Received: 7 July 2021 | Revised: 25 January 2022 | Accepted: 2 February 2022

DOI: 10.1111/ddi.13495

Diversity and Distributions WILEY

METHOD

A new statistical approach for identifying rare species under imperfect detection

Jafet Belmont | Claire Miller | Marian Scott | Craig Wilkie

School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

Correspondence
Jafet Belmont, School of Mathematics and Statistics, University of Glasgow, University Place, G12 8QQ Glasgow, UK.
Email: j.belmont-osuna.1@research.gla.ac.uk

Funding information
Consejo Nacional de Ciencia y Tecnología, Grant/Award Number: 494334; Natural Environment Research Council, Grant/Award Number: NE/N005740/1

Editor: Raimundo Real

Abstract

Aim: Species rarity is often used as a measure to assess the risk of extinction of species, and thus, different methods have been developed to describe the composition of rare species in biological communities. These methods usually depend on species attributes that are not always available and very often ignore imperfect species detection. In this work, we developed a new method to characterize species rarity in a community when species are detected imperfectly. Our modelling framework is based on Bayesian occupancy models to estimate species distributions under imperfect detection using presence-nondetection data.

Innovation: We propose a finite mixture occupancy model to identify rare species based on their occupancy and class-membership probabilities. Here, we explored a two-class finite mixture model to distinguish between rare and common species classes and presented the general modelling framework for a problem with more than two classes. By using simulations, we were able to compare our model results under different scenarios obtaining a high-classification performance across all of them. Additionally, we applied our model to a data set of Odonata occurrence records that were partially observed due to imperfect detection and quantified the proportion of rare species on a national scale across waterbodies in the United Kingdom.

Main conclusions: Nowadays, biodiversity conservation involves monitoring programmes that target multiple species within a community where individual species responses may vary widely. This high variability makes the task of identifying the ecological processes that drive distributions of rare species difficult. Thus, our method represents a new approach to characterize the composition of a community in terms of species rarity while correcting for detectability bias. Our modelling framework also suggests lines of research and future developments for the understanding of how species rarity can be measured in a wide range of scenarios.

KEYWORDS
classification, community ecology, detectability, occupancy model, Odonata, rare species

1 | INTRODUCTION

The need to understand how species distributions are influenced by environmental changes has motivated the development of a wide range of species distribution models that allow for identification of the most important areas for biodiversity conservation (Elith & Leathwick, 2009). These methods are usually based on scoring procedures that describe different attributes of an assemblage of species in a community (Leroy et al., 2012). However, individual species responses may vary widely within a community, making the task of identifying the ecological processes of interest that drive occupancy patterns difficult, especially for rare and elusive species (Bailey et al., 2014).

Rare species are often a point of interest for conservationists, as they represent the most vulnerable species to environmental changes (Leroy et al., 2013). A species is considered to be rare when the range of distribution or their abundances is low with respect to the range or distributional properties of other species within a comparable taxa (Blackburn & Gaston, 1997). Thus, species rarity is considered to be a scale-dependent, relative-emerging property for a set of species (Hartley & Kunin, 2003).

Species with limited distributions are prone to experiencing local extinctions influenced greatly by a reduction in the genetic pool due to population isolation (Ellstrand & Elam, 1993; Karron, 1997), environmental changes and demographic stochastic processes (e.g. variability in birth or death rates and fluctuating sex ratios (Lande, 1993; Lee et al., 2011)) that cause random fluctuations in the species' survival and reproduction rates (Melbourne & Hastings, 2008). However, estimating the distributions of rare species is a challenging task because occurrence records are scarce and underestimate the true species distribution due to imperfect detection (MacKenzie et al., 2005).

To correct for detection bias, multispecies occupancy models enable probabilities of species occurrence and detection to be estimated simultaneously. This class of models has proven to be a powerful tool for estimating attributes of biological communities such as size, composition and species richness when the number of species in the community is known (Gelfand et al., 2005) or unknown (Dorazio & Royle, 2005; Dorazio et al., 2006). However, their potential application to classify species is a less studied subject (Pacifici et al., 2014), and very few studies in community ecology have analysed the species classification problem while accommodating imperfect detection. For instance, Pacifici et al. (2014) proposed using multispecies occupancy models to investigate the variability in occupancy probabilities after classifying species from two distinct avian communities into different groups defined by landscape features, habitat requirements and species diet. Grouping species can reveal group-level responses to habitat covariates that would otherwise be difficult to observe if the community was analysed as a whole. However, results can be sensitive to these predefined classes (Pacifici et al., 2014), and the class labels are assumed to be known *a priori* rather than being estimated from the data.

Imperfect detection has become an increasing area of research over the last decade (Devarajan et al., 2020; Kellner & Swihart, 2014), but current methods proposed to quantify the rarity of a species at the community level have not yet accounted for detectability bias and very often rely on population density parameters, which are not always available for the less studied species. For instance, Rabinowitz (1981) proposed a classification system that has been applied in several conservation studies (e.g. Broennimann et al., 2005; Isaac et al., 2009; Maciel, 2021; Yu & Dobson, 2000) to categorize species by different types of rarity and commonness based on the local population density, the area of the species range and the number of different types of habitats each species occupies. However, abundance and habitat specificity data are not always available, especially for small invertebrates (Leroy et al., 2013).

Hence, scoring procedures based on species occurrences have been used to assess the rarity of assemblages of species where the only information available comes from occurrence data sets (Leroy et al., 2012, 2013). In this work, we also propose an occurrence-based approach to characterize the rarity of a given species. Here, our modelling framework assesses species rarity based on the relative occupancy probabilities on a National scale. We propose a Bayesian occupancy mixture model for multiple species to quantify species rarity in a community when species are detected imperfectly to help determine sites where the incidence of rare species is higher and, thus, relevant for conservation and management. Mixture models have been explored in a wide range of fields because of their flexibility to model situations in which the population of interest is a mixture of subpopulations for which subpopulation membership is not known (McLachlan & Basford, 1988). For instance, finite mixtures in Capture-Recapture models have been used to account for heterogeneity in individual capture probabilities (Norris & Pollock, 1996; Pledger, 2000). Within the context of occupancy models, mixture models have been used to model heterogeneity in detection probabilities (Royle, 2006), to account for false-positive errors (Royle & Link, 2006), and more recently, to characterize the structure of a community based on latent groups of species that have similar responses to environmental conditions (Sollmann et al., 2021).

In this work, we propose using finite mixtures as a classification tool that enables a community's structure to be described in terms of rarity/commonness. Specifically, we look at UK Odonata communities and characterize them based on a two-class finite mixture into rare and common species while accommodating imperfect detection. Odonata, a taxonomic order comprised dragonflies and damselflies, has served as an important bioindicator to assess water body quality and ecosystem integrity (Golfieri et al., 2016). Thus, ecological studies of the distributions of these species are crucial for the management and restoration of freshwater ecosystems.

2 | METHODS

2.1 | Occupancy and species trait data

Data were compiled by hydroscape (web: hydroscapeblog.wordpress.com), a project investigating how anthropogenic stressors and connectivity interact to influence biodiversity in UK freshwaters.

Odonata occurrence records (for over 4000 [1 km] grid cells defined by the presence of a waterbody) from 2000 to 2016 were taken for 39 noninvasive species from the British Dragonfly Society Recording Scheme (2020). Species-specific covariates that may affect the species detection probability were taken from Powney et al. (2014). These covariates were (1) median body size, (2) flight duration (difference in months between the start and the end of the flying period) and (3) number of different habitats that each species occupies (e.g. lowland rivers and canals, bogs moorland and lowland wet heath, ponds, lakes and woodlands).

2.2 | Generating presence-absence data

Our proposed modelling approach relies on presence-absence (a.k.a. detection-nondetection) data. However, Odonata occurrences are derived from presence-only records where there is no information regarding species nondetections. Thus, individual species nondetection records were inferred based on the information of sightings of other dragonfly and damselflies species by following K´ery et al. (2010) and Termaat et al. (2019), i.e. sighting of species i confirmed its presence at that site and it was deemed as undetected at those sites where any species other than i was recorded.

2.3 | Statistical methods

2.3.1 | Bayesian mixture occupancy model to identify rare species

Our modelling approach is based on occupancy models, a special class of methods that enable species distribution to be estimated under imperfect detection (MacKenzie et al., 2002; Tyre et al., 2003). These models are defined by specifying (i) an ecological process from which the occupancy state of whether a site is truly occupied or not (i.e. $z_j = 1$, if site j is occupied and 0 otherwise) is drawn based on the estimated species occupancy probability (ψ), and (ii) an observational process conditioned on the occupancy state that relates the observed species occurrences on repeated sampling occasions (i.e. $y_{jk} = 1$, if species is detected in site j during visit k) to the estimated detection probability (p). In this work, we developed a mixture occupancy model that can be used to classify species rarity based on their relative occurrences while accounting for the imperfect detection derived from species' nondetections.

Particularly, finite mixture models assume that a random vector $\mathbf{x} = x_1, \ldots, x_n$ characterized by a set of unknown parameters ϑ is distributed among H different nonoverlapping groups with probability π_h of being in group h (such that $\sum_{h=1}^{H} \pi_h = 1$). Thus, the density function describing the distribution of vector $\mathbf{x}$ is given by:

$$f(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\vartheta}) = \sum_{h=1}^{H} \pi_h f(\mathbf{x}|\vartheta_i). \tag{1}$$

Bayesian estimation using MCMC methods is based on specifying a latent class model that associates each x_i with a latent variable $\zeta_i \in \{1, \ldots, H\}$. The latent class variable ζ_i is sampled from a categorical distribution, i.e. $\zeta_i \sim \text{Categorical}(H, \boldsymbol{\pi})$, resulting in the likelihood defined as:

$$f(x_i, \zeta_{hi}|\pi_h, \vartheta_{hi}) = \prod_{h=1}^{H} \left[\pi_h f(x_i|\vartheta_{hi})\right]^{\zeta_{hi}}. \tag{2}$$

This leads to the joint posterior

$$p(\pi, \vartheta|x, \zeta) \propto p(x, \zeta|\pi, \vartheta) p(\pi)\, p(\vartheta), \tag{3}$$

from which a Gibbs sampler can be implemented by setting priors for ϑ and π (see section 2.3.2 for details). The hierarchical structure of the occupancy model allows for incorporation of the latent variable $\zeta_i = h$ for those species belonging to class $h \in (1, 2, \ldots, H)$. By grouping species in classes, the relative frequency of the number of species in each class can be expressed as a proportion of the total number of species estimated to be present at each site (see equation (7)).

In a standard multispecies occupancy model, species-level parameters are drawn from a common distribution characterized by hyperparameters representing the parameter's average value among all the species in a community (and also the scale parameters for normal hyperpriors) (Dorazio & Royle, 2005; Dorazio et al., 2006). Typically, logit-normal models are used to describe species heterogeneity in occupancy and detection probabilities (Coull & Agresti, 1999; Dorazio & Andrew Royle, 2003; Royle, 2006). In addition to the species heterogeneity described by the logit-normal model, adding the finite mixture component enables occupancy parameters of the ecological process to be defined based on the latent classes to which each species belongs. This allows the composition of a community to be characterized in terms of how common or rare the species are.

Thus, a multiple-species occupancy model can be formulated as follows:

$$\begin{aligned} \zeta_i &\sim \text{Categorical}\left(H, \pi_h\right) && (a)\\ z_{ij}|\zeta_i, \psi &\sim \text{Bernoulli}\left(\psi_{hi}\right) \end{aligned}$$

$$\sum_{K_j} y_{ij\cdot}|z_{ij}, \zeta_i, p_i \sim \text{Binomial}\left(K_j, p_i z_{ij}\right) \quad (b)$$

$$\begin{aligned} \text{logit}\left(\psi_{hi}\right) &\sim \text{Normal}\left(\mu_{\psi_h}, \sigma^2_{\psi_h}\right) && (c)\\ \text{logit}\left(p_i\right) &\sim \text{Normal}\left(\mu_p, \sigma^2_{p_i}\right) \end{aligned} \tag{4}$$

Equation 4 (a) denotes the ecological/state process where the latent variable ζ is a categorical random variable relating the i-th species' occupancy (z_{ij} for $i = 1, \ldots, S$ species and $j = 1, \ldots, M$ sites) given by the occupancy probability ψ_{hi} for a specific class h such that $\Pr(\zeta = h) = \pi_h$ for $h = 1, \ldots, H$ and $\pi = \pi_1, \ldots, \pi_H$, allowing for the class structure in the community to be estimated by linking each

species-level parameter to a latent class from the finite mixture component, so that species are grouped in terms of their rarity.

The observational process driven by the detection probability p_i (Equation 4 (b)) can then be formulated as the total number of times species i was detected at site j across K_j visits (where the number of visits are defined by the different sampling occasions in which a species was recorded). This model was developed under the assumption that species' occupancy and detection probabilities are constant through time and space unless site- and time-varying covariates are specified. (See section 2.3.2 for details on how this assumption could be relaxed). Finally, the species heterogeneity model (Equation 4 (c)) is characterized by the individual species logit-scaled occupancy and detection probabilities drawn from the same normal prior distribution with hyperparameters describing the overall community response. It is important to notice that logit-scale Normal distributions can overlap to different degrees depending on the mean and variance of each class, which affects the performance of the model in correctly identifying the members of each latent class (Sollmann et al., 2021). Issues with the model's performance can also arise if the density mass is heavily skewed due to an imbalanced number of observations allocated to certain classes. Thus, in section 3, we present a simulation study to investigate the effect of different degrees of overlapping and class imbalance on the model's performance.

Mixture model parameters also suffer from a lack of identifiability due to the invariance of the posterior distribution to permutations of the group labels (Redner & Walker, 1984; Richardson & Green, 1997). To make model parameters identifiable and avoid label-switching issues, component mean values can be ordered, i.e. $\mu_{\psi,1} < \mu_{\psi,2} < \ldots < \mu_{\psi,H}$.

By tracking ζ on each MCMC sample draw s, then, the posterior probability of each species being classified into the h-th category is given by:

$$\Pr(\text{species i belongs to h}) = \Pr(\zeta_i = h|\pi, z, \psi)$$

$$\approx \frac{1}{S}\sum_{s}^{S} \mathrm{I}\left(\zeta_i^{(s)} = h\right), \tag{5}$$

where $\mathrm{I}\left(\zeta_i^{(s)} = h\right)$ denotes an indicator variable that takes the value of one if species i belongs to cluster h and zero otherwise. Species can be clustered by assigning them to a class based on these posterior probabilities as follows:

$$\text{Species } i \text{ belongs to } h = \text{argmax}_h \Pr(\zeta_i = h|\pi, z, \psi) \tag{6}$$

Additionally, the relative class frequency at each site (η_{hj}), expressed as a proportion of the local species richness (i.e. $\sum_i z_{ij}$), can be computed as a derived quantity by tracking the h-th class-membership and the occupancy status for every species on each MCMC draw:

$$\eta_{hj} = \sum_i z_{ij}^{-1} \sum_i z_{ij} \mathrm{I}\left(\zeta_i = h\right), \text{ such that } \eta_{hj} \in [0, 1]. \tag{7}$$

Note that a finite mixture density could also be included for the detection probabilities in the observational model by fitting a multivariate density for the joint distribution of (ψ, p) by using inverse Wishart priors for the covariance-variance matrix. However, we found identifiability issues in some of our simulations due to label-switching among clusters. (Similar issues were reported by Sollmann et al. (2021) when using infinite mixtures for modelling the joint distribution of occupancy parameters, especially when variation among clusters was high). Thus, in this work, we have specified finite mixtures for the state process only because our primary goal is to identify species based on their rarity rather than their elusiveness.

Moreover, by adopting a Bayesian inference approach, the proportion of species in each class can be computed as a derived quantity of the occupancy model at each of the sampled sites in the study. In the Odonata case study, this enables inference about the occupancy pattern of rare species to be limited to only the sites in the sample. This is of particular interest since these sites represent lakes and ponds, which are key components of the hydrological network in the United Kingdom.

2.3.2 | Introducing species-specific effects and site-level covariates

The Odonata case study contains information about species-specific traits that could be associated with the species detection. Thus, these species-specific effects can be specified as a linear model for the detection hyperparameters (Eqn. 4 (b)) as follows:

$$\mu_p = \gamma_0 + \sum_{m=1}^{T} \gamma_m w_{im}. \tag{8}$$

Here, the mean detection probability μ_p of each species is a linear function of $T = 3$ regression terms γ_m associated with each trait covariate w_{im} and an intercept γ_0 representing the baseline detection probability.

The mixture occupancy model described in Eqn.(4) has great flexibility and can be adapted to different scenarios depending on the question of interest. For example, time- and space-varying occupancy and detection probabilities can be incorporated through the logit function in Eqn. 4 (c) by either including a site and year random effects (see Outhwaite et al. (2018)) or by having distinct site-level predictors that affect species occupancy and detection (Wintle & Bardos, 2006) (e.g. $\text{logit}\left(\psi_{hij}\right) = \beta_{0hi} + \sum_{g=1}^{L} \beta_{ghi} x_{gj}$ where logit-scaled occupancy probabilities are defined as a function of L site-level covariates $x_{1j}, \ldots, x_{Lj}$ and $\beta_{0hi}, \beta_{1hi}, \ldots, \beta_{Lhi}$ species-specific terms for each class h).

A similar approach by Dunstan et al. (2011), implemented finite mixture models in a frequentist setting to capture heterogeneity in species responses to environmental gradients among different latent classes. However, by adopting a Bayesian inference approach, we can retain the latent variables while accommodating imperfect detection. Moreover, the hierarchical structure of the occupancy model enables

the information among the species in the community to be shared within each class. Therefore, by assuming all species within a particular class (e.g. rare species) are related to one another by being part of the same biological community, the parameters of species with sparse occurrence records can be estimated (Dorazio et al., 2011).

To estimate the parameters of the occupancy mixture model, Dirichlet conjugate priors are specified for π (i.e. $p(\pi) \propto \prod_i \pi_h^{\alpha_{h-1}}$) such that the posterior is sampled from $\pi|\zeta \sim \pi\text{Dirichlet}, \left(\sum_i \zeta_{1,i} + \alpha_1, \ldots, \sum_i \zeta_{H,i} + \alpha_H\right)$, where $\sum_i \zeta_{h,i}$ is the number of species assigned to each class. Vague normal priors, logistic(0,1) or weakly informative zero-centered t-distributed priors with scale parameter of 1.566 and degrees of freedom 7.763 can then be specified for the mean hyperparameters and inverse-gamma (conjugate prior), Uniform(0,5) or Half-Cauchy priors for the variance hyperparameters (Outhwaite et al., 2018). (See Northrup and Gerber (2018) for a detailed discussion on prior specifications).

2.3.3 | Fitting a two-class finite mixture

Note that for this work, we will be addressing a binary classification problem only, since our aim is to distinguish rare from common species, but the method can easily be generalized to multiclass problems. We work under the assumption that the number of classes is fixed and known before conducting the analysis. (The choice of the number of classes is based on the ecological context of the problem only). For a two-class problem, the likelihood in (2) can be simplified to

$$f(x_i) = \pi f(x_i|\vartheta_{1i}) + (1-\pi) f(x_i|\vartheta_{2i}), \tag{9}$$

allowing Beta (a_1, a_2) priors to be specified for π.

For this work, the sensitivity of the priors was tested by comparing our model results under the different aforementioned prior parametrizations. Our results were consistent when either logistic(0,1) and zero-centered t($\sigma = 1.566, \nu = 7.763$) distributed priors were specified for the mean hyperparameters. Specifying such priors instead of vague normal priors avoids the need for calibrating the precision parameter of the normal prior, which can often be a problem when a logit-scale transformation is used, as vague normal priors (e.g. Normal [0500]) lead to a high probability density around zero and one on a probability scale. Moreover, Uniform(0,5) and Half-Cauchy priors for the variance hyperparameters showed an overall better mixing and lower autocorrelation of the MCMC chains compared with inverse-gamma priors. Finally, we also specified Dirlichet (10,10) priors for the mixing parameters. Graphical diagnostics for convergence of our analysis are available in the Appendix S1.

For each analysis, we ran a total of 50,000 iterations with a burnin period of 10,000 and a thinning of 10 (for memory optimization purposes only) on three independent Markov chains (approximate run time <30 min). The algorithm's convergence was assessed through conventional graphical diagnostics (i.e. traceplots of the posterior samples showing overall good mixing with low posterior autocorrelation, and Gelman-Rubin between-within chains variance ratio <1.1 (Gelman & Rubin, 1992)). All of our modelling work and data manipulation was implemented in R version 4.0.0 (R Core Team, 2020). Our models were run in R Nimble (de Valpine et al., 2017).

3 | SIMULATION STUDY

We designed a simulation study to test the performance of our occupancy mixture model in which four distinct species classes were defined to simulate a community made up of a combination of common, rare, elusive and nonelusive species. We tested the ability of our model to differentiate between common and rare species under varying occupancy, detection and mixing probabilities, i.e. we assessed if the proposed model could identify those elusive common species that could be mislabeled as rare due to their low detection probabilities.

To generate the aforementioned community, a total of $S = 50$ species were simulated in $M = 300$ sites visited on 4 different occasions. Species-specific occupancy and detection probabilities were drawn from a multivariate normal distribution as follows:

$$\Omega \sim \sum_{h=1}^{4} \pi_h \text{Normal}\left(\mu_h = \begin{pmatrix}\mu_{\psi h}\\ \mu_{ph}\end{pmatrix}, \Sigma_h\right), \tag{10}$$

where $\text{logit}^{-1}(\Omega) = (\psi_{hi}, p_{hi})$ is the species-specific occupancy and detection probabilities with π_h being the mixing probabilities that determine the proportion of species belonging to each class. These species-specific parameters were drawn from the community logit-scaled baseline occupancy and detection probabilities μ_{ψ_h} and μ_{p_h} respectively. Note that the simulated occupancy and detection probabilities ranged between 0.05 and 0.95 across species. This captures a reasonably wide span of different species responses that matches what we observed in the Odonata case study.

Σ_h is the covariance-variance matrix for the h-th class with diagonal elements $\left(\sigma^2_{\psi_h}, \sigma^2_{p_h}\right)$ corresponding to variances for the logit-scaled community mean occupancy and detection probabilities respectively. The off-diagonal elements of the covariance-variance matrix Σ_h were set to zero to remove the abundance-induced detection effect (i.e. when detection probabilities are influenced by the high abundance of widespread species), which is assumed to be captured by specifying the number of different habitats that each species occupies (Eqn. 8). Figure 1 illustrates the general framework used to simulate the three following scenarios (details of each scenario can be found in the Appendix S1):

3.1 | Simulation scenario 1: nonoverlapping with constant variance and proportional allocation

First, four well-separated classes were simulated by specifying a reasonable distance between the mean value μ_h of each group and

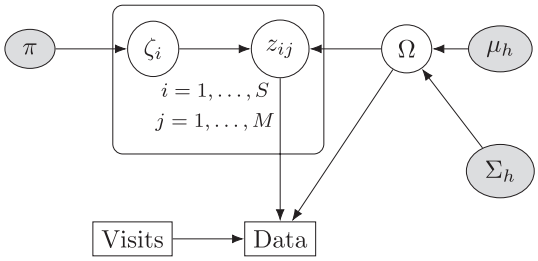

FIGURE 1 Direct acyclic graph illustrating the simulation scheme under varying stochastic parameters indicated by the shaded nodes. The square box represents the model's latent state process for S=50 species and M=300 sites. Square nodes represent simulated data while circle nodes are the stochastic parameters varying for simulation

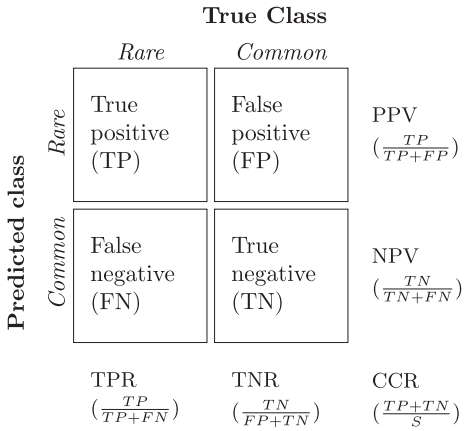

FIGURE 2 Confusion matrix and standard classification metrics for a two-class problem. TPR denotes the sensitivity or true positive rate, TNR denotes the specificity or true negative rate, PPV and NPV denote the positive and negative predictive values, respectively, and CCR denotes the correct classification rate computed over the total number of species S

constant variance ($\Sigma_h = \Sigma \forall h$) while retaining the same proportion of species assigned to each class (i.e. $\pi_h = 1/4 \forall h$).

3.2 | Simulation scenario 2: moderate overlapping with variance heterogeneity under constant and varying mixing probabilities

For a second simulation, we allowed for a moderate degree of overlapping between rare/nonelusive species and common/elusive species while specifying different variances for each class (see Appendix S1 for the specification of variance heterogeneity). Furthermore, we explored the model performance under constant and varying mixing probabilities such that the number of species in common and elusive classes was larger than the number of species in the rare and nonelusive classes respectively.

3.3 | Simulation scenario 3: strong overlapping with variance and mixing heterogeneity

For the third scenario, a greater degree of overlapping was induced by setting (i) similar detection and occupancy probabilities for each class, (ii) variance heterogeneity and (iii) different proportions of species allocated to each class. The mixing coefficients and the variance-covariance matrix were the same as those described for scenario 2.

3.4 | Assessing model performance

Model 4 was fitted to the simulated data for each scenario using the settings described in section 2.3.3. The model performance was assessed by using standard metrics calculated based on the confusion matrix shown in Figure 2. The diagonal elements of the confusion matrix are given by the number of rare and common species that have been correctly predicted as such (true positives (TP) and negatives (TN) respectively). The off-diagonal elements of the confusion matrix contain the false positives (FP $=\Pr(\hat{\zeta}_i = 2|\zeta_i = 1)$) and false negatives (FN $=\Pr(\hat{\zeta}_i = 1|\zeta_i = 2)$) errors. The different metrics that were compared are also presented in Figure 2. Here, the correct classification rate (*CCR*) describes the overall proportion of species that were classified correctly. Sensitivity (*TPR*) and specificity (*TNR*) denote the proportion of rare/common species that have been classified correctly among the rare and common classes respectively. The model's precision or positive predictive value (PPV) represents the proportion of truly rare species among all the species that have been predicted to be rare. Similarly, the negative predictive value (NPV) describes how many species predicted to be common are truly common. Cohen's Kappa statistic (κ) was also calculated as a measure of the model's performance relative to what would be expected by chance. According to Fleiss and Cohen (1973), a κ value between 0.40 – 0.60 indicates a moderate performance, while values of 0.61 – 0.80 and 0.81 – 1 suggest a substantial and almost perfect performance respectively. Finally, the F-score was calculated as $F = 2 \times (TPR \times PPV) / (TPR + PPV)$, and the balanced accuracy ($TPR + TNR/2$) was used as metrics to assess the overall model performance under each scenario. The F-score, on the one hand, accounts for both false-positive and false-negative errors and gives equal importance to sensitivity and precision (i.e. focuses primarily on detecting correctly rare species). The balanced accuracy, on the other hand, is useful when comparing classes with an unbalanced number of observations, as it also considers the true negatives and, thus, gives equal importance to predicting correctly both rare and common species.

4 | RESULTS

4.1 | Simulation study results

Our simulation study results in Table 1 show the standard metrics of classification performance (section 3.1) for each simulated scenario. Overall, our results suggest good model performance in identifying

TABLE 1 Occupancy mixture model classification performance metrics under different simulated scenarios

Scenario	Proportion of species per class	Overlapping	CCR	TPR	TNR	PPV	NPV	κ	Balanced accuracy	F-Score
1	Equal proportion	No	1	1	1	1	1	1	1	1
2	Equal proportion	Moderate	0.96	0.93	1	1	0.92	0.92	0.96	0.96
	More common than rare		0.88	1	0.83	0.70	1.00	0.74	0.92	0.82
	More elusive than nonelusive		0.94	0.92	0.96	0.96	0.93	0.88	0.94	0.94
3	Equal proportion	Strong	0.80	0.64	1	1	0.69	0.61	0.82	0.78
	More common than rare		0.76	0.86	0.72	0.55	0.93	0.49	0.79	0.67
	More elusive than nonelusive		0.78	0.67	0.89	0.84	0.74	0.56	0.78	0.74

those rare and widespread species with an ≈ 80\% of accuracy across all scenarios. For scenario 1, our model correctly classified all rare and common species with respect to the true categories. In scenario 2, the model classification performance was affected by groups with an unbalanced number of species. For instance, with constant mixing probabilities for each class, the CCR was 0.96 and the TNR and PPV indicated that all common species were correctly identified as such, and all the predicted rare species were truly rare. Moreover, the κ statistic (>0.80) suggested an almost perfect performance relative to what can be expected by chance, and the balanced accuracy and F-score values above 0.90 indicated a very good predictive performance. However, when the proportion of common species was greater than the proportion of rare species, the TNR and PPV (0.83 and 0.70 respectively) suggested that 83% of species were classified as common out of the total number of common species because some common species have been predicted as rare (i.e. only 70% of predicted rare species were truly rare). In this situation, the F-score shows a lower value compared to CCR and balanced accuracy because it does not consider the true negatives, unlike the balanced accuracy, which suggests a very good classification performance when either an equal or unequal proportion of species are allocated to each class. On the third scenario, the accuracy decreased to ≈ 80%. The different classification performance metrics were approximately 10%–20% lower than the metrics in scenario 2. We also found that having an unequal number of species for each class yields a much lower accuracy. Particularly, when the number of common species is greater than the number of rare species, both the PPV and TNR suggest an overestimation of the true number of rare species. On the other hand, when the number of elusive species was higher than the number of nonelusive species, the TPR decreased, suggesting that several rare species are being classified as common (only 67% of the species are classified as rare out of the total number of rare species). In summary, the accuracy, F-score and κ statistic suggest a reasonably good performance of the mixture occupancy model as a classifier when there is an balanced number of observations for each class. When there is an unbalanced number of observations, the balanced accuracy suggests a very good performance of our model for scenarios 1 and 2 and a moderate performance under scenario 3. Particularly, our model performance under scenario 3 depends on the different class proportions, i.e. a greater number of common species yield an overestimation of the true number of rare species, while having more elusive and nonelusive species results in an underestimation of rare species.

4.2 | Odonata case study results

We fitted the proposed mixture occupancy model (Equations 1 and 8) to quantify the proportion of rare species of Odonata across freshwaters in the United Kingdom. The proportion of rare species at each site was calculated according to Eqn. 7. Figure 3 shows the predicted classes for each of the Odonata species, with approximately half of the species being categorized as rare. Note that the credible intervals show that the uncertainty for detection probabilities is larger for rare species than for common species. However, at high levels of occupancy, the uncertainty of the estimated occupancy for widespread species becomes larger than for rare species. This pattern could be related to the distribution of occupancy being right-skewed (a small number of common species show relatively high-occupancy probabilities >0.75), so that uncertainty around the estimates of species with high-occupancy probability will be larger than estimates for species with low occupancy. In addition, when the occupancy is low, we have less information from which the probabilities of detection are estimated resulting in greater uncertainty around these estimates.

The proportion of rare species across the different sites is shown in Figure 4. Overall, the proportion of rare species and its estimated standard deviation are low and homogeneous across space. There are, however, some areas in the north and south west where the proportion of rare species is greater than the proportion in central areas, which are dominated almost exclusively by widespread species. These proportions were calculated with a reasonably small error (SD <0.2 for most sites), though there are some sites where uncertainty is larger possibly due to the very low number of Odonata occurrence records at those sites (see Figure S1).

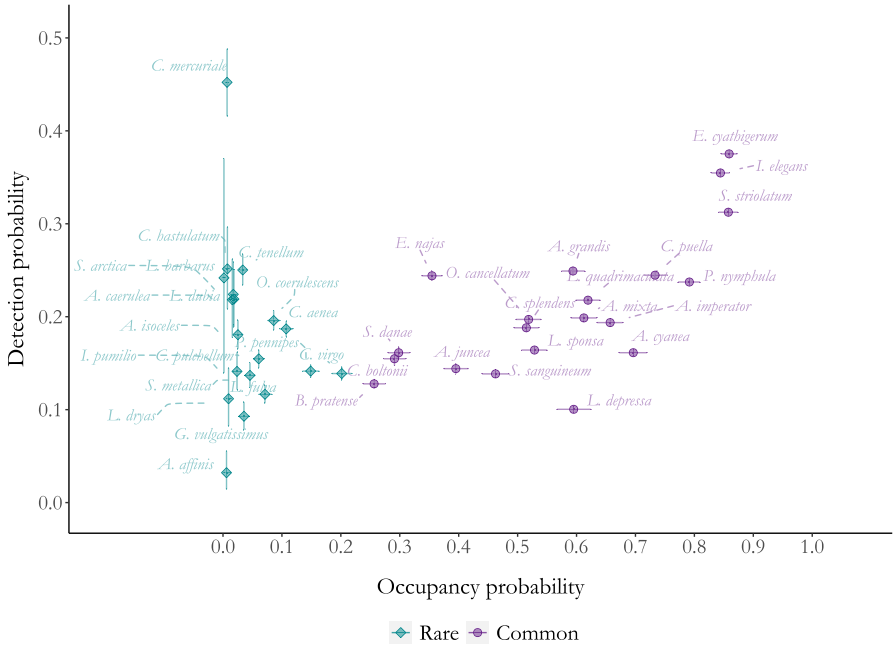

FIGURE 3 Estimated occupancy, detection probabilities and predicted class for the 39 British Odonata species in our study. Error bars represent 95% credible intervals for the corresponding individual species occupancy/detection probability

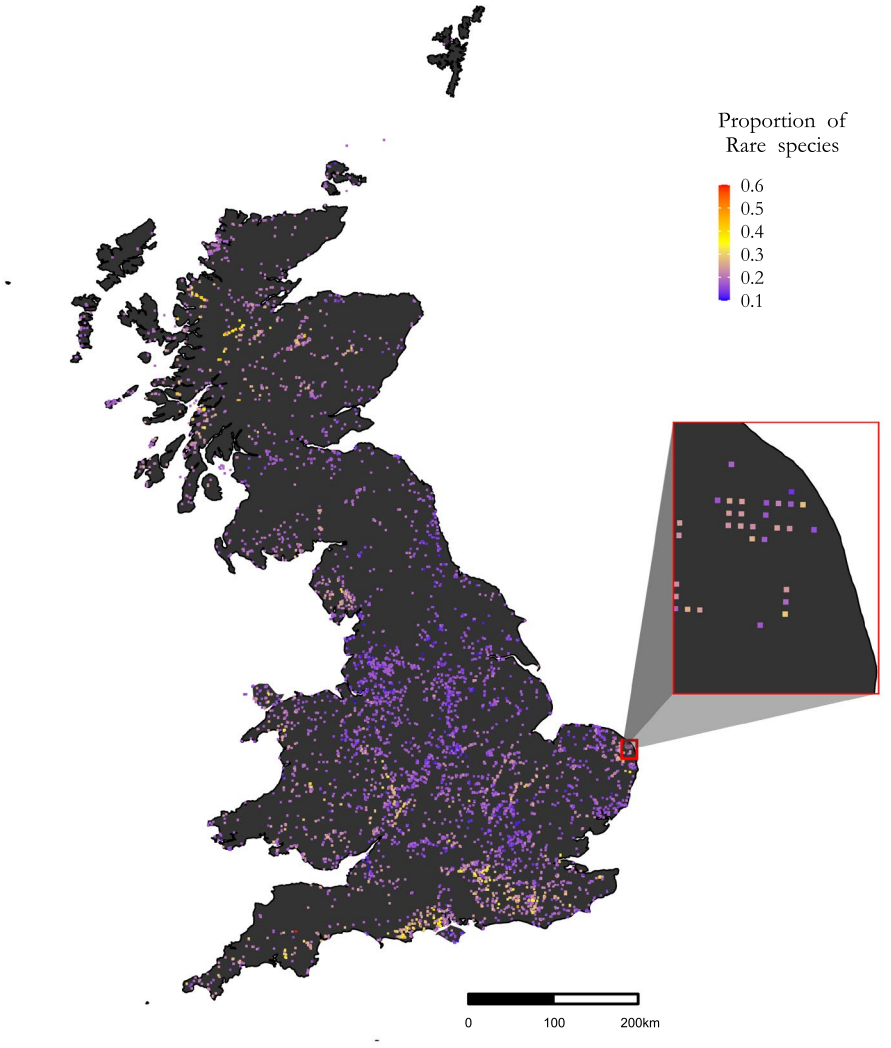

FIGURE 4 Posterior mean of the proportion of rare species at each site across the United Kingdom where each location represents a freshwater body associated with a catchment within the hydrological network

5 | DISCUSSION AND CONCLUSION

In this work, we have presented a new approach to characterize the composition of a community in terms of how rare or common species are. Our model differs from previous methods by considering imperfect detection. Moreover, our modelling framework enables using standard designs of occupancy models for which information about the species' distribution ranges, population densities or habitat specificity are not required to estimate species distributions. In fact, any species-specific data of this nature could be incorporated in either or both latent state and observational processes of the model as described in section 2.3.2.

We constructed our model under the assumption that the species distributions remain constant across different sampling occasions. We used the single-season occupancy model as the starting point based on some of our preliminary research that suggests that the distribution of Odonata species shows little evidence of major temporal changes within the time period of this study (2000–2016). However, any extension to the occupancy model could potentially be developed within our mixture modelling approach. For example, our model could be integrated with an explicit spatiotemporal occupancy model (see Rushing et al. (2019)) to estimate the temporal changes in occupancy that a species of a specific class experiences. This could be potentially useful for monitoring the distributions and temporal dynamics of invasive species at different stages after their introductions and to identify the tipping point when a focal species experiences a major increase or reduction in its distribution given by the change in its latent class.

The simulation analysis of our mixture occupancy model showed an overall good classification performance, specifically when the proportion of individual species is similar between classes. However, when the proportion of common species was set to be greater than the proportion of rare species, a larger proportion of common species were classified as rare. Under this unbalanced scenario, the posterior mass density from which the occupancy community parameters are drawn will be greater for common than for rare species. Thus, uncertainty around the occupancy community mean for rare species will be larger and pulled towards the density mass of common species, resulting in some common species being classified as rare, yet producing precise predictions for the common species (i.e. the sensitivity and the proportion of predicted common species that are truly common will be high). Note that for this simulation analysis, the large difference in the proportions of species belonging to each class could very well portray a real-life situation where biological communities are composed primarily of widespread species. Thus, to avoid the overestimation or underestimation of rare species caused by unbalanced classes, informative Dirichlet priors (or beta priors for a two-class classification problem) could be specified to reflect our prior judgement about the proportion of species of each class determining the community structure.

It is important to notice that using an informative prior causes the classification of species rarity to change and, hence, introduces a risk of underestimating the number of truly rare species if the prior weight is greater for common species, or overestimating the rare species if the prior assigns more weight to this class. However, as mentioned by Leroy et al. (2012), it is preferable to conduct an analysis to characterize species rarity rather than ignore it and thus potentially overlook hot spots relevant for conservation. Therefore, the choice of priors depends on (1) the previous knowledge about the target species distribution and (2) the conservation targets. For instance, if the conservation study focuses on rare species, more weight could be assigned to the prior of a species being classified as rare to ensure that most of the species will be correctly classified as rare.

Our approach enables the class membership for multiple species based on a fixed number of classes to be estimated. If the number of classes is unknown, a Bayesian nonparametric mixture model involving Dirichlet processes could be specified (Hjort et al., 2010). For instance, Johnson and Sinclair (2017) implemented a reversible jump MCMC algorithm (RJMCMC; Richardson and Green (1997)) to make Bayesian inference on a multiple-species distribution model that uses Chinese Restaurant Process for clustering species into guilds. RJMCMC allows for the number of components to vary between iterations of the Markov chain. Unfortunately, selecting appropriate proposal densities is challenging and difficult to implement (Nasserinejad et al., 2017). Recently, Sollmann et al. (2021) proposed a truncated stick-breaking Dirichlet process to describe the community structure in terms of the estimated classes and the similarity in species responses to environmental conditions. However, Sollmann et al. (2021) reported an overestimation of the number of classes and poor mixing due to label-switching issues attributed to high dimensionality and variability among clusters. In our work with a two-class finite mixture, we did not find such issues because of the small number of fixed classes (even under moderate heterogeneity among clusters).

Having a set of candidate models, each with a small number of predefined classes, makes the problem of selecting an appropriate number of classes more tractable and easier to implement. For example, WAIC (Watanabe-Akaike information criteria; Watanabe and Opper (2010)) can be used to compare and select models with a different number of fixed groups. WAIC is fully Bayesian criteria computed based on (1) the log-pointwise predictive density, which is an average of the model fit across all the posterior draws and (2) the variance of the log-likelihood across MCMC samples representing the number of effective parameters. An alternative approach that has proved to be efficient for identifying the number of classes in Bayesian finite mixture models is based on specifying an overparametrized model that converges to the true number of components if the α_h parameters of the Dirichlet prior are less than half of the number of class-specific parameters (Rousseau & Mengersen, 2011). By specifying such vague priors, the overfitted classes will asymptotically become empty during MCMC sampling and can then be discarded to find the true number of groups. This approach can be easily implemented in JAGS or Nimble, and extensions to this approach can be found in Malsiner-Walli et al. (2016) and in Nasserinejad et al. (2017).

In the case study analysed in this work, the proportion of rare Odonata species is generally low (below 20%) and homogeneous across all of the region. There are, however, some areas in the north of Scotland and in the south of England where the proportion of rare species is above 50%. This could be explained by the occurrences of species at very local scales such as *Coenagrion hastulatum* and *Somatochlora arctica*, which are confined to a few particular lakes in Scotland and south west Ireland (source: https://british-dragonflies.org.uk). Consequently, the distribution patterns for some of these rare species might not be evident on a national scale and could be limited to specific habitats where species are exposed to different environmental conditions and extinction processes (Hartley & Kunin, 2003). Hence, because of the occurrence-based approach implemented here, where the proportion of rare species on a national scale is derived from the relative mean occupancy probabilities in each class, the estimation of rarity depends on the study spatial scale. Previous studies by Hartley and Kunin (2003) and Leroy et al. (2013) have shown how species rarity can be a scale-dependent process and have proposed different multiscale metrics to quantify the species rarity. However, producing such metrics under imperfect detection is a lesser studied subject for small invertebrates (Leroy et al., 2012, 2013) and a very interesting area for future research. For example, our modelling framework could be integrated with multiscale occupancy modelling designs and sampling schemes like the one proposed by Nichols et al. (2008) to estimate the proportion of rare species at varying spatial scales under imperfect detection.

Other occurrence-based methods that assess species rarity (such as the Leroy et al. (2012) rarity index) rely greatly on the sampling methods, as uneven sampling effort might induce an artificial rarity for those species recorded in poorly sampled sites. Accounting for sampling effort and detection probabilities in the observation process is of major importance, specifically when working with large citizen-science data sets where observations are collected without a standardized field sampling protocol for a frequently arbitrary selection of sites. This variation in observation effort causes detection probabilities to vary over time and space (K´ery et al., 2010). The two-component structure in our model allows us to incorporate terms that correct for uneven sampling effort. In our models, we only used the number of visits at each site as a proxy for the sampling efficiency. Thus, the correction for uneven sampling effort occurs automatically through the number of sampling days representing the number of Bernoulli trials in the distribution of p rather than as a covariate in the model. Alternatively, a common approach to the analysis of citizen-science data is to define the logit-scale detection probabilities as a function of (1) the day when each site was visited and (2) the daily species lists to account for heterogeneity in detection probabilities caused by seasonal variation in the flying periods of species and unequal observation effort (K´ery et al., 2010; van Strien et al., 2010, 2013).

The approach presented in this work provides a new method to understand the species rarity assemblage in a community when species are detected imperfectly. It can be potentially applied to different situations such as multiscale studies and spatiotemporal modelling. We have applied our model on a national scale to estimate the proportion of rare species of Odonata, a taxonomic group (along with other small invertebrates) that has a significant under-representation in studies accounting for imperfect detection (Devarajan et al., 2020; Kellner & Swihart, 2014). We believe this work provides a substantial improvement to how rare species are identified and to the understanding of how species rarity assemblages can be quantified.

ACKNOWLEDGEMENTS

The authors thank Dr. Stephen J. Brooks, Dr. Tom August and the Hydroscape team for their insights and comments. The authors thank NERC funding (grant NE/N005740/1) for making this project possible. JB was supported by CONACyT scholarship (494334).

CONFLICT OF INTEREST

We warrant that this manuscript is the original work of the authors listed and has not previously been published. All the authors listed made substantial contributions to the manuscript and qualify for authorship, and no authors have been omitted. We warrant that none of the authors has any conflict of interest with regard to this manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13495.

DATA AVAILABILITY STATEMENT

All data sources used for this work are publicly available and have been properly cited in the manuscript. Odonata occurrences data are available from the British Dragonfly Society Recording Scheme (2020). doi: https://doi.org/10.15468/cuyjyi. Odonata species traits data from Powney et al. (2014) are stored in the online data repository GitHub https://github.com/BiologicalRecordsCentre/Odonata_traits. Nimble code for fitting the Odonata mixture occupancy model can be found in the Supporting Information.

ORCID

Jafet Belmont https://orcid.org/0000-0002-6879-4412

REFERENCES

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, *5*, 1269–1279. https://doi.org/10.1111/2041-210X.12100

Blackburn, T. M., & Gaston, K. J. (1997). Who is rare? Artefacts and complexities of rarity determination. In W. E. Kunin, & K. J. Gaston (Eds.), *The Biology of Rarity* (pp. 48–60). Springer.

British Dragonfly Society Recording Scheme (2020). *Dragonfly records from the British Dragonfly Society Recording Scheme*. https://doi.org/10.15468/cuyjyi

Broennimann, O., Vittoz, P., Moser, D., & Guisan, A. (2005). Rarity types among plant species with high conservation priority in Switzerland. *Botanica Helvetica*, *115*, 95–108. https://doi.org/10.1007/s00035-005-0713-z

Coull, B. A., & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, *55*(1), 294–301. https://doi.org/10.1111/j.0006-341X.1999.00294.x

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical

algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Devarajan, K., Morelli, T. L., & Tenan, S. (2020). Multi-species occupancy models: Review, roadmap, and recommendations. *Ecography*, *43*, 1612–1624. https://doi.org/10.1111/ecog.04957

Dorazio, R. M., & Andrew Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, *59*(2), 351–364. https://doi.org/10.1111/1541-0420.00042

Dorazio, R. M., Gotelli, N. J., & Ellison, A. M. (2011). Modern methods of estimating biodiversity from presence-absence surveys. In O. Grillo, & G. Venora (Eds.), *Biodiversity Loss in a Changing Planet* (pp. 277–302). IntechOpen.

Dorazio, R. M., & Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, *100*(470), 389–398. https://doi.org/10.1198/016214505000000015

Dorazio, R. M., Royle, J. A., Söderström, B. O., & Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, *87*(4), 842–854. https://doi.org/10.1890/0012-9658(2006)87[842:ESRAAB]2.0.CO;2

Dunstan, P. K., Foster, S. D., & Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, *222*(4), 955–963. https://doi.org/10.1016/j.ecolmodel.2010.11.030

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ellstrand, N. C., & Elam, D. R. (1993). Population genetic consequences of small population size: Implications for plant conservation. *Annual Review of Ecology and Systematics*, *24*(1), 217–242. https://doi.org/10.1146/annurev.es.24.110193.001245

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. https://doi.org/10.1177/001316447303300309

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A. Jr, Latimer, A., & Rebelo, A. G. (2005). Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(1), 1–20.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Golfieri, B., Hardersen, S., Maiolini, B., & Surian, N. (2016). Odonates as indicators of the ecological integrity of the river corridor: Development and application of the odonate river index (ORI) in northern Italy. *Ecological Indicators*, *61*, 234–247. https://doi.org/10.1016/j.ecolind.2015.09.022

Hartley, S., & Kunin, W. E. (2003). Scale dependency of rarity, extinction risk, and conservation priority. *Conservation Biology*, *17*(6), 1559–1570.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, Vol. *28*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802478

Isaac, J. L., Vanderwal, J., Johnson, C. N., & Williams, S. E. (2009). Resistance and resilience: Quantifying relative extinction risk in a diverse assemblage of Australian tropical rainforest vertebrates. *Diversity and Distributions*, *15*, 280–288. https://doi.org/10.1111/j.1472-4642.2008.00531.x

Johnson, D. S., & Sinclair, E. H. (2017). Modeling joint abundance of multiple species using Dirichlet process mixtures. *Environmetrics*, *28*(3), e2440. https://doi.org/10.1002/env.2440

Karron, J. D. (1997). Genetic consequences of different patterns of distribution and abundance. In W. E. Kunin, & K. J. Gaston (Eds.), *The Biology of Rarity* (pp. 174–189). Springer.

Kellner, K. F., & Swihart, R. K. (2014). Accounting for imperfect detection in ecology: A quantitative review. *PLoS One*, *9*, e111436. https://doi.org/10.1371/journal.pone.0111436

Kéry, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., & Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Con- Servation Biology*, *24*(5), 1388–1397. https://doi.org/10.1111/j.1523-1739.2010.01479.x

Lande, R. (1993). Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *The American Naturalist*, *142*(6), 911–927. https://doi.org/10.1086/285580

Lee, A. M., Sæther, B.-E., & Engen, S. (2011). Demographic stochasticity, Allee effects, and extinction: The influence of mating system and sex ratio. *The American Naturalist*, *177*(3), 301–313. https://doi.org/10.1086/658344

Leroy, B., Canard, A., & Ysnel, F. (2013). Integrating multiple scales in rarity assessments of invertebrate taxa. *Diversity and Distributions*, *19*, 794–803. https://doi.org/10.1111/ddi.12040

Leroy, B., Petillon, J., Gallon, R., Canard, A., & Ysnel, F. (2012). Improving occurrence-based rarity metrics in conservation studies by including multiple rarity cut-off points. *Insect Conservation and Diversity*, *5*(2), 159–168. https://doi.org/10.1111/j.1752-4598.2011.00148.x

Maciel, E. A. (2021). An index for assessing the rare species of a community. *Ecological Indicators*, *124*, 107424. https://doi.org/10.1016/j.ecolind.2021.107424

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., & Bailey, L. L. (2005). Improving inferences in population studies of rare species that are detected imperfectly. *Ecology*, *86*, 1101–1113. https://doi.org/10.1890/04-1060

Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, *26*(1–2), 303–324. https://doi.org/10.1007/s11222-014-9500-2

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, (*38*). M. Dekker New York.

Melbourne, B. A., & Hastings, A. (2008). Extinction risk depends strongly on factors contributing to stochasticity. *Nature*, *454*(7200), 100–103. https://doi.org/10.1038/nature06922

Nasserinejad, K., van Rosmalen, J., de Kort, W., & Lesaffre, E. (2017). Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PLoS One*, *12*(1), e0168838. https://doi.org/10.1371/journal.pone.0168838

Nichols, J. D., Bailey, L. L., O'Connell, A. F. Jr, Talancy, N. W., Campbell Grant, E. H., Gilbert, A. T., Annand, E. M., Husband, T. P., & Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, *45*(5), 1321–1329. https://doi.org/10.1111/j.1365-2664.2008.01509.x

Norris, J. III, & Pollock, K. H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, *52*(2), 639–649. https://doi.org/10.2307/2532902

Northrup, J. M., & Gerber, B. D. (2018). A comment on priors for Bayesian occupancy models. *PLoS One*, *13*, e0192819. https://doi.org/10.1371/journal.pone.0192819

Outhwaite, C. L., Chandler, R. E., Powney, G. D., Collen, B., Gregory, R. D., & Isaac, N. J. (2018). Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data. *Ecological Indicators*, *93*, 333–343. https://doi.org/10.1016/j.ecolind.2018.05.010

Pacifici, K., Zipkin, E. F., Collazo, J. A., Irizarry, J. I., & DeWan, A. (2014). Guidelines for a priori grouping of species in hierarchical community models. *Ecology and Evolution*, *4*(7), 877–888. https://doi.org/10.1002/ece3.976

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, *56*(2), 434–442.

Powney, G. D., Brooks, S. J., Barwell, L. J., Bowles, P., Fitt, R. N. L., Pavitt, A., Spriggs, R. A., & Isaac, N. J. B. (2014). Data from: Morphological and Geographical Traits of the British Odonata. *Biological Records Centre Repository*, *2*, e1041. https://doi.org/10.3897/BDJ.2.e1041

R Core Team (2020). *R: A Language and Environment for Statistical Computing [Computer software manual]*. Retrieved from https://www.R-project.org/

Rabinowitz, D. (1981). Seven forms of rarity. In H. Synge (Ed.), *The biological aspects of rare plant conservation* (pp. 205–217). Wiley.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195–239. https://doi.org/10.1137/1026034

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792. https://doi.org/10.1111/1467-9868.00095

Rousseau, J., & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x

Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, *62*(1), 97–102. https://doi.org/10.1111/j.1541-0420.2005.00439.x

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*(4), 835–841.

Rushing, C. S., Royle, J. A., Ziolkowski, D. J., & Pardieck, K. L. (2019). Modeling spatially and temporally complex range dynamics when detection is imperfect. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-48851-5

Sollmann, R., Eaton, M. J., Link, W. A., Mulondo, P., Ayebare, S., Prinsloo, S., Plumptre, A. J., & Johnson, D. S. (2021). A Bayesian Dirichlet process community occupancy model to estimate community structure and species similarity. *Ecological Applications*, *31*(2), e02249. https://doi.org/10.1002/eap.2249

Termaat, T., van Strien, A. J., van Grunsven, R. H., De Knijf, G., Bjelke, U., Burbach, K., Conze, K. J., Goffart, P., Hepper, D., Kalkman, V. J., Motte, G., Prins, M. D., Prunier, F., Sparrow, D., van den Top, G. G., Vanappelghem, C., Winterholler, M., & WallisDeVries, M. F. (2019). Distribution trends of European dragonflies under climate change. *Diversity and Distributions*, *25*, 936–950. https://doi.org/10.1111/ddi.12913

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*(6), 1790–1801. https://doi.org/10.1890/02-5078

van Strien, A. J., Termaat, T., Groenendijk, D., Mensing, V., & Kéry, M. (2010). Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic and Applied Ecology*, *11*(6), 495–503. https://doi.org/10.1016/j.baae.2010.05.003

van Strien, A. J., van Swaay, C. A., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, *50*(6), 1450–1458. https://doi.org/10.1111/1365-2664.12158

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.

Wintle, B., & Bardos, D. (2006). Modeling species–habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, *16*(5), 1945–1958.

Yu, J., & Dobson, F. S. (2000). Seven forms of rarity in mammals. *Journal of Biogeography*, *27*(1), 131–139. https://doi.org/10.1046/j.1365-2699.2000.00366.x

BIOSKETCH

The research team's main interests are in developing and applying statistical methods in environmental and ecological sciences. Specifically, the authors work focuses on the development of environmental indicators to quantify the state of the environment, nonparametric methods for spatiotemporal data, design of monitoring networks and species distribution modelling for citizen data projects.

Author contributions: CM and MS directed the research, reviewed the data and the manuscript critically and directed revisions. JB conducted the research, analysed the data and drafted the manuscript. CW reviewed the results and the manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Belmont, J., Miller, C., Scott, M., & Wilkie, C. (2022). A new statistical approach for identifying rare species under imperfect detection. *Diversity and Distributions*, 28, 882–893. https://doi.org/10.1111/ddi.13495