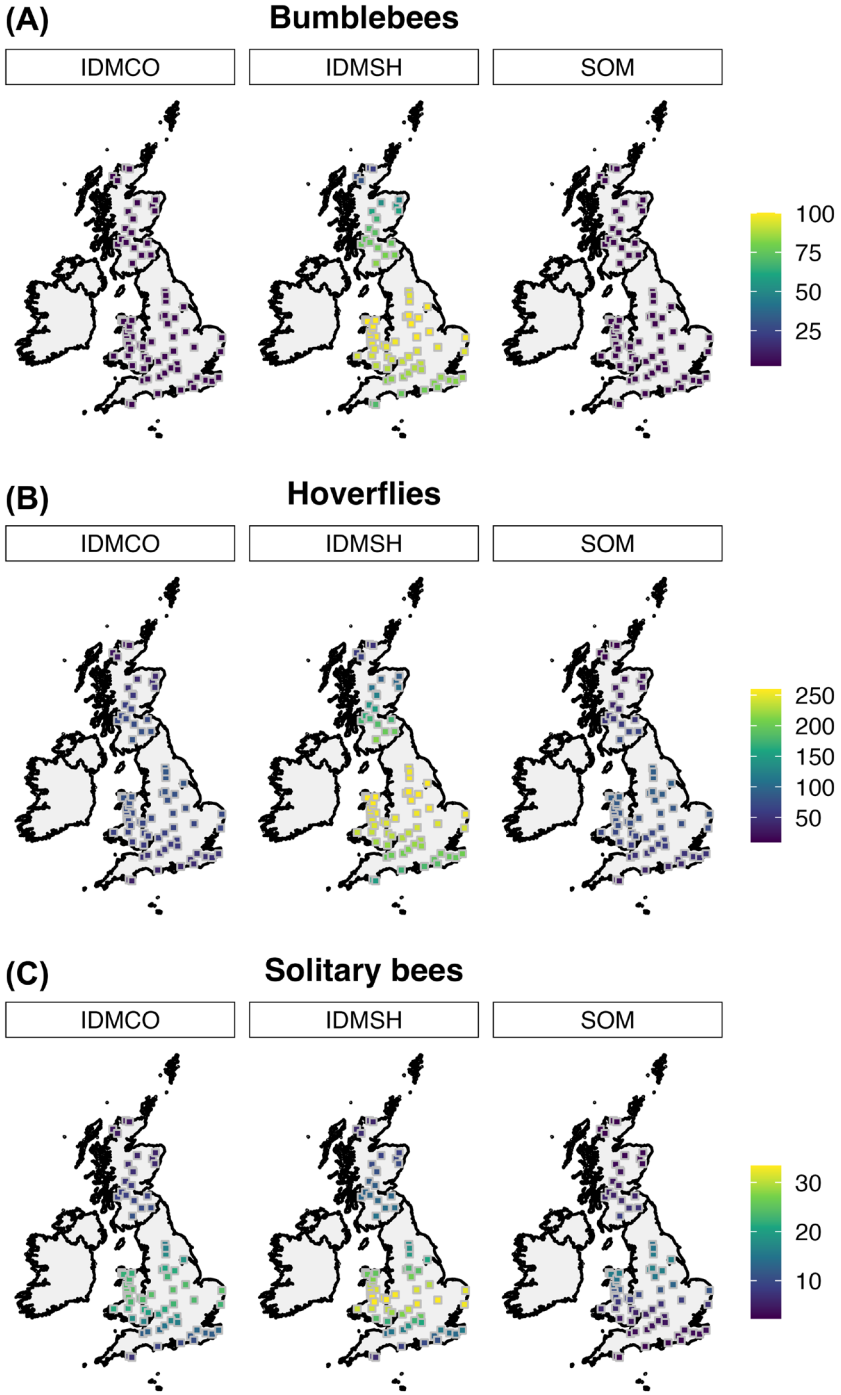

Figure 3. Precision of the Shannon diversity (H') estimates for each of the 74 PoMS sites from the five models in this study summarised in Table (2) for each of the insect groups: (A) bumblebees, (B) hoverflies and (C) solitary bees.

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-dataset 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.

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 a multi-species 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-dataset models in most cases.

Previous studies have shown that IDMs perform better than single-dataset models in estimating state variable parameters and prediction accuracy of new datasets (Pacifici et al. 2017, Miller et al. 2019, Doser et al. 2022, Strebel et al. 2022). Miller et al. (2019) noted that IDMs present opportunities to model community dynamics and diversity from multiple datasets, and Zipkin et al. (2023) presented an integrated community model to explore such opportunities. Our IDM shares information among species and provides estimates for parameters at both species and community levels (Zipkin et al. 2023). Our work provides further advancements in such integrated community models by providing an IDM that combines data from different taxonomic levels to estimate alpha diversity in a community. Our simulation and case studies showed that the IDMs outperform the single-dataset models in producing precise alpha diversity estimates in a community if both datasets share information between them (Fig. 2, Table 3, Supporting information). 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 (Miller et al. 2019, Isaac et al. 2020).

Furthermore, the proposed IDMs outperform the single-dataset models' prediction accuracy of new datasets for some insect groups. From our model assessment of the PoMS data using twofold cross-validation, IDMs outperformed the single-dataset 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 independent-intercept and complete-parameter-sharing 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 complete-parameter-sharing formulation ensured that all state variables were shared between both datasets. The independent-intercept formulation allowed some flexibility in sharing the state model definition by allowing each dataset to have a unique intercept. Previous studies on IDMs, using either independent-intercept or complete-parameter-sharing formulation, have all shown that IDMs have higher prediction accuracy than single-dataset models (Fletcher et al. 2016, 2019, Koshkina et al. 2017, Pacifici et al. 2017, Simmonds et al. 2020, Adde et al. 2021, Ahmad Suhaimi et al. 2021, Zulian 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-dataset models since all the IDMs we tested performed comparably better than the single-dataset 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 (Fig. 2–3, 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.

Acknowledgements – PoMS is indebted to the many volunteers who carry out surveys and contribute data to the scheme, as well as to those who allow access to their land for surveys, and the taxonomists identifying specimens. We thank Nadine Mitschunas and Martin Harvey for coordinating PoMS surveys and data publication. We also acknowledge the assistance of Daniel Turek in the data analysis. *Funding* – This study is part of the Transforming Citizen Science for Biodiversity project funded by the Digital Transformation initiative of the Norwegian University of Science and Technology. The UK Pollinator Monitoring Scheme (PoMS) is a partnership funded jointly by UKCEH and JNCC (through funding from Defra, Scottish Government, Welsh Government, and DAERA). UKCEH's contribution is funded by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

Author contributions

Kwaku Peprah Adjei: Conceptualization (equal); Formal analysis (lead); Methodology (equal); Writing – original draft (lead); Writing – review and editing (lead). **Claire Carvell:** Conceptualization (equal); Data curation (equal); Funding acquisition (lead); Project administration (lead); Writing – review and editing (equal). **Nick J. B. Isaac:** Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **Francesca Mancini:** Data curation (equal); Visualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Robert B. O'Hara:** Conceptualization (equal); Funding acquisition (lead); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://publons.com/publon/10.1111/ecog.07182.

Data availability statement

Data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.34tmpg4s0 (Adjei et al. 2024) and the code are available from Zenodo: https://doi.org/10.5281/zenodo.8424494 (Adjei et al. 2023).

Supporting information

The Supporting information associated with this article is available with the online version.

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.

Adjei, K. P., Carvell, C., Isaac, N. J. B., Mancini, F. and O'Hara R. B 2023. Code for: Integrating data from different taxonomic resolutions to better estimate community alpha diversity (v1.0.0). – Zenodo Digital Resository, https://doi.org/10.5281/zenodo.8424494

Adjei, K. P., Carvell, C., Isaac, N. J. B., Mancini, F. and O'Hara, R. B. 2024. Data from: Integrating data from different taxonomic resolutions to better estimate community alpha diversity. – Dryad Digital Repository, https://doi.org/10.5061/dryad.34tmpg4s0.

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. – Divers. Distrib. 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.

Altwegg, R. and Nichols, J. D. 2019. Occupancy models for citizen-science data. – Methods Ecol. Evol. 10: 8–21.

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. – Global Ecol. Biogeogr. 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 98: 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., Pecl, G. T., Barrett, N. and Frusher, S. 2014. Statistical solutions for error and bias in global citizen science datasets. – Biol. Conserv. 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 Ecol. Evol. 32: 477–487.

Bowler, D. E., Nilsen, E. B., Bischof, R., O'Hara, R. B., Yu, T. T., Oo, T., Aung, M. and Linnell, J. D. 2019. Integrating data from

different survey types for population monitoring of an endangered species: the case of the eld's deer. – Sci. Rep. 9: 7766.
Breeze, T. D. et al. 2021. Pollinator monitoring more than pays for itself. – J. Appl. Ecol. 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 Ecol. Evol. 6: 99–108.
Brooks, S. P. and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. – J. Comp. Graph. Stat. 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. – Ecol. Evol. 4: 1389–1397.
Chao, A. and Jost, L. 2015. Estimating diversity and entropy profiles via discovery rates of new species. – Methods Ecol. Evol. 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. – Ecol. Appl. 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 Ecol. 44: 105–113.
Dorazio, R. M. and Royle, J. A. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. – J. Am. Stat. Assoc. 100: 389–398.
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 Ecol. Evol. 13: 919–932.
Durante, D. 2017. A note on the multiplicative gamma process. – Stat. Probab. Lett. 122: 198–204.
Fernández, M. X. 2016. ggmcmc: analysis of MCMC samples and Bayesian inference. – J. Stat. Softw. 70 9: 1–20. doi:10.18637/jss.v070.i09
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 Ecol. 31: 1369–1382.
Fletcher, R. J. Jr., 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. – Ecol. Modell. 424: 109020.
Gelman, A., Hwang, J. and Vehtari, A. 2014. Understanding predictive information criteria for bayesian models. – Stat. Comput. 24: 997–1016.
Gotelli, N. J. and Chao, A. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. – Encyclopedia of biodiversity pp. 195–211.
Guillera-Arroita, G., Kéry, M. and Lahoz-Monfort, J. J. 2019. Inferring species richness using multispecies occupancy modeling: estimation performance and interpretation. – Ecol. Evol. 9: 780–792.
Haase, P. et al. 2023. The recovery of European freshwater biodiversity has come to a halt. – Nature 620: 582–588.
Harremoës, P. 2001. Binomial and poisson distributions as maximum entropy distributions. – IEEE Trans. Inform. Theor. 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. – Ecol. Monogr. 85: 3–28.
Isaac, N. J. B., 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., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G. and O'Hara, R. B. 2020. Data integration for large-scale models of species distributions. – Trends Ecol. Evol. 35: 56–67.
Itô, Y. 2007. Recommendations for the use of species diversity indices with reference to a recently published article as an example. – Ecol. Res. 22: 703–705.
Jacod, J. 1975. Two dependent poisson processes whose sum is still a poisson process. – J. Appl. Probab. 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. – Divers. Distrib. 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: volume 1: Prelude and static models. – Elsevier.
Kéry, M. and Royle, J. A. 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 Ecol. Evol. 8: 420–430.
Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – Oikos 76: 5–13.
Lauret, V., Labach, H., David, L., Authier, M. and Gimenez, O. 2023. Using integrated multispecies occupancy models to map co-occurrence between bottlenose dolphins and fisheries in the gulf of lion, french Mediterranean sea. – Oikos 2023: e10270.
Loeys, T., Moerkerke, B., De Smet, O. and Buysse, A. 2012. The analysis of zero-inflated count data: beyond zero-inflated poisson regression. – Br. J. Math. Stat. Psychol. 65: 163–180.
MacKenzie, D. I. and Bailey, L. L. 2004. Assessing the fit of site-occupancy models. – J. Agric. Biol. Environ. Stat. 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.
MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G. and Franklin, A. B. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. – Ecology 84: 2200–2207.
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 Ecol. Evol. 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. – Appl. Geogr. 22: 175–186.
Nicenboim, B., Schad, D. and Vasishth, S. 2021. An introduction to bayesian data analysis for cognitive science. – Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series.

O'Connor, R. S., Kunin, W. E., Garratt, M. P. D., Potts, S. G., Roy, H. E., Andrews, C., Jones, C. M., Peyton, J. M., Savage, J., Harvey, M. C., Morris, R. K. A., Roberts, S. P. M., Wright, I., Vanbergen, A. J. and Carvell, C. 2019. Monitoring insect pollinators and flower visitation: the effectiveness and feasibility of different survey methods. – Methods Ecol. Evol. 10: 2129–2140.

O'Hara, R. B. 2005. Species richness estimators: how many species can dance on the head of a pin? – J. Anim. Ecol. 74: 375–386.

Ovaskainen, O. and Abrego, N. 2020. Joint species distribution modelling: biotic interactions. – Ecology, biodiversity and conservation. Cambridge Univ. Press, pp. 142–183.

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. W. and Pease, B. S. 2019. Resolving misaligned spatial data with integrated species distribution models. – Ecology 100: e02709.

Potts, S. et al. 2020. Proposal for an EU Pollinator Monitoring Scheme. – Publications Office of the European Union.

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. – Divers. Distrib. 21: 1329–1338.

Powney, G. D., Carvell, C., Edwards, M., Morris, R. K. A., Roy, H. E., Woodcock, B. A. and Isaac, N. J. B. 2019. Widespread losses of pollinating insects in Britain. – Nat. Commun. 10: 1018.

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.

Royle, J. A. and Nichols, J. D. 2003. Estimating abundance from repeated presence–absence data or point counts. – Ecology 84: 777–790.

Schaub, M., Gimenez, O., Sierro, A. and Arlettaz, R. 2007. Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. – Conserv. Biol. 21: 945–955.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B. 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 Biogeosci. For. 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. – J. Biogeogr. 49: 563–575.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J. and Ovaskainen, O. 2020. Joint species distribution modelling with the r-package hmsc. – Methods Ecol. Evol. 11: 442–447.

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

UK Pollinator Monitoring Scheme 2022a. Flower-insect timed count data UK Pollinator Monitoring Scheme, 2017–2020 ver. 2. NERC EDS Environmental Information Data Centre. doi.org/10.5285/13aed7ac-334f-4bb7-b476-4f1c3da45a13.

UK Pollinator Monitoring Scheme 2022b. Pan-trap survey data from the UK Pollinator Monitoring Scheme, 2017–2020. NERC EDS Environmental Information Data Centre. doi.org/10.5285/2c43ba3c-d821-442c-989b-754451d72091.

UK Pollinator Monitoring Scheme 2023. The UK PoMS annual report 2022. UK Centre for Ecology & Hydrology and Joint Nature Conservation Committee.

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. – J. Comp. Graph. Stat. 26: 403–413.

Vehtari, A., Gelman, A. and Gabry, J. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. – Stat. Comput. 27: 1413–1432.

Wright, W. J., Irvine, K. M. and Rodhouse, T. J. 2016. A goodness-of-fit test for occupancy models with correlated within-season revisits. – Ecol. Evol. 6: 5404–5415.

Zipkin, E. F., DeWan, A. and Andrew Royle, J. 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. – J. Appl. Ecol. 46: 815–822.

Zipkin, E. F., Doser, J. W., Davis, C. L., Leuenberger, W., Ayebare, S. and Davis, K. L. 2023. Integrated community models: a framework combining multispecies data sources to estimate the status, trends and dynamics of biodiversity. – J. Anim. Ecol. 92: 2248–2262.

Zulian, V., Miller, D. A. W. and Ferraz, G. 2021. Integrating citizen-science and planned-survey data improves species distribution estimates. – Divers. Distrib. 27: 2498–2509.