Received: 12 December 2022 | Revised: 21 August 2023 | Accepted: 26 October 2023

DOI: 10.1002/eap.2941

ARTICLE

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

Ecologically informed priors improve Bayesian model estimates of species richness and occupancy for undetected species

Emily M. Beasley

Department of Biology, University of Vermont, Burlington, Vermont, USA

Correspondence
Emily M. Beasley
Email: beasley.em@gmail.com

Present address
Emily M. Beasley, Département de sciences biologiques, Université de Montréal, Montréal, Quebec, Canada.

Funding information
American Museum of Natural History Theodore Roosevelt Memorial Grant

Handling Editor: Juan C. Corley

Abstract

Detection error can bias observations of ecological processes, especially when some species are never detected during sampling. In many communities, the probable identity of these missing species is known from previous research and natural history collections, but this information is rarely incorporated into subsequent models. Here, I present prior aggregation as a method for including information from external sources in Bayesian hierarchical detection models. Prior aggregation combines information from multiple prior distributions, in this case, an ecologically informative, species-level prior, and an uninformative community-level prior. This approach incorporates external information into the model without sacrificing the advantages of modeling species in the context of the community. Using simulated data supplied to a multispecies occupancy model, I demonstrated that prior aggregation improves estimates of (1) metacommunity richness and (2) environmental covariates were associated with species-specific occupancy probabilities. When applied to a dataset of small mammals in Vermont, prior aggregation allowed the model to estimate occupancy correlates of the Eastern cottontail *Sylvilagus floridanus*, a species observed at several sites in the region but never captured. Prior aggregation can be used to improve the analysis of several important metrics in population and community ecology, including abundance, survivorship, and diversity.

KEYWORDS

Bayesian models, detection error, hierarchical models, informative priors, prior aggregation

INTRODUCTION

Estimates of biodiversity and other population and community metrics are often biased due to observer error. Biases or errors, especially detection errors, can be introduced by characteristics of the target species, study design, or observer (Iknayan et al., 2014; Kellner & Swihart, 2014). When species richness or species occupancy is of interest, detection error results in richness and occupancy estimates that are biased low (Benoit et al., 2018; Iknayan et al., 2014) and adds "noise" to the data in the form of false negatives, making it more difficult to evaluate the importance of environmental covariates (Gu & Swihart, 2004). While good study design can reduce survey bias (Banks-Leite et al., 2014) and statistical methods such as the Chao index (Chao, 1984) or bootstrapping (Burnham & Overton, 1979) can correct species richness counts, optimal study design is not always feasible, and traditional statistical estimates are biased when detection rates vary spatially or when the community contains many rare species (New & Handel, 2015).

More recent approaches to account for detection error include hierarchical occupancy models (MacKenzie et al., 2002), specifically the multispecies occupancy model (MSOM). MSOMs yield less biased estimates than traditional methods by jointly analyzing an ecological model of occurrence and an observation model of detection. This strategy allows the model to explicitly differentiate between the true state of the ecological metric and detection error (Dorazio & Royle, 2005; Iknayan et al., 2014; New & Handel, 2015). In a Bayesian framework, MSOMs are also able to efficiently model data-poor species, either by assuming that all species are ecologically comparable (Link & Sauer, 1996) or by using informative priors with information drawn from sources such as previous studies or natural history collections (McCarthy & Masters, 2005). However, the structure of MSOMs often renders these two approaches incompatible. This paper presents a method that combines the above approaches to model rare or undetected species with little to no associated data.

Unlike single-species models, MSOMs assume that species in the community are ecologically similar (i.e., exchangeable) such that species-level parameters can be drawn from a common prior distribution (Figure 1a). In other words, all species are analyzed in the context of the full community. A community-level approach means that rare or hard-to-detect species, which may not yield sufficient data to model individually, can be analyzed by "borrowing" information from common species (Ferrier & Guisan, 2006; Link & Sauer, 1996). Although "borrowing" information results in estimates of rare species that are closer to the community mean (Iknayan et al., 2014; Kéry & Schaub, 2011), MSOMs yield more accurate and more precise estimates of rare species than single-species models.

The use of community-level distributions in Bayesian MSOMs also implies that species that are known to occur in the study region, but were never detected during sampling, can be included in the model using a method called data augmentation (Royle et al., 2007; Royle & Dorazio, 2012; Figure 1b), in which a series of zeros are appended to the original data set to represent species in a community that may have been undetected (Royle et al., 2007). Data augmentation yields reliable community-level estimates when model assumptions are not severely violated and few species in the community are missed (Guillera-Arroita et al., 2019). However, the lack of data for augmented species means that estimates of occupancy or covariate responses for undetected species are inevitably "pulled" toward the center of the community distribution (Link & Sauer, 1996). In practice, this lack of data means the model does not have enough information to accurately estimate specific occupancy or detection probabilities for particular undetected species, so these species can only be used to calculate an asymptotic species richness estimate for the study region (Guillera-Arroita et al., 2019).

In a Bayesian framework, the information needed to estimate specific parameters for particular undetected species can be readily incorporated using an informative species level prior. When priors and data are consistent,

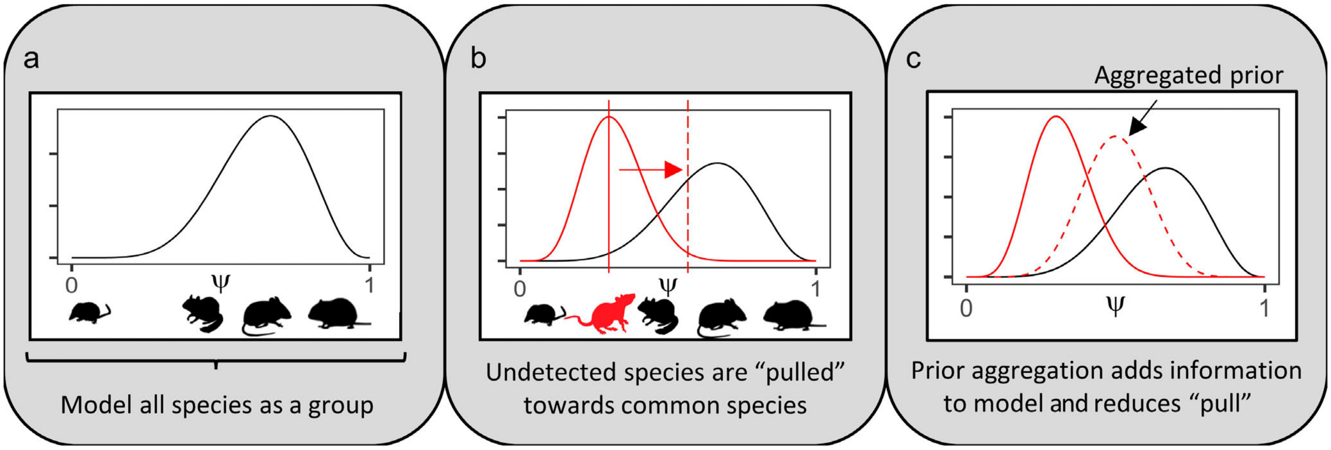
a
0
Ψ
1
Model all species as a group
b
0
Ψ
1
Undetected species are "pulled" towards common species
c
Aggregated prior
0
Ψ
1
Prior aggregation adds information to model and reduces "pull"

FIGURE 1 (a) Community-based detection models account for rare or undetected species by assuming all species-level parameters, such as occupancy probability Ψ, are drawn from a common probability distribution. (b) Species that were never detected during sampling (red) can be analyzed by adding a set of zeros to the data. However, a lack of data means the estimated parameters (dashed red line) are "pulled" to the center of the distribution and away from the true value (solid red line). (c) When the identities of undetected species are known, the hyperprior (black line) can be combined with species-level information (solid red line) to form an aggregated prior (dashed red line) to reduce the "pull" of the hyperprior and more accurately model these species. Silhouette images of species were sourced from PhyloPic with attribution and license information available at https://www.phylopic.org/permalinks/6af9bf48239139440440f6f2e4ea1c641cf9f2b3c5e88de96b2332e9569a612f.

using informative priors in ecological models tends to increase the confidence in conclusions (i.e., narrower credible intervals; McCarthy & Masters, 2005), particularly when data are scarce (Low Choy et al., 2009). In the context of hierarchical detection models, other authors have demonstrated that "weakly informative" priors can be used to stabilize the model and prevent coefficients from taking extreme values (Lemoine, 2019; Northrup & Gerber, 2018), but the use of ecologically informative priors is much rarer. Ecologically informative priors may be rare in part because replacing the uninformative species level prior with an informative prior means that the species is no longer described by the community-level distribution, and the advantages of modeling species in the context of the community are lost.

A potential solution to this problem is prior aggregation applied to species-level parameters for undetected species. Originally used to combine multiple expert opinions (Genest et al., 1984), prior aggregation can be applied to MSOMs to combine the community prior and an ecologically informative prior into a single prior distribution (Figure 1c). With an aggregated prior distribution, the model analyzes undetected species in the context of the community while allowing researchers to retain the identity of each undetected species and reduce the pull of the community prior. However, to my knowledge, prior aggregation has never been used in the context of hierarchical occupancy models and its effects on model performance remain unknown.

Using simulated data with known parameter values, I tested whether ecologically informative, aggregated priors for undetected species improved the estimates of MSOMs. I compared the posterior estimates of metacommunity richness, local species richness, and species-level regression coefficients from MSOMs with (1) nonaggregated uninformative priors, (2) informative priors for undetected species, and (3) misspecified priors for undetected species. I also varied the relative contribution of the informative or misspecified priors to the aggregated prior to determining whether prior strength influenced model estimates. Finally, I applied prior aggregation to an empirical dataset of small mammal communities in Vermont to model occupancy probability and species-level coefficients of an undetected species known to occur in the study region.

METHODS

Data simulation

I simulated 50 metacommunities of $i = 1, 2, \dots 22$ species which potentially occupy $j = 1, 2, \dots, 30$ sites. The first step in the simulation was to generate raw species-level occupancy probabilities, Ψ_i, from a beta distribution (Equation 1) for the first 20 species in the community. The remaining two species, which would represent undetected species that were present in the community, had species-level occupancy probabilities that were fixed at 0.1 and 0.4, respectively, to facilitate comparison across simulations. These probabilities represent relatively rare species, which are more likely to be undetected during sampling due to their low occurrence (MacKenzie et al., 2005). The raw occupancy probabilities were logit transformed to create the latent state of the species-level occupancy intercepts, $\alpha 0_i$, in which i designated the simulated species. Logit-transforming the raw probabilities allows the inclusion of environmental covariates that may influence site-level occupancy:

$$\Psi_i \sim \text{Beta}\,(\alpha = 2, \beta = 4), \tag{1}$$

$$\alpha 0_\text{i} = \text{logit}(\Psi_\text{i}). \tag{2}$$

The second step was simulating site-level occupancy probability as a logit-linear function of species-level intercepts and a continuous covariate (Equation 3). Species in each metacommunity were assigned coefficient values drawn from normal distributions with a mean of 0, 3, or −3, representing no response, a strong positive response, or a strong negative response to the environmental covariate. Coefficients were randomly assigned to the 20 detected species, whereas responses for the undetected species were fixed at 0 and −3. These values were chosen to examine if prior aggregation scenarios differed in their ability to detect both significant (−3) and nonsignificant (0) coefficient values, with the significant coefficient selected as an effect of equal magnitude of the detected species. For each metacommunity, the true occupancy state of each species Z_{ij}, denoted 1 if the species was present at the site and 0 if absent, was modeled as the outcome of a Bernoulli trial with the site-level occupancy probability as the probability of success (Equation 4). Species that did not occupy any site in the initial simulation were assigned a value of 1 to the site with the highest occupancy probability to ensure there were 22 species in each metacommunity simulation:

$$\text{logit}\left(\Psi_{ij}\right) = \alpha 0_i + \alpha 1_i \text{cov}_{j,} \tag{3}$$

$$Z_{ij} \sim \text{Bern}\left(\Psi_{ij}\right). \tag{4}$$

After generating site-level occupancy values, I simulated survey data by generating species-level detection probabilities p_i using a beta distribution $p_i \sim \text{Beta}(\alpha = 2, \beta = 8)$, resulting in low-to-moderate detection probabilities for the 20 detected species (95% interval 0.028–0.482).

Undetected species were assigned a species-level probability of 0. Detection of a species during a survey was modeled as the outcome of a Bernoulli trial with the species-level detection probability as the probability of success, conditional on the species being present at the site. If any of the 20 "detected" species were not detected during any survey, I assigned a value of 1 to the survey with the highest detection probability to ensure there were 20 detected and 2 undetected species in each simulation. Values for the beta distributions were chosen to create a scenario in which data augmentation is effective: when some observed species in the community have low occupancy and/or low detection probabilities, it is more reasonable to expect that some species were absent from all sampled sites or missed during all surveys (Guillera-Arroita et al., 2019).

Multispecies occupancy model

I analyzed the data using a single-season Bayesian MSOM (Dorazio & Royle, 2005, Appendix S1: Figure S1). This modeling framework consists of three levels; the first of which represents the true occupancy state w_i of all observed and potentially unobserved species i in the metacommunity (Equation 5). The dataset of observed species n can be augmented by m all-zero encounter histories representing species that may or may not be present in the metacommunity. The choice of m is somewhat arbitrary, but should be large enough that the posterior distribution for estimated metacommunity richness N is not truncated but not so large as to be computationally prohibitive (Guillera-Arroita et al., 2019). The parameter w_i is then modeled as a Bernoulli trial such that $w_i = 0$ for species that were not present in the metacommunity and $w_i = 1$ for species that were either directly observed or were not observed but were likely to be available for sampling in the metacommunity (Dorazio & Royle, 2005), in which the parameter Ω represents the probability a species is available for sampling in the metacommunity. I augmented the simulated dataset with five undetected species ($M = 5$), two of which were present in the simulated metacommunity but not detected. The remaining three augmented species were included as a control to ensure prior aggregation had minimal influence on other detection histories that may be included in the dataset.

The second level of the model represents the ecological quantity of interest; in this case, site-level occupancy. Site-level occupancy Z_{ij} takes the value of 1 when species i is present at site j, provided the species is available for sampling in the metacommunity. Occupancy is modeled as the outcome of a Bernoulli trial with the probability of success defined as the product of site-level occupancy probability Ψ_{ij} and the metacommunity parameter w_i (Equation 6). Thus, a species cannot occupy a site if it is not available for sampling in the metacommunity.

In empirical datasets, site-level occupancy Z_{ij} is often imperfectly observed due to detection errors associated with the sampling process. By sampling each site multiple times $k = 1, 2, 3$ over a short period, the model can estimate the probability of detecting a species that occupies the site during a given survey and better estimate the true occupancy state (Dorazio & Royle, 2005). Detection of a given species at a site during a given sampling period (x_{ijk}) is modeled as a Bernoulli process conditional on the species i occupying the site j (Equation 7). Similar to the model for site occupancy, the probability of success is defined as the product of detection probability during a given sampling period p_{ijk} and the true occupancy state Z_{ij}, meaning a species cannot be detected at a site where it is not present:

$$w_i \sim \text{Bernoulli}(\Omega), \tag{5}$$

$$Z_{ij} | w_i \sim \text{Bernoulli}(\Psi_{ij} \times w_i), \tag{6}$$

$$x_{ijk} | Z_{ij} \sim \text{Bernoulli}\left(p_{ijk} \times Z_{ij}\right). \tag{7}$$

Environmental covariates can be used to accurately estimate occupancy and detection probabilities using a logit link function. I used the simulated covariate described above to estimate site-level occupancy probability Ψ_{ij} (Equation 8):

$$\text{logit}\left(\Psi_{ij}\right) = \alpha 0_i + \alpha 1_i \text{cov}_j. \tag{8}$$

Species-level values for model intercepts ($a0$, $b0$) and covariate coefficients ($a1$) for all detected species were modeled using uninformative priors (e.g., Equation 9). The parameters of the community-level distribution from which species were drawn, called hyperparameters, were in turn drawn from a hyperprior distribution (Equations 10–12):

$$\alpha 0_i \sim N(\mu_{\alpha 0}, \tau_{\alpha 0}), \tag{9}$$

$$\tau_{\alpha 0} \sim \text{Gamma}(0.1, 0.1), \tag{10}$$

$$\mu_{\alpha 0} = \log(\text{mean}.a0) - \log(1 - \text{mean}.a0), \tag{11}$$

$$\text{mean}.a0 \sim U(0, 1). \tag{12}$$

Hyperprior specification was derived from Zipkin et al. (2009) and Kéry and Royle (2009); these hyperpriors allow for heterogeneity in occupancy probabilities (Coull & Agresti, 1999; Royle et al., 2007). The parameter tau (τ) in the equations above represents precision, and is used instead of standard deviation σ in the JAGS programming language (Plummer, 2017).

Although the use of a hyperprior allows species with little or no data to be modeled in the context of the full community, model estimates for these species are disproportionately "pulled" to the center of the hyperprior distribution due to a lack of data. However, solely modeling these species using highly informative priors results in a loss of the advantages gained by modeling rare species in the context of the community. Prior aggregation is a promising tool for resisting the "pull" of the hyperprior when modeling undetected species, while also retaining the advantages of modeling undetected species in the context of the community. In brief, prior aggregation involves combining two or more prior distributions using a defined pooling method (Genest et al., 1984), typically as a way to account for multiple differing expert opinions. In the context of modeling undetected species, one can aggregate (1) the hyperprior distribution, toward the center of which undetected species are pulled, and (2) a prior distribution based on information about the undetected species that is not present in the dataset.

I calculated aggregated priors for the two undetected species for the parameters $a0$ and $a1$ using logarithmic pooling for Gaussian distributions (de Carvalho et al., 2015, Equations 13–15):

$$\boldsymbol{w}^* = \frac{\boldsymbol{\alpha}}{\boldsymbol{\sigma}^2}, \tag{13}$$

$$\sigma^2_{\text{pooled}} = \frac{1}{\sum \boldsymbol{w}^*}, \tag{14}$$

$$\mu_{\text{pooled}} = \sigma^2_{\text{pooled}} \times \sum \left(\boldsymbol{w}^* \times \boldsymbol{\mu} \right). \tag{15}$$

In which $\boldsymbol{\sigma}^2$ is a vector of variances of the initial prior distributions, $\boldsymbol{\mu}$ a vector of means, and $\boldsymbol{\alpha}$ a vector of pooling weights (see below). The parameter $a0$ was an aggregate of the community prior $N(\mu_{a0}, \tau_{a0})$ and an ecological prior $N(\mu_{\text{True}}, \tau_{\text{True}})$. For models with informative priors, μ_{True} was the true, simulated occupancy probability that was logit transformed and rounded to the nearest integer; models with misspecified priors used the opposite sign as the true value. Similarly, parameter $a1$ was an aggregate of the community prior and the ecological prior. Models with informative priors used the true value for the species-level coefficient rounded to the nearest whole number as the mean of the distribution, whereas misspecified models used a value with the opposite sign (or a value of −3 if the true coefficient value was 0). The precision parameter $\boldsymbol{\tau}$ was assigned a value of 0.5 for all ecological priors.

Pooling weights (Equation 13) define the relative contribution of individual priors to the aggregated distribution. The weight assigned to each prior distribution represents the relative degree of confidence in the information it contains (Genest et al., 1984). Methods for systematically assigning prior weights have been developed (e.g., de Carvalho et al., 2015); however, these methods are typically used for aggregating multiple expert opinions, leaving weight assignment in other situations somewhat arbitrary (French, 1983). I assigned the ecological priors for weakly informative models a weight of 0.15 and the community prior a weight of 0.85; for moderately informative models, both priors were assigned a weight of 0.5; and for strongly informative models, a weight of 0.85 for the ecological prior and a weight of 0.15 for the community prior. The vector of weights for each aggregated prior must sum to 1.

I compared models with possible prior combinations (informative/misspecified × weakly/moderately/strongly) to one another and to a single model with uninformative priors, resulting in seven different models. I compared estimates of (1) regional species richness, (2) site-level species richness, and (3) species-level regression coefficients across the seven models. To ensure that prior aggregation has minimal influence on estimates of (1) community-level parameters and (2) species for which an informative prior was not used, I visually compared the posterior distributions of (1) community-level hyperparameters and (2) species-level model coefficients. I estimated all model parameters using Bayesian analysis in the program JAGS (Plummer, 2017) and the R package R2jags (R Core Team, 2020; Su & Yajima, 2015). I ran each model using three Markov chains and assessed convergence using the R-hat statistic, which compares between-chain and within-chain parameter estimates for each of the Markov chains (Gelman & Rubin, 1992), and by visual examination of the trace plots. An R-hat less than 1.1 and trace plots with a stable, "grassy" appearance (i.e., well mixed, Gelman et al., 2013) were considered converged. Values for the length of the Markov chains, burn-in period, and thinning were chosen on a trial-and-error basis until model convergence was achieved. A tutorial of the prior aggregation method using R and JAGS can be found in Appendix S2; data and code associated with the analysis are archived on Zenodo (Beasley, 2023).

Application to real data

In addition to the simulation analyses, I applied the prior aggregation method to an empirical dataset of small mammal trapping surveys collected in Vermont from

May to July 2019. Sampling occurred in 30 sites located in forests, uncultivated fields, and active farms (Appendix S1: Figure S2). Trapping transects were 300 m long, with trap stations 10 m apart, with two traps per station placed to maximize capture efficiency (e.g., along fallen logs or rock ledges). Traps were baited with sunflower seeds and supplemented with batting and mealworms to reduce cold-related mortality (Do et al., 2013). Traps were opened in the evening and checked the following morning for a period of 3 consecutive days. I marked captured mammals with an ear tag, identified them to species, and released them unharmed at the point of capture.

I collected vegetation data at every third trap station along each transect for a total of 10 samples per site. Vegetation metrics included (1) composition, measured as the proportion of each cover type in a 0.5 × 0.5 m grid, (2) vertical structure, measured using the point-touch method described in Wiens (1969), and (3) canopy cover, measured using a spherical convex densiometer. I reduced the dimensionality of the data using a principal components analysis (PCA). I incorporated the first principal component as a covariate in the MSOM because multiple variables had relatively large loadings and the principal component was ecologically interpretable. Should the principal component not be interpretable or dominated by one variable, the variable with the highest loading could be chosen as the environmental covariate.

I examined how the use of informative priors affected estimates of real datasets in a similar manner to the procedure described in the previous section. The dataset was augmented with two all-zero encounter histories; with one representing the Eastern cottontail *Sylvilagus floridanus*, a species common in the study region and visually observed at some sampling sites, but with low catchability (and therefore detectability) in Sherman live traps.

I applied aggregated priors to the occupancy intercept $\alpha0_i$ and the species-level coefficient $\alpha1_i$. Prior information was derived from the literature and field notes taken during sampling (Table 1). I ran one model with weakly informative priors and another with moderately informative priors using the relative weights defined in the previous section; these models were compared with a model with uninformative priors. I compared estimates of (1) regional species richness and (2) species-level coefficients across these three models. Model specifications such as the number of Markov chains, model iterations, and evaluation of model convergence were determined in the manner described in the previous section.

RESULTS

Simulated data

Models with informative priors generally yielded more accurate estimates of metacommunity richness than models with uninformative or misspecified priors (Figures 2 and 3). Specifically, models with strongly informative priors (i.e., the contribution of the ecologically informative distribution to the aggregate prior was high compared with the community distribution) yielded higher posterior probabilities of 22 species in the metacommunity compared with the models with uninformative or misspecified priors (Figure 2). Models with weakly informative and moderately informative priors also tended to yield higher posterior probabilities of 22 species than models with uninformative priors (Figure 2). Models with moderately informative and strongly informative priors yielded higher estimated probabilities of at least one of the undetected species being present in the metacommunity (e.g., $N = 21$ or $N = 22$; Figure 3). All models were equally likely to overestimate regional richness (Appendix S1: Figure S3).

At the site level, models with moderately and strongly informative priors yielded richness estimates that were closer to true values than models with uninformative priors (Figure 4). Models with weakly and moderately misspecified priors generally estimated site-level richness as accurately as models with uninformative priors. Models with weakly informative and strongly misspecified priors yielded richness estimates that deviated the most from the true values.

TABLE 1 Sources of ecologically informative priors for *Sylvilagus floridanus*.

Parameter	Source	Description
α_0	Field notes	*Sylvilagus floridanus* was visually observed at 20% of sites; the mean of the prior distribution was set at logit (0.2) with a variance of 0.5
α_1	Chapman et al., 1980; DeGraaf & Yamasaki, 2001	Old fields and grasslands are preferred habitat in the northeastern United States, interpreted as a negative response to PC1. The mean of the prior distribution was set at −2.
	Field notes	All visual observations of this species occurred in old fields or active farms.

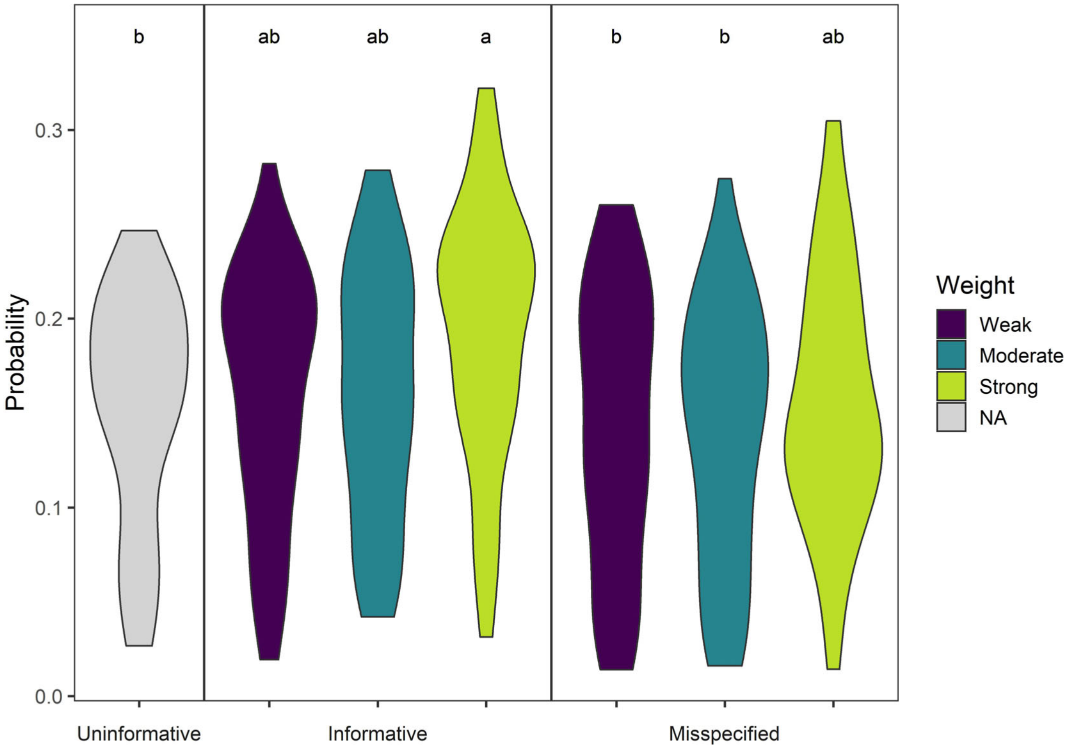

FIGURE 2 Estimated posterior probabilities of a metacommunity richness $N = 22$ species across models with uninformative, informative, and misspecified priors. Letters denote groups as assigned using a Tukey test. Models with strongly informative priors yielded higher estimated probabilities for the true metacommunity richness of 22 species than models with uninformative priors. Models with weakly and moderately informative priors also performed marginally better than models with uninformative priors. Models with misspecified priors performed similarly to models with uninformative priors.

The model with uninformative priors correctly estimated a nonsignificant coefficient for one undetected species but failed to detect a significant coefficient for the second undetected species (Figure 5). Models with informative and misspecified priors generally yielded more precise estimates of coefficients for undetected species than models with uninformative priors, although the accuracy varied depending on the accuracy of the information supplied to the model (Figure 5). The models with weakly informative priors, weakly misspecified priors, and moderately misspecified priors yielded estimates qualitatively similar to the model with uninformed priors, and models with moderately and strongly informative priors correctly estimated model coefficients for both undetected species (Figure 5). The model with strongly misspecified priors incorrectly estimated a positive regression coefficient for both undetected species (Figure 5). The improvement in coefficient estimates probably caused the improvement in site-level occupancy estimates for undetected species in models with informative priors (Appendix S1: Figures S4 and S5).

The prior aggregation had minimal influence on the posterior estimates of community-level hyperparameters (Appendix S1: Figure S6) and species-level model coefficients (Appendix S1: Figure S7).

Vermont small mammals

I captured 89 individuals representing 10 species. The most common species were the white-footed mouse *Peromyscus leucopus* with 33 individuals, the meadow jumping mouse *Zapus hudsonius* with 17 individuals, and the woodland jumping mouse *Napaeozapus insignis* with 16 individuals. All other species were represented by fewer than 10 individuals.

The first principal component from the PCA of the vegetation data explained 82.4% of the variation in the data. This principal component was included in the model as an environmental covariate, capturing a gradient from mostly grassy cover (low PCA scores) to cover that is predominately leaf litter and other dead vegetation (high PCA scores; Appendix S1: Figure S8).

The model with uninformed priors yielded a metacommunity richness estimate of 10 species, while models with informed priors yielded estimates of 11 species (Appendix S1: Figure S9). At the species level, the

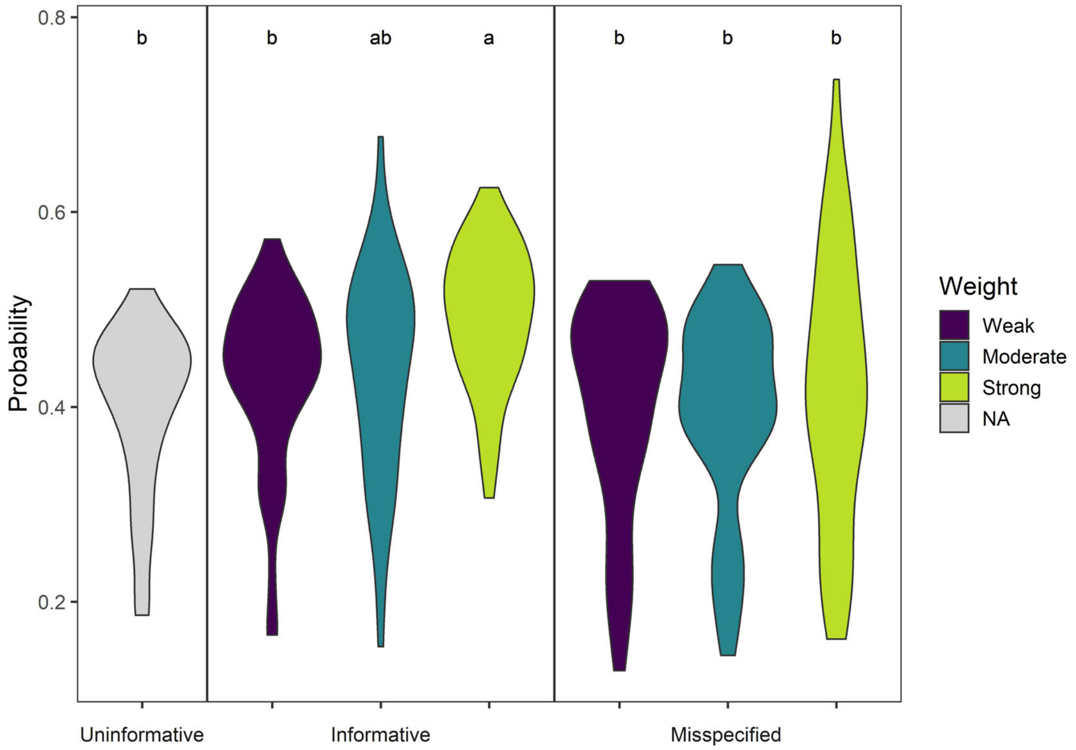

FIGURE 3 Estimated posterior probabilities of a metacommunity richness $N = 21$ or $N = 22$ species across models with uninformative, informative, and misspecified priors. Letters denote groups as assigned using the Tukey test. Models with strongly informative priors yielded higher estimated probabilities for the true metacommunity richness of 22 species than models with uninformative priors. Models with moderately informative priors also performed marginally better than models with uninformative priors. Models with misspecified priors performed similarly to models with uninformative priors.

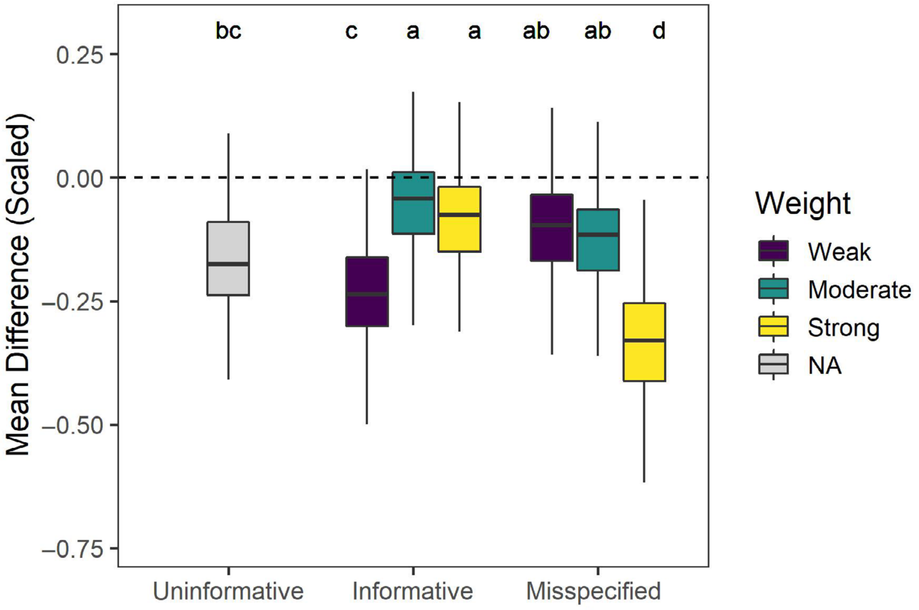

FIGURE 4 Mean differences between true and estimated site-level richness of the simulated datasets, standardized as a percent error. Models with moderately and strongly informative priors outperformed models with uninformative, weakly informative, and strongly misspecified priors. Models with strongly misspecified priors performed less well than models with uninformative priors. Letters denote groups as assigned using the Tukey test.

augmented species *S. floridanus* was not predicted to have a covariate response significantly different from 0 in any model; however, the species-level estimate from models with informed priors was more precise than the model with uninformed priors (Figure 6).

DISCUSSION

These results suggest that using prior aggregation to model undetected species improves estimates of multiple model parameters, provided the information supplied to

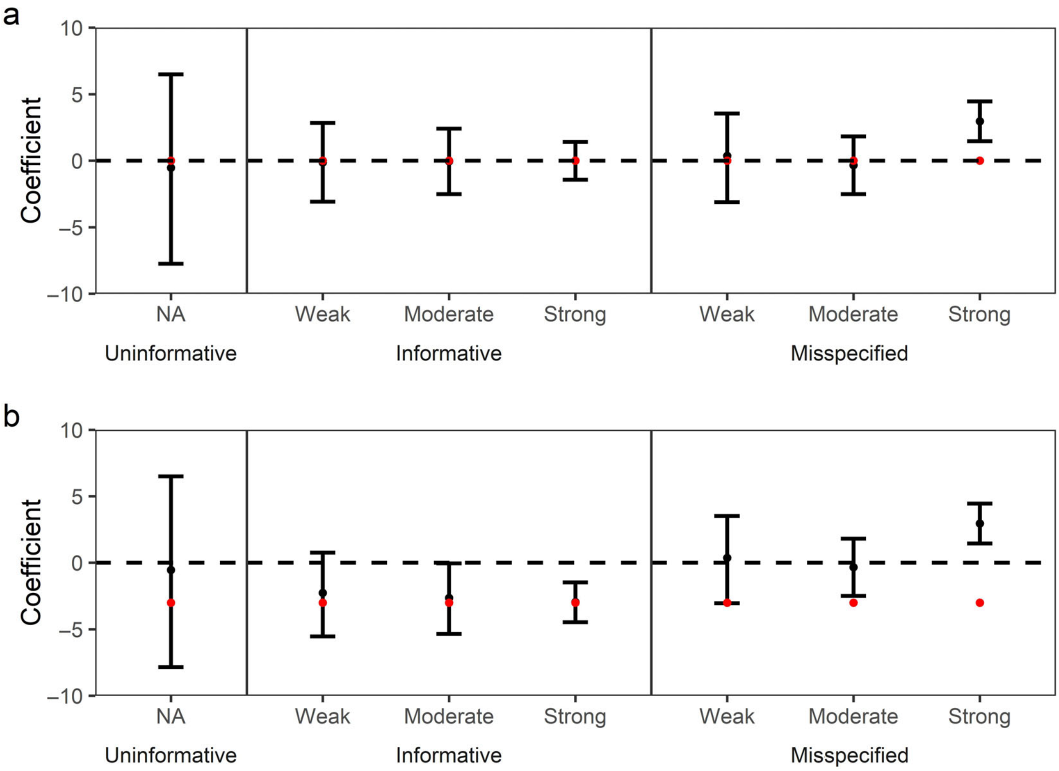

FIGURE 5 Estimated responses of two undetected species to simulated covariates. Error bars represent the 95% CI; error bars that did not overlap 0 (dashed line) were considered significant. Red dots represent the true value of the simulated coefficient. Increasing the relative weight of the species level prior increased the precision of model estimates, regardless of the accuracy of the prior. The model correctly estimated a nonresponse to the covariate in all models except models with strongly misspecified priors for the first undetected species (a). Models with moderately and strongly informative priors correctly estimated a significant, negative response to the covariate for the second undetected species (b).

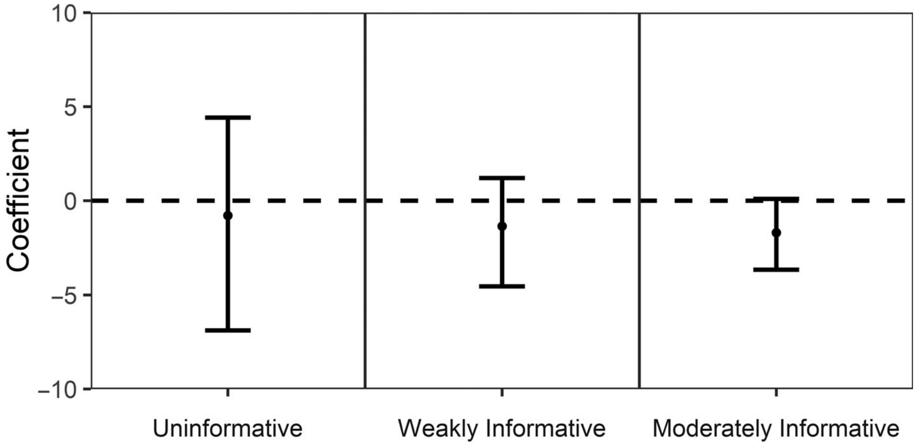

FIGURE 6 Species-specific responses to the vegetation covariate for the augmented species *Sylvilagus floridanus*. Error bars represent the 95% credible interval; bars that do not overlap 0 (dashed line) were considered significant. No model yielded significant covariate estimates.

the model is correct (Figures 3 and 5). However, the degree of improvement depends on the parameter in question: prior aggregation tends to have a larger effect on species-level coefficient estimates than site-level or regional richness estimates. These findings align with previous work suggesting that informative priors in Bayesian models tended to improve model estimates when priors are appropriately specified (Lemoine, 2019; McCarthy & Masters, 2005; Morris et al., 2015; Northrup & Gerber, 2018). In addition, prior aggregation tends to result in more ecologically meaningful conclusions for undetected species by reducing the pull of the community prior and retaining information about species with particular characteristics rather than hypothetical species of unknown identity.

Species are more likely to be undetected when they are rare (McCarthy et al., 2013). Rare species are often of conservation concern (Cunningham & Lindenmayer, 2005; Fagan et al., 2002; MacKenzie et al., 2005) and can drive site-level variation in metrics such as species richness, beta diversity, and functional diversity (Leitão et al., 2016; Mao & Colwell, 2005; Routledge, 1977, but see Lennon et al., 2004). A common strategy for addressing this problem is to use common, closely related species as a proxy for rare relatives (Gaston & Kunin, 1997); however, rare and common species are often ecologically different (Kunin & Gaston, 1993; Leitão et al., 2016). Using a hierarchical MSOM is an improvement over the use of rare species as a proxy because prior estimates of the community are derived from data from all species in the community, rare and common. However, using a community prior still uses data from species that may differ from the undetected species of interest in ecologically important ways. Accounting for these differences with ecologically informative priors can lead to more accurate estimates on which to base management decisions.

One consideration when applying aggregated priors to inform species-level coefficients is that when the ecologically informed prior is assigned a high weight relative to the community prior, the posterior for the coefficient is essentially estimated from the prior (Figure 5). When the information supplied to the prior is incorrect, this results in severe bias in the coefficient estimate (Figure 5), which can lead to poor management decisions. Bias can also be introduced to the aggregated prior via a mismatch between data types (e.g., qualitative prior information to quantitative covariates, Kuhnert et al., 2010) or when multiple covariates are considered in context (Chen et al., 1999). The former issue is applicable to the Vermont small mammals dataset: the first principal component of the vegetation PCA was a reasonable match for the "open habitat" preference of undetected species *S. floridanus* from historical data (Chapman et al., 1980; DeGraaf & Yamasaki, 2001). However, covariates measured using modern techniques (e.g., LiDAR) or modified using more recent statistical methods (e.g., PCA) might not be as compatible with historical records. Although highly informative priors can markedly improve estimates of species-level coefficients (Figure 5), it is important to ensure that the information in the prior is compatible with the environmental covariate before inducing an informative prior on the species-level coefficient.

From a management perspective, estimates for specific sites may be just as important as regional or species-level estimates, especially for targeted management actions such as habitat restoration efforts or reserve design (Cabeza et al., 2004). My results suggest that prior aggregation only improves site-level estimates when the relative weight of the informative prior is moderate or strong (Figure 4), and therefore prior aggregation may not be beneficial when characteristics of the site are of primary interest. However, the lack of improvement of site-level estimates may be due to characteristics of the simulated data and model structure rather than characteristics of the priors. The simulated dataset included species with positive and negative covariate responses, which could be affecting the accuracy of site-level richness estimates compared with a covariate with more uniform effects, such as patch area. Models for the simulated and empirical datasets also assumed stochastic detection errors. Detectability in real communities is often influenced by site-level or species-level characteristics (Iknayan et al., 2014) and accounting for site-level variation in detectability using model covariates tends to improve estimates (New & Handel, 2015). In systems in which detectability varies by site and is modeled using a covariate, the use of prior aggregation may improve site-level richness estimates compared with models with uninformative priors.

A key component of prior aggregation is assigning weights to each of the contributing priors. Weight choice determines how much of the ecologically informed prior contributes to the final aggregate and should reflect the reliability of the source of information (Genest et al., 1984). Defining the reliability of a source is difficult, and in practice the choice of prior weight is somewhat arbitrary (French, 1983). That said, methods for choosing weights in a more meaningful way have been developed (Abbas, 2009; Myung et al., 1996; Rufo, Martin, & Pérez, 2012; Rufo, Pérez, & Martin, 2012) and a few of these also account for uncertainty about the weights (de Carvalho et al., 2015; Poole & Raftery, 2000). A possible avenue for future research would include adapting these methods for use in MSOMs.

Although prior aggregation shows promise as a method for using external data to model undetected species more accurately, inducing strong priors on these species can potentially lead to erratic model results. The following practices will allow a user to better detect a scenario in which prior aggregation may cause unanticipated results and provide insights into model behavior. First, as with all Bayesian hierarchical models, a prior sensitivity analysis is a must to determine the effects of prior choice on the posterior distribution (Banner et al., 2020; Cressie et al., 2009; Lele & Dennis, 2009; Northrup & Gerber, 2018). In the context of prior aggregation, this includes adjusting prior weights, along with other parameters that potentially influence the "informativeness" of the prior, such as τ.

Second, if including covariate scenarios that were not tested here, it is best to first test a toy model using simulated data before application to a real data set. Such scenarios include, but are not limited to, the inclusion of detection covariates or additional occupancy covariates, "dummy" variables for modeling the effects of a categorial covariate, or covariate interactions. Third, there can be difficulty mapping historical data onto modern priors. For example, the results of a PCA might not be readily interpretable or, if interpretable, may not be compatible with previous studies using raw environmental variables. Testing increasingly complex modeling scenarios is beyond the scope of this paper, but previous work suggests that informative priors can be used in mortality, survival, or occupancy models with multiple covariates to improve the precision and accuracy of model estimates if the priors are appropriately specified (Morris et al., 2015; Parlato et al., 2021; Rodhouse et al., 2019).

The concept of "borrowing" data from multiple sources is not new in ecology (McCarthy & Masters, 2005), and pooling information across species within a dataset is a common practice in hierarchical detection models (Iknayan et al., 2014; Link & Sauer, 1996). Using prior aggregation to incorporate data from external sources, such as previous studies or natural history collections, to improve model accuracy is an extension of this concept. In addition to the Vermont small mammals case study discussed previously, prior aggregation may be used to bridge gaps in monitoring (Rodhouse et al., 2019) in communities where species may have gone locally extinct; or may be used in conjunction with Bayes factors to compare hypotheses about communities in which species composition is uncertain (Kary et al., 2016; Vanpaemel & Lee, 2012). The flexibility of hierarchical detection models means that prior aggregation is not limited to questions of species richness or occupancy: prior aggregation can potentially be used to add information about missing individuals in a population (Royle & Dorazio, 2012) leading to more accurate estimates of abundance, survival rates, or diversity estimates. Despite the continuing challenges of choosing meaningful prior weights (de Carvalho et al., 2015; Genest et al., 1984) and prior selection in Bayesian ecological models in general (Banner et al., 2020; Lemoine, 2019; Northrup & Gerber, 2018), prior aggregation is a promising tool for using external data to generate more reasonable estimates in systems for which nondetection is of ecological concern.

ACKNOWLEDGMENTS

This work was conducted on land stewarded by ongoing generations of Indigenous peoples, in particular the Abenaki. I thank Bailey Luter, Riley O'Halloran, and Tabitha Myers for their assistance in the field, and Nicholas Gotelli, Brittany Mosher, and four anonymous reviewers for their helpful comments on the manuscript. The research was funded by the American Museum of Natural History Theodore Roosevelt Memorial Grant. Permission for trapping small mammals was granted by the Vermont Fish and Wildlife Department (permit no. SR-2018-25), the National Park Service (OMB no. 1024-0236), and other farmers and land managers. Methods were approved by the University of Vermont IACUC (protocol no. PROTO202000114).

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data and code (Beasley, 2023) are available in Zenodo at https://doi.org/10.5281/zenodo.10054955.

ORCID

Emily M. Beasley https://orcid.org/0000-0003-0593-0760

REFERENCES

Abbas, A. E. 2009. "A Kullback–Leibler View of Linear and Log-Linear Pools." *Decision Analysis* 6: 25–37.

Banks-Leite, C., R. Pardini, D. Boscolo, C. R. Cassano, T. Püttker, C. S. Barros, and J. Barlow. 2014. "Assessing the Utility of Statistical Adjustments for Imperfect Detection in Tropical Conservation Science." *Journal of Applied Ecology* 51: 849–859.

Banner, K. M., K. M. Irvine, and T. J. Rodhouse. 2020. "The Use of Bayesian Priors in Ecology: The Good, the Bad and the Not Great." *Methods in Ecology and Evolution* 11: 882–89.

Beasley, E. M. 2023. "Beasley015/Beasley2023_Ecologically_informed_priors_undetected_species: Publication Version (v1.0.1)." Zenodo. https://doi.org/10.5281/zenodo.10054955.

Benoit, D., D. A. Jackson, and M. S. Ridgway. 2018. "Assessing the Impacts of Imperfect Detection on Estimates of Diversity and Community Structure through Multispecies Occupancy Modeling." *Ecology and Evolution* 8: 4676–84.

Burnham, K. P., and W. S. Overton. 1979. "Robust Estimation of Population Size when Capture Probabilities Vary among Animals." *Ecology* 60: 927–936.

Cabeza, M., M. B. Araújo, R. J. Wilson, C. D. Thomas, M. J. R. Cowley, and A. Moilanen. 2004. "Combining Probabilities of Occurrence with Spatial Reserve Design." *Journal of Applied Ecology* 41: 252–262.

Chao, A. 1984. "Nonparametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics* 11: 265–270.

Chapman, J. A., J. G. Hockman, and M. M. Ojeda. 1980. "*Sylvilagus floridanus*." *Mammalian Species* 136: 1–8.

Chen, M.-H., J. G. Ibrahim, and C. Yiannoutsos. 1999. "Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61: 223–242.

Coull, B. A., and A. Agresti. 1999. "The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies." *Biometrics* 55: 294–301.

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef, and C. K. Wikle. 2009. "Accounting for Uncertainty in Ecological Analysis: The Strengths and Limitations of Hierarchical Statistical Modeling." *Ecological Applications* 19: 553–570.

Cunningham, R. B., and D. B. Lindenmayer. 2005. "Modeling Count Data of Rare Species: Some Statistical Issues." *Ecology* 86: 1135–42.

de Carvalho, L. M., D. A. M. Villela, F. C. Coelho, and L. S. Bastos. 2015. "Combining Probability Distributions: Extending the Logarithmic Pooling Approach." *ArXiv*. https://doi.org/10.48550/ARXIV.1502.04206.

DeGraaf, R. M., and M. Yamasaki. 2001. *New England Wildlife: Habitat, Natural History, and Distribution*. Lebanon, NH: UPNE.

Do, R., J. Shonfield, and A. G. McAdam. 2013. "Reducing Accidental Shrew Mortality Associated with Small-Mammal Livetrapping II: A Field Experiment with Bait Supplementation." *Journal of Mammalogy* 94(4): 754–760. https://doi.org/10.1644/12-mamm-a-242.1.

Dorazio, R. M., and J. A. Royle. 2005. "Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species." *Journal of the American Statistical Association* 100: 389–398.

Fagan, W. F., P. J. Unmack, C. Burgess, and W. L. Minckley. 2002. "Rarity, Fragmentation, and Extinction Risk in Desert Fishes." *Ecology* 83: 3250–56.

Ferrier, S., and A. Guisan. 2006. "Spatial Modelling of Biodiversity at the Community Level." *Journal of Applied Ecology* 43: 393–404.

French, S. 1983. *Group Consensus Probability Distributions: A Critical Survey*. Manchester, UK: University of Manchester. Department of Decision Theory.

Gaston, K., and W. E. Kunin. 1997. "Concluding Comments." In *The Biology of Rarity*, edited by K. Gaston and W. E. Kunin. London: Chapman & Hall.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

Gelman, A., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7: 457–511.

Genest, C., S. Weerahandi, and J. V. Zidek. 1984. "Aggregating Opinions through Logarithmic Pooling." *Theory and Decision* 17: 61–70.

Gu, W., and R. K. Swihart. 2004. "Absent or Undetected? Effects of Non-detection of Species Occurrence on Wildlife–Habitat Models." *Biological Conservation* 116: 195–203.

Guillera-Arroita, G., M. Kéry, and J. J. Lahoz-Monfort. 2019. "Inferring Species Richness Using Multispecies Occupancy Modeling: Estimation Performance and Interpretation." *Ecology and Evolution* 9: 780–792.

Iknayan, K. J., M. W. Tingley, B. J. Furnas, and S. R. Beissinger. 2014. "Detecting Diversity: Emerging Methods to Estimate Species Diversity." *Trends in Ecology & Evolution* 29: 97–106.

Kary, A., R. Taylor, and C. Donkin. 2016. "Using Bayes Factors to Test the Predictions of Models: A Case Study in Visual Working Memory." *Journal of Mathematical Psychology* 72: 210–19.

Kellner, K. F., and R. K. Swihart. 2014. "Accounting for Imperfect Detection in Ecology: A Quantitative Review." *PLoS One* 9: e111436.

Kéry, M., and J. A. Royle. 2009. "Inference about Species Richness and Community Structure Using Species-Specific Occupancy Models in the National Swiss Breeding Bird Survey MHB." In *Modeling Demographic Processes in Marked Populations*, edited by D. L. Thomson, E. G. Cooch, and M. J. Conroy, 639–656. Boston: Springer.

Kéry, M., and M. Schaub. 2011. *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Cambridge, MA: Academic Press.

Kuhnert, P. M., T. G. Martin, and S. P. Griffiths. 2010. "A Guide to Eliciting and Using Expert Knowledge in Bayesian Ecological Models." *Ecology Letters* 13: 900–914.

Kunin, W. E., and K. J. Gaston. 1993. "The Biology of Rarity: Patterns, Causes and Consequences." *Trends in Ecology & Evolution* 8: 298–301.

Leitão, R. P., J. Zuanon, S. Villéger, S. E. Williams, C. Baraloto, C. Fortunel, F. P. Mendonça, and D. Mouillot. 2016. "Rare Species Contribute Disproportionately to the Functional Structure of Species Assemblages." *Proceedings of the Royal Society B: Biological Sciences* 283: 20160084.

Lele, S. R., and B. Dennis. 2009. "Bayesian Methods for Hierarchical Models: Are Ecologists Making a Faustian Bargain?" *Ecological Applications* 19: 581–84.

Lemoine, N. P. 2019. "Moving beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses." *Oikos* 128: 912–928.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. 2004. "Contribution of Rarity and Commonness to Patterns of Species Richness." *Ecology Letters* 7: 81–87.

Link, W. A., and J. R. Sauer. 1996. "Extremes in Ecology: Avoiding the Misleading Effects of Sampling Variation in Summary Analyses." *Ecology* 77: 1633–40.

Low Choy, S., R. O'Leary, and K. Mengersen. 2009. "Elicitation by Design in Ecology: Using Expert Opinion to Inform Priors for Bayesian Statistical Models." *Ecology* 90: 265–277.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. "Estimating Site Occupancy Rates when Detection Probabilities Are Less than One." *Ecology* 83: 2248–55.

MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. "Improving Inferences in Population Studies of Rare Species that Are Detected Imperfectly." *Ecology* 86: 1101–13.

Mao, C. X., and R. K. Colwell. 2005. “Estimation of Species Richness: Mixture Models, the Role of Rare Species, and Inferential Challenges.” *Ecology* 86: 1143–53.

McCarthy, M. A., and P. Masters. 2005. “Profiting from Prior Information in Bayesian Analyses of Ecological Data.” *Journal of Applied Ecology* 42: 1012–19.

McCarthy, M. A., J. L. Moore, W. K. Morris, K. M. Parris, G. E. Garrard, P. A. Vesk, L. Rumpff, et al. 2013. “The Influence of Abundance on Detectability.” *Oikos* 122: 717–726.

Morris, W. K., P. A. Vesk, M. A. McCarthy, S. Bunyavejchewin, and P. J. Baker. 2015. “The Neglected Tool in the Bayesian Ecologist’s Shed: A Case Study Testing Informative Priors’ Effect on Model Accuracy.” *Ecology and Evolution* 5: 102–8.

Myung, I. J., S. Ramamoorti, and A. D. Bailey. 1996. “Maximum Entropy Aggregation of Expert Predictions.” *Management Science* 42: 1420–36.

New, L. B. M., and C. M. Handel. 2015. “Evaluating Species Richness: Biased Ecological Inference Results from Spatial Heterogeneity in Detection Probabilities.” *Ecological Applications* 25: 1669–80.

Northrup, J. M., and B. D. Gerber. 2018. “A Comment on Priors for Bayesian Occupancy Models.” *PLoS One* 13: e0192819.

Parlato, E. H., J. G. Ewen, M. McCready, F. Gordon, K. A. Parker, and D. P. Armstrong. 2021. “Incorporating Data-Based Estimates of Temporal Variation into Projections for Newly Monitored Populations.” *Animal Conservation* 24: 1001–12.

Plummer, M. 2017. “JAGS: Just another Gibbs Sampler.” https://mcmc-jags.sourceforge.io/.

Poole, D., and A. E. Raftery. 2000. “Inference for Deterministic Simulation Models: The Bayesian Melding Approach.” *Journal of the American Statistical Association* 95: 1244–55.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rodhouse, T. J., R. M. Rodriguez, K. M. Banner, P. C. Ormsbee, J. Barnett, and K. M. Irvine. 2019. “Evidence of Region-Wide Bat Population Decline from Long-Term Monitoring and Bayesian Occupancy Models with Empirically Informed Priors.” *Ecology and Evolution* 9: 11078–88.

Routledge, R. D. 1977. “On Whittaker’s Components of Diversity.” *Ecology* 58: 1120–27.

Royle, J. A., and R. M. Dorazio. 2012. “Parameter-Expanded Data Augmentation for Bayesian Analysis of Capture–Recapture Models.” *Journal of Ornithology* 152: 521–537.

Royle, J. A., R. M. Dorazio, and W. A. Link. 2007. “Analysis of Multinomial Models with Unknown Index Using Data Augmentation.” *Journal of Computational and Graphical Statistics* 16: 19–85.

Rufo, M. J., J. Martin, and C. J. Pérez. 2012. “Log-Linear Pool to Combine Prior Distributions: A Suggestion for a Calibration-Based Approach.” *Bayesian Analysis* 7: 411–438.

Rufo, M. J., C. J. Pérez, and J. Martin. 2012. “A Bayesian Approach to Aggregate Experts’ Initial Information.” *Electronic Journal of Statistics* 6: 2362–82.

Su, Y.-S., and M. Yajima. 2015. “R2jags: Using R to Run “JAGS”.” https://cran.r-project.org/web/packages/R2jags/R2jags.pdf.

Vanpaemel, W., and M. D. Lee. 2012. “Using Priors to Formalize Theory: Optimal Attention and the Generalized Context Model.” *Psychonomic Bulletin & Review* 19: 1047–56.

Wiens, J. A. 1969. “An Approach to the Study of Ecological Relationships among Grassland Birds.” *Ornithological Monographs* 8: 1–93. https://doi.org/10.2307/40166677.

Zipkin, E. F., A. DeWan, and J. Andrew Royle. 2009. “Impacts of Forest Fragmentation on Species Richness: A Hierarchical Approach to Community Modelling.” *Journal of Applied Ecology* 46: 815–822.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Beasley, Emily M. 2024. “Ecologically Informed Priors Improve Bayesian Model Estimates of Species Richness and Occupancy For Undetected Species.” *Ecological Applications* 34(2): e2941. https://doi.org/10.1002/eap.2941