DR. RAHEL SOLLMANN (Orcid ID : 0000-0002-1607-2039)

DR. MITCHELL JOSEPH EATON (Orcid ID : 0000-0001-7324-6333)

Article type : Articles

Journal: Ecological Applications

Manuscript type: Article

Running Head: Dirichlet process occupancy model

A Bayesian Dirichlet process community occupancy model to estimate community structure and species similarity

R. Sollmann,[1,7] M.J. Eaton,[2] W.A. Link,[3] P. Mulondo,[4] S. Ayebare,[4] S. Prinsloo,[4] A.J. Plumptre,[5,8] and D.S. Johnson[6]

[1] University of California Davis, Department of Wildlife, Fish and Conservation Biology, 1088 Academic Surge, One Shields Ave, Davis, CA 95616, USA

[2] U.S. Geological Survey, Southeast Climate Adaptation Science Center, N.C. State University, Raleigh, NC, USA

[3] US Geological Survey Patuxent Wildlife Research Center, Laurel, MD, USA

[4] Wildlife Conservation Society, PO Box 7487, Kampala, Uganda.

[5] KBA Secretariat, c/o BirdLife International, David Attenborough Building, Pembroke Street,

Cambridge, UK.
[6] Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98115, USA.
[7] Corresponding author. Email: rsollmann@ucdavis.edu
[8] Current address, Wildlife Conservation Society, 2300 Southern Boulevard, Bronx, NY

Manuscript received 12 June 2020; accepted 17 August 2020; final version received 29 October 2020.

Abstract

Community occupancy models estimate species-specific parameters while sharing information across species by treating parameters as sampled from a common distribution. When communities consist of discrete groups, shrinkage of estimates towards the community mean can mask differences among groups. Infinite mixture models using a Dirichlet process (DP) distribution, in which the number of latent groups is estimated from the data, have been proposed as a solution. In addition to community structure, these models estimate species similarity, which allows testing hypotheses about whether traits drive species response to environmental conditions. We develop a community occupancy model (COM) using a DP distribution to model species-level parameters. Because clustering algorithms are sensitive to dimensionality and distinctiveness of clusters, we conducted a simulation study to explore performance of the DP-COM with different dimensions (i.e., different numbers of model parameters with species-level DP random effects) and under varying cluster differences. Because the DP-COM is computationally expensive, we compared its estimates to a COM with a normal random species effect. We further applied the DP-COM model to a bird dataset from Uganda. Estimates of the number of clusters and species cluster identity improved with increasing difference among clusters and increasing dimensions of the DP; but the number of clusters was always overestimated. Estimates of number of sites occupied and species and community level covariate coefficients on occupancy probability were generally unbiased with (near-) nominal 95% Bayesian Credible Interval coverage. Accuracy of estimates from the normal and the DP-COM were similar. The DP-COM clustered 166 bird species into 27 clusters regarding their affiliation with open or woodland habitat and distance to oil wells. Estimates of covariate coefficients were

similar between a normal and the DP-COM. Except sunbirds, species within a family were not more similar in their response to these covariates than the overall community. Given that estimates were consistent between the normal and the DP-COM, and considering the computational burden for the DP models, we recommend using the DP-COM only when the analysis focuses on community structure and species similarity, as these quantities can only be obtained under the DP-COM.

Key words: bird point-counts, clustering, community occupancy model, dimensionality, Dirichlet process, latent groups, infinite mixture models

Introduction

Occupancy models (MacKenzie et al. 2002) have rapidly gained popularity in wildlife research, because they offer a means of estimating ecologically relevant parameters (species occurrence, association with covariates, colonization/extinction rates) while accounting for imperfect detection (MacKenzie et al. 2017) using relatively inexpensive species detection/non-detection data. The basic single-season single-species occupancy model has seen many modifications, including the joint modeling of multiple species in a community modeling framework (Dorazio and Royle 2005). Community models share information across species while maintaining the ability to estimate species-specific parameters by assuming that all parameters come from a common distribution. This distribution, in turn, is governed by hyperparameters, which reflect community-level patterns or processes; this model formulation is equivalent to including a species-level random effect. The community modeling approach has been combined with single-season (e.g., Zipkin et al. 2009, Sollmann et al. 2017) and dynamic (e.g., Dorazio et al. 2010) occupancy models, as well as with other hierarchical modeling frameworks such as distance sampling (Sollmann et al. 2016), or N-mixture modeling (Yamaura et al. 2016).

Choice of the specific distribution used to model species level parameters entails assumptions about how the community is structured. A common choice is the normal distribution, postulating that variation in parameter values across species can be described using a bell-shaped curve (Sauer and Link 2002, Kéry and Royle 2008, Zipkin et al. 2009). Particularly for data-sparse species, parameter estimates are pulled closer to the overall mean. Although the ability to derive

more precise parameter estimates for rarely observed species is a significant benefit of community models, this shrinkage of parameters towards the mean can mask effects that are present only in subgroups of the entire community (Pacifici et al. 2014). This problem can be circumvented by grouping species a priori and analyzing groups, rather than entire communities. This approach, however, reduces overall sample size and thus, precision of parameter estimates. Additionally, results can be sensitive to a priori grouping of species (Pacifici et al. 2014).

Finite mixture models, in which species are assigned probabilistically to a pre-defined number of latent groups, are an alternative to a priori grouping, and have been employed in a community modeling context (Dunstan et al. 2011, 2013). Building on the idea that communities consist of latent groups of species, Johnson and Sinclair (2017) proposed an infinite mixture approach for the joint modeling of multi-species abundance data using a Dirichlet process (DP) prior. In this approach, the number of clusters into which species group is unknown and must be estimated. Briefly, the DP consists of a base distribution from which cluster-specific parameter values are generated, and a concentration parameter α, which determines the amount of clustering. In the context of community models, species are allocated to clusters based on cluster probabilities, which are generated with an algorithm governed by α (for details, see Methods). All species in a cluster share the same parameter value, which serves to reduce the number of model parameters (Escobar and West 1995). Compared to normally distributed random effects, this semiparametric approach also increases the flexibility to capture patterns in parameter distribution within the community of interest (Dorazio et al. 2008). In addition, the approach provides information on community structure (number of clusters in the community), as well as the degree of similarity of species (how often two species belong to the same cluster) (Johnson and Sinclair 2017). The ability to estimate the degree of similarity in how species occurrence responds to covariates holds potential to address questions of ecological and conservation interest: the degree of similarity among species with similar functional traits can be used to quantify a community's response diversity, defined as the variation of responses to environmental change, which is a key determinant of ecosystem resilience (Mori et al. 2013). Further, estimates of similarity in habitat use can be contrasted with phylogenetic relatedness to investigate questions of coexistence and niche partitioning among closely related species, a topic of ongoing debate in ecology (Hutchinson 1959, Gotelli 2000, Graham et al. 2004).

In clustering algorithms, the cluster identity of objects is estimated based on multivariate data measured for each object. Clustering algorithms identify cluster identity with greater accuracy when more dimensions (i.e., more variables) are used to describe objects, as long as added dimensions contain information about clusters (e.g., Azizyan et al. 2013). Further, clustering in high-dimensional data (with 100s or 1000s of dimensions) suffers from the "curse of dimensionality" – the fact that in high dimensional space, volume expands so rapidly that data appear sparse and dissimilar, causing common clustering algorithms to be inefficient (Bellman 1957, Houle et al. 2010). Given the dependency of clustering algorithms on the dimensions of the data, the performance of a community model using a DP prior to cluster species likely also depends on the dimensions of the DP process. To our knowledge, the effect of dimensionality on the ability of the DP to recover information on clustering of and similarity among objects has not been explored in the context of ecological modeling.

In this study, we develop a community occupancy model (COM) with a multivariate DP distribution for species level parameters (DP-COM). Using a simulation study, we first assess the model's ability to recover community structure (number of clusters and species similarity) and estimate parameters of ecological interest in occupancy modeling (number of sites occupied and coefficients describing the relationship between occupancy probability and environmental variables). We set up the simulation to test how the dimensionality of the DP and differences among clusters affect these estimates. We then apply the DP-COM to bird survey data from Murchison Falls National Park, Uganda, to illustrate the modeling approach and its ability to address questions of species similarity. Finally, because DP priors are computationally expensive (Johnson and Sinclair 2017), tradeoffs between their use and traditional normal random effects models should be considered. We therefore compared accuracy of estimates from the DP-COM with that of a COM using a customary normal random species level effect (normal COM).

Methods

Model development

Under the hierarchical formulation (Royle and Dorazio 2008) of single-species single-season occupancy models (MacKenzie et al. 2002), whether or not a site j is occupied by the species of interest, z_j, is a Bernoulli random variable governed by occupancy probability ψ, which can be modeled as a function of site-specific covariates on an appropriate link scale f (for example, logit

or probit):

$$z_j \sim Bernoulli(\psi_j)$$

$$f(\psi_j) = \mathbf{X}_j'\boldsymbol{\beta}$$

Here, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{X}_j'$ is a vector with measures of the corresponding site-level covariates for site *j*. Sites are visited on *k* occasions, and binary observations of the focal species, y_{jk}, are treated as Bernoulli random variables governed by the detection probability *p*, which is conditional on the latent true occupancy state z_j and either adopts the value $p_{jk,}$ when $z_j = 1$, or a value of 0 when $z_j = 0$.

$$y_{jk} \sim Bernoulli(p_{jk} z_j)$$

Analogous to ψ, *p* can be modeled as a function of both site and occasion specific covariates.

To extend this to a community occupancy model, the parameters and latent variables of the model described above are further indexed by species, *i*. Rather than treating species-level parameters as independent, we assume that parameters come from a common distribution, governed by community (or hyper-) parameters (Dorazio and Royle 2005, Dorazio et al. 2006). This model formulation constitutes a form of information sharing, which allows us to include species with sparse data into the analysis.

A normal distribution is a common choice to describe species level parameters; however, this entails parametric assumptions of unimodality and symmetry in the community. In contrast, the semi-parametric DP allows fitting infinite mixture models that treat species as belonging to latent clusters and lets the data govern the specific cluster structure of the community. The DP consists of a base distribution, G_0, which generates cluster-level parameters, and a concentration parameter α, which governs the amount of clustering. Under this formulation, the probability distribution for species-level parameters is a random draw from a DP [for a formal description of the DP, see Sethuraman 1994]. There are multiple means of implementing a DP; we opted for the Stick Breaking Algorithm (Sethuraman 1994), because it can be readily implemented in JAGS (Ohlssen et al. 2007). In the Stick Breaking Algorithm, cluster probabilities are generated using a sequence of auxiliary variables $v \sim \text{Beta}(1,\alpha)$, with mean $E(v) = 1/(1+\alpha)$. The variable v can be thought of as the proportion that is broken off a stick. The proportion v_1 corresponds to the probability of cluster 1, π_1; v_2 is the proportion broken off the remaining stick, and can be

translated into π_2 by scaling it back to the size of the original stick, $\pi_2 = v_2(1 - v_1)$, and so forth for the remaining clusters. The *n* species are then assigned to a cluster using a Multinomial(*n*, $\boldsymbol{\pi}$) distribution. If α is large, only small pieces are broken off, leading to many clusters *K* and a distribution of species-level parameters that approximates G_0. If α is small, large pieces are broken off, resulting in few clusters and a distribution of species-level parameters that can look very different from G_0.

It is common practice (though not exclusive) in community models to ascribe separate univariate hyperdistributions to each set of species-specific parameters. To take advantage of the relationship between the number of dimensions of multivariate data and the ability to identify clusters in the data, we followed Johnson and Sinclair (2017) and specified G_0 in the DP-COM as a multivariate normal (*MVN*) distribution. Here, the *MVN* means correspond to the community hyperparameters $\boldsymbol{\beta}$, which determine the distribution of parameters across clusters. Rather than estimating the *MVN* means directly, we estimated them as separate fixed parameters and parameterized the *MVN* G_0 in terms of deviations from the community mean effect, $\boldsymbol{\delta}_i^*$. This allowed us to center the *MVN* on 0 for identifiability (Johnson and Sinclair, 2017):

$$\boldsymbol{\delta}_i^* \sim DP(G_0, \alpha)$$

$$G_0 = MVN(\mathbf{0}, \Omega)$$

Note that this is equivalent to $\boldsymbol{\delta}_k \sim MVN(\mathbf{0}, \Omega)$, where $\boldsymbol{\delta}_k$ are cluster-level deviations from the community means. Species-level coefficients $\boldsymbol{\beta}_i^*$ can be derived as

$$\boldsymbol{\beta}_i^* = \boldsymbol{\beta} + \boldsymbol{\delta}_{k[g[i]]},$$

where g*[i]* is the cluster identity of species *i*, estimated using the cluster probabilities generated under the Stick Breaking Algorithm.

The number of dimensions of the *MVN* and thus the DP is determined by the number of parameters that are modeled with random species effects. As an example, when the intercept and all coefficients for *m* covariates are modeled as having species-level random effects, then the multivariate DP for $\boldsymbol{\delta}_i^*$ has $m + 1$ dimensions. Occupancy models are composed of an observational (detection) and an ecological (occupancy) component, and researchers are likely interested in understanding species similarities with respect to each component separately (i.e., which species are ecologically similar vs which species are detected similarly). We therefore

specified separate DPs for each model component. Though not necessary, this choice also allows for efficient priors for $\boldsymbol{\delta}_k$ (see Simulation study below).

Simulation study

To evaluate the effect of the dimensionality of the multivariate DP on the model's ability to recover community structure, we set up a simulation study. We simulated occupancy and detection data for a community of $n = 30$ species, grouped into $K = 5$ clusters (10, 8, 6, 4 and 2 species per cluster) across $J = 35$ sampling locations and $T = 5$ sampling occasions. We held detection probability constant across species, sites and occasions at $p = 0.24$ but allowed for cluster-specific intercepts and coefficients in the predictor of occupancy (not adding the DP structure to the detection component made the models run faster and thus made the simulation study viable). We considered 5 scenarios of dimensionality, using 0 to 4 predictor variables for occupancy, corresponding to $m = 1$ to 5 regression parameters (intercept and coefficient(s)) and, therefore, dimensions of the multivariate DP. Predictor variables were simulated as independent random variables following a *Normal*(0,1) distribution and we modeled their effect on occupancy probability using a probit link function. We set community hyperparameters (intercept followed by covariate coefficients) $\boldsymbol{\beta} = \{0{,}1, -0.5, 0.5, -1\}$. Following Johnson and Sinclair (2017), we modeled cluster-specific deviations from community level parameters, $\boldsymbol{\delta}_k$, as a *MVN*(0, Ω), where $\Omega = \omega^2(\mathbf{H}'\mathbf{H})^{-1}$, ω determines the amount of variation among cluster-specific coefficients and $\mathbf{H}$ is a $J \times m$ matrix of predictors measured at each sampling site (including an intercept term). This *MVN* corresponds to a g-prior (Tiao and Zellner 1964), which is often used for regression coefficients, because of its property that with a single parameter, ω, it controls the scale of variance and covariance based on the variance and correlation of predictor variables.

Because it is intuitive and has been shown (Johnson and Sinclair, 2017) that the differences among clusters influence how well a DP model can reproduce community structure, we considered three levels of among-cluster variation, ω = 1, 2 and 5, for each dimensionality scenario, yielding a total of 15 scenarios. We generated 50 data sets for each scenario, fitting the generated data to the above described DP-COM using the same covariates as the data-generating model.

We fit models in a Bayesian framework using a *Beta*(1,1) prior for p and priors suggested by Johnson and Sinclair (2017) for parameters of the DP component of the model, namely:

(1) for the DP concentration parameter α, a *Gamma*(*a, b*) prior where a and b are chosen depending on n so that $[k] \approx 1/k$, thus favoring smaller number of clusters (i.e., a more parsimonious model);

(2) for $\boldsymbol{\beta}$, a *MVN* g-prior with $\boldsymbol{\mu} = 0$ and $\Sigma = 10{,}000(\mathbf{X}'\mathbf{X})^{-1}$, where $\mathbf{X}$ is the design matrix for community-level effects and the specific multiplicative factor ensures sufficient variance to create a vague prior for our specific data.

(3) For ω, a scaled half-T distribution with φ=1 and df=1, which corresponds to a half-Cauchy prior distribution.

We simulated and analyzed data using the software R version 3.5.1 (R Core Team 2018). We fit models in JAGS 4.3.0 (Plummer 2003), accessed through the R package jagsUI 1.5.0 (Kellner 2019). We ran three parallel chains with 30,000 iterations of which we discarded 10,000 as burn-in. We thinned chains by 10 to reduce output size. We used the posterior mean as a point estimate, except for the number of clusters (K) and the number of sites occupied by species i (N_i, derived as $\sum_{j=1}^{J} z_{ij}$), for which we used the posterior mode (a more representative quantity in skewed posterior distributions typical for positive integer variables with small values). From model output we derived species-specific occupancy coefficients $\boldsymbol{\beta}_i^*$. We further calculated pairwise species clustering rates as the proportion of MCMC iterations in which two species were estimated to be in the same cluster. This $n \times n$ matrix can also be viewed as a species similarity matrix with respect to occupancy coefficients. We used the similarity matrix to calculate true and false pairwise clustering rates: first, we constructed a species-by-species matrix from the simulated data, in which species pairs received an entry of 1 if they were in the same cluster, and an entry of 0 otherwise. Then, we calculated the average pairwise clustering rate from the model output for all true species pairs (i.e., pairs with an entry of 1 in the data matrix) as true clustering rate, and the average pairwise clustering rate for all false species pairs (pairs with an entry of 0 in the data matrix) as false clustering rate.

We assessed convergence of parallel chains using the Gelman-Rubin statistic, R-hat (Gelman and Hill 2006). However, this statistic was not devised for a DP-type mixture model in which cluster labels switch (i.e., cluster 1 does not have the same identity throughout all iterations), and

as a result, cluster level parameters also switch. We were therefore more liberal in our assessment of convergence. We considered that we had achieved convergence when all structural parameters (α, ω, p, $\boldsymbol{\beta}$) as well as all species-level coefficients, $\boldsymbol{\beta}_i^*$, had an R-hat value <1.5 and excluded iterations that did not meet this criterion. We inspected chain plots for several cases of 1.1<Rhat<1.5 and found that generally, parallel chains fluctuated around the same average value, but that mixing was poor. Because these models are time intensive, we opted against running chains for more iterations, as this would have made the simulation study unfeasible.

To evaluate the performance of the DP model under the different scenarios, we calculated median bias (absolute bias, $\hat{x} - x$, for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_i^*$, because true values were often close/equal to 0; relative bias, $(\hat{x} - x)/x$ for all other parameters), median coefficient of variation (CV; posterior standard deviation divided by point estimate), median true and false clustering rates, and 95% Bayesian Credible Interval (BCI) coverage (percentage of iteration in which the 95% BCI included the true parameter value; henceforth just coverage) across all iterations that reached convergence. We used the median across iterations rather than the mean, because for some parameters, particularly the number of clusters K, the distribution across iterations was highly skewed, most likely due to poor identifiability particularly in scenarios with low ω and m.

To evaluate whether we lose accuracy in parameter estimates when using the normal-COM on a clustered community, we also ran a normal COM using the same data generated under the 15 scenarios described above and compared median bias and CV of estimates of N_i and $\boldsymbol{\beta}_i^*$ between the two approaches. We used the same g-priors for $\boldsymbol{\beta}$ and $\boldsymbol{\delta_k}$ (which correspond to $\boldsymbol{\delta}_i^*$ in the normal COM where each species forms its own cluster) and half-Cauchy prior on ω, the same MCMC settings and applied the same convergence criteria as for the DP-COM.

Application: Bird survey data from Uganda

Avian point-count data were collected from Murchison Falls National Park (MFNP) in western Uganda. The park covers nearly 4,000 km^2 in East Africa's Albertine Rift Valley, an area containing the highest vertebrate biodiversity on the African continent (Plumptre et al. 2007). Elevations in MFNP range from 620 m at the shore of Lake Albert to nearly 1,300 m in the southeast. The park experiences two rainy seasons (March – June and August – November), with an average annual rainfall of 1,100 mm.

297 Between 2010 and 2011, the Wildlife Conservation Society conducted bird surveys at elevations
298 ranging between 650 and 720 m. The Ugandan government recently granted access to MFNP for
299 oil exploration, and bird survey transects were established relative to the location of oil drilling
300 platforms with the goal of evaluating the effects of drilling activities on bird populations. The
301 survey area contained a mosaic of grasslands, dense and open borassus palm (*Borassus*
302 *aethiopum*) woodland, dense and open woodland, and bush habitat. Transects measuring 2000 m
303 were located in an easterly or westerly direction on either side of four oil-well pads (Appendix
304 S1: Figure S1). Twenty-one point-count locations were established along each transect. The first
305 point was located adjacent to the pad perimeter fence and subsequent points were spaced every
306 100 m. Transects were visited on average once every 2.5 days. Following a 2-minute rest period
307 upon arrival at a point-count location, the survey team leader (accompanied by 1 or 2 assistants)
308 recorded all birds seen or heard over a 5-minute period, within an estimated radius of 500 m.
309 Data collected included time of day, number of each bird species detected, estimated distance to
310 observer, elevation of point-count location and habitat type. Surveys took place between
311 February and September of 2010 and March to June of 2011. We selected a subset of the data
312 that included 62 survey dates between 22 February to 4 May 2010 (corresponding to the early
313 wet season). During that time, 149 points were visited at least once, with a mean of 23.3 (SD =
314 3.2) visits per point, resulting in 3464 visits across all points. We assumed that bird populations
315 were demographically closed during this period.

316 For each sampling location, we classified habitats into a binary variable of either open habitat
317 (grassland, bush; 72 locations) or woodland (Borassus and other woodland; 77 locations) and
318 determined the distance to the nearest oil well. In addition, for each visit, we had information on
319 observer experience. This was evaluated qualitatively by the lead field investigator (AJP) based
320 on years of experience, ability to identify species by call and accuracy in determining number of
321 individuals and distance from observation point. Although all observers were competent in
322 species identification, there was variation in experience and lead observers were ranked from 1
323 to 3, as most to least experienced, respectively.

324 To construct a species level detection-non-detection matrix, we considered each visit a sampling
325 occasion and reduced observations to binary species-level detection-non-detection data. We
326 excluded species from analysis that had fewer than 5 observations, resulting in a data set of 166

species. Species were encountered, on average, during 121 (SD 206) visits, at 39 (SD 35) sampling locations.

We included the binary habitat information (open versus closed) and scaled distance to oil well as covariates on occupancy probability. Detection probability was modeled as a function of the experience of the survey team leader; because implementing the DP community occupancy model was very time consuming and our dataset had many occasions, we calculated the average experience score of a site across all visits to avoid having to model detection probability as varying by occasion. The resulting values were almost binary (either 2 or >2); we therefore included average experience as a binary covariate on detection probability (2 = intermediate experience; >2 = high experience). We modeled occupancy intercept and regression coefficients as species specific, with a multivariate DP (see below). We modeled the detection intercept with a normal random effect and the effect of observer experience on detection as fixed across all species. We opted for a normal random effect in the detection component because our simulations indicated that a univariate DP performed poorly at estimating the cluster structure of a community (see Results). Our model ignored the potential spatial autocorrelation in occupancy stemming from the surveys recording birds up to 500 m from survey points spaced 100 m apart. In practice, 94% of all observations in our datasets were within 200 m from the survey point, and 78% were within 100 m. As this case study serves to demonstrate the DP-COM, rather than as an in-depth analysis of bird community ecology, we felt comfortable with the choice to ignore spatial autocorrelation.

For parameters of the occupancy component, we used the same priors as described for the simulation study, except that we set the multiplicative factor for the g-prior on $\boldsymbol{\beta}$ to 100,000 (to avoid overly low values in the prior variance-covariance matrix). We used a *Normal*(0, 10) prior on the mean and a *Gamma*(0.01, 0.01) prior on the standard deviation of the normal random effect on the intercept of probit(p). To improve computational speed, we used an upper bound of 100 for K. Imposing an upper bound on K is an accepted approximation of the infinite-mixture DP as long as it is set sufficiently high (Reich and Bondell 2011). Upper 95 BCI limits for the estimate of K were well below 100 (see Results), suggesting our choice of this upper limit did not affect estimates.

We implemented the models using the same software as for the simulation study. We ran three parallel chains with 50,000 iterations, of which we discarded 20,000 as burn-in. We thinned the remaining iterations by 30 (to avoid unwieldy model output). All model parameters except 5 δ_k and 1 β had Rhat<1.5 and in spite of these convergence issues, all species-specific β_i^* had Rhat $\leq$ 1.1. As we focus on species-level parameters and species similarity, these convergence problems should not impact our inference. Running this model took about 5.5 days on an IBM HS22 virtual BladeCenter server with an allocation of 3 logical cores using Intel Xenon E5645 processors at 2.4GHz and 1 GB RAM running ESXi. Further, we fit a normal-COM with both covariates to the data and compared estimates of $\boldsymbol{\beta}_i^*$ and N_i.

Finally, we explored the information provided by the DP-COM on bird community structure: First, to provide context for the amount of clustering suggested by the DP models, we compared the estimated number of clusters as well as the average pairwise clustering rate across the community to the respective expected values if species clustered at random. We generated these numbers by simulating draws from a Multinomial distribution with K=100 categories and equal cell probabilities ($\pi = 1/K$). The number of categories with at least one species corresponds to the number of random clusters. For each simulated set of cluster identities, we constructed a pairwise species clustering matrix, as described above. We simulated 3,000 such multinomial draws. We present the mean, SD, and range for the number of clusters; and the average (across all species) proportion of iterations that two species fell in the same cluster. Further, we contrasted average pairwise clustering rate of all families with at least 5 member species against community-wide average pairwise clustering rate, to investigate whether closely related taxa tended to respond to covariates more or less similarly than the entire community.

Results

Species in the simulated data sets occupied, on average (across species, iterations and scenarios), 17.06 of the 35 sites (average range: 9.23 – 24.87). They were detected, on average, 20.43 times (average range: 8.75 – 34.17) and at 12.71 sites (average range: 5.98 – 20.01). Across scenarios, for the DP-COM we excluded between 1 and 13 of the 50 iterations due to convergence problems; the number of excluded iterations increased with increasing number of parameters m and decreasing among-cluster variation ω. In comparison, for the normal COM we excluded

between 0 and 9 iterations due to convergence issues (Appendix S2: Table S1). When using the customary cut-off of $\text{Rhat} \leq 1.1$ for the normal COM, this number rose to between 0 and 23 iterations (Appendix S3).

Community structure, species similarity in simulated communities

Bias in estimates of *K* generally decreased with increasing ω (i.e., with increasing variation among clusters) and *m* (i.e., dimensions of the DP) (Figure 1a). For ω = 1, *K* (true value of 5) was consistently underestimated, with a median estimate of 1 (*m* = 1) to 2 (*m* > 1) clusters. For almost all other scenarios, *K* was overestimated, up to $\hat{K} = 12$. At ω = 5 and $m \geq 2$, the median estimate of *K* was consistently at 7, but variability in estimates across iterations decreased with increasing *m*. Precision of estimates of *K* increased with increasing ω (from a maximum CV of 2.38 at ω = 1 to a minimum of 0.29 at ω = 5). There was no evident relationship of the CV with *m*. Finally, coverage was nominal or near nominal (at least 93%) for all scenarios with ω < 5 but dropped to between 80 and 88% for ω = 5 and $m \geq 2$ (Appendix S2: Table S2).

The rate at which two species were correctly estimated as being in the same cluster (true clustering rate) ranged from 42% to 76% across scenarios (Figure 1b; Appendix S2: Table S2). The highest true clustering rate was attained at ω = 1 and *m* = 1; however, the rate at which two species were incorrectly estimated to be in the same cluster (false clustering rate) for that scenario was almost as high (72%), consistent with an average estimate of *K*=1 for this scenario. For most other scenarios, true clustering rate was <60%. Only for ω = 5 did the true clustering rate tend to increase with increasing *m*, and within ω = 5, only for $m \geq 3$ did the correct clustering rate exceed 60%. False clustering rate decreased with increasing ω, ranging from 36% to 72% at ω = 1, from 21% to 38% at ω = 2, and from 6% to 18% at ω = 5. Only for ω = 5 did the false clustering rate continuously decrease with increasing *m*.

Occupancy in simulated communities

Across all scenarios, the number of sites occupied by a given species was estimated without bias (Figure 2a), though in some rare species-iteration combinations, bias reached 100%. The median CV of the number of sites occupied ranged from 9% to 15%; values decreased slightly with increasing *m* and decreasing ω. (Figure 2a) The incidence of extreme CVs (at or above 100%) for specific species-iteration combinations increased with increasing ω. Coverage was nominal

for all scenarios (Appendix S2: Table S3).

Estimates of community level regression coefficients **β** showed low to moderate absolute bias. For example, depending on the scenario, the median estimate of the community intercept (true value of 0) ranged from -0.05 to 0.15, with most scenarios having median estimates <|0.10|. There were no apparent patterns in bias with respect to *m*, but bias tended to increase with increasing ω. Estimates were less precise (i.e., had higher CVs) with increasing *m*, except for the community intercept. Coverage was nominal for all parameters and scenarios (Appendix S2: Table S4).

Median bias (across species and iterations) in species-specific regression coefficients $\boldsymbol{\beta}_i^*$ was low to moderate. Median bias, as well as the incidence (i.e., particular species-iteration combinations) of strong bias, increased with increasing ω and increasing *m*, though the latter was less pronounced. CVs increased with increasing ω, with the exception of the intercept, where CVs decreased with increasing ω. There was no discernable relationship between CVs and *m*. Coverage ranged from 89% to 97% and increased with increasing ω (see Appendix S2: Figure S1 for an example plot and Table S5 for details of simulation results).

Comparison with normal COM in simulated communities

Bias and CV for estimates of N_i were very similar across the DP and the normal COM (Figure 2, Appendix S2: Table S6); across scenarios, the DP model tended to have lower median CVs but only by 1 or 2 percentage points. For **β**, median bias was similar between both models across parameters and scenarios, but parameters from the DP model had considerably higher CVs (Appendix S2: Figure S2, Table S7). For $\boldsymbol{\beta}_i^*$, both bias and CVs were very similar between the two modeling approaches (Appendix S2: Figure S3, Table S8). These patterns were the same when applying the $\text{Rhat} \leq 1.1$ cut-off to the normal COM results (Appendix S3).

Other parameters: ω, α and p in simulated communities

Detection probability *p* was estimated with minimal bias (-2% to 1%), 4-5% CV and BCI coverage between 86% and 97% (Appendix S2: Table S9). Median estimates of the DP concentration parameter α ranged from 4.29 (ω = 5 and *m* = 5) to 12.64 (ω = 2, *m* = 2) (Appendix S2: Table S10).

Estimates of ω were most biased for ω = 1 (-28% to -84%). For all other scenarios, relative bias was low to moderate, ranging from -2% to 16%. The CV of ω increased with increasing ω and *m*. Coverage ranged from 87% to 100%, except for ω = 1 and *m* = 1, where coverage was 0% (Appendix S2: Table S11).

Bird case study

For 5 out of 300 δ_k and one β, R-hat > 1.5; however, all (derived) species-specific regression coefficients $\boldsymbol{\beta}_i^*$ had R-hat<1.1. We visually checked chains for the non-converged δ_k and β, which appeared to be strongly autocorrelated but oscillated around the same average value; we therefore felt confident to use the estimates.

For the occupancy component of the model (with a multivariate DP for the coefficients of the probit-linear predictor of occupancy probability), species comprised 27 clusters (SD = 4.16, 95% BCI 22 - 37; Figure 3). Probabilities of two species clustering together ranged from 0 to 0.92. The estimate of ω for the full model was 8.60 (SD 1.04, 95BCI 6.85 – 10.98), indicating considerable variation in regression coefficients among clusters.

The data set contained ten families with at least 5 species, comprising 90 species total. When looking at pairwise clustering rates for these families, we found that most families showed clustering probabilities similar to those of the entire community. However, the sunbirds (Nectariniidae, 5 species) had considerably higher clustering probabilities, whereas the Cisticolidae (12 species) and the bee-eaters (Meropidae, 5 species) had lower clustering probabilities (Figure 4).

Species were estimated to occupy between 1 and 147 of the 149 sample sites. We observed strong effects (i.e., with 95% BCI not overlapping 0) of woodland habitat for 57 species, with 24 negative and 33 positive coefficients. For distance from oil well, 14 species showed strong negative and 12 species showed strong positive effects (Figure 5). Species with positive associations with woodland habitat tended to have positive associations with distance to oil as well (52 species), and vice versa (72 species).

When comparing estimates of $\boldsymbol{\beta}_i^*$ and N_i between the DP-COM and the normal COM, both modeling approaches produced very similar results (Appendix S1: Figure S2).

Discussion

In wildlife research, DP priors have been used to model genetic population structure (Reich and Bondell 2011), spatial variation in abundance (Dorazio et al. 2008, Dorazio 2009), spatial clustering of population trends (Johnson et al. 2013), and clustering of species with respect to habitat coefficients in the context of community distribution models (Johnson and Sinclair 2017). Our simulation study showed that a community occupancy model with a DP, instead of the customary normal random species effect, was able to retrieve aspects of community structure when differences among clusters and the number of parameters making up the multinomial DP were sufficient. Applied to data for a bird community, the model led to a considerable reduction in the number of parameters estimated, grouping 166 species into 27 clusters. This suggests that detection/non-detection data contain information on the similarity of species in a community that can be exploited with a DP model. Major shortcomings of the approach were its computational expense, poor mixing and difficulty with convergence of MCMC chains due to label switching among clusters, and its reduced performance in retrieving community structure when cluster parameters were similar and/or few parameters were used in the DP. These drawbacks may appear particularly off-putting given that there are no a priori tests that would indicate whether the existence of, and differences among, clusters warrant exploring a "costly" DP-COM. Moreover, the customary model with a normal random effect performed similarly to the DP-COM, even when applied to data from a clustered community, suggesting that a normal random effect is flexible enough to capture variation in parameters that do not follow a normal distribution. For analyses focused on community and species-level responses in occurrence (and/or detection) to covariates, or simply the estimation of occupancy probabilities in the absence of covariates, we recommend the more efficient normal COM. Only the DP-COM, however, returns estimates of community structure and species similarity in their response to covariates; for analyses aimed at testing hypotheses regarding these measures, the additional time investment needed to fit a DP-COM seems worthwhile.

Factors affecting the performance of the DP-COM

We found that both the variability among clusters and the dimensionality of the DP affected the ability of the model to retrieve information on community structure. Median bias in K, the

incidence of large bias and the incidence of large CVs all declined with increasing number of dimensions of the DP; when variation among clusters was high ($\omega = 5$), increased dimensionality also led to higher true and lower false clustering rates. Across levels of among-cluster variation, univariate DPs did the worst in terms of clustering rates and estimating K. All of this indicates improved ability of the model to identify cluster identity of species with increased dimensionality. Estimates of community structure may not be reliable when only a single dimension is considered. As such, the DP-COM may be more useful for data sets with sufficient replication to support modeling of multiple covariates. It is possible, however, that if variation among clusters is stronger than what we considered in the simulation, a univariate DP may be able to identify clusters. Even though the effect of dimensionality on the performance of clustering algorithms is known (e.g., "curse of dimensionality"; Bellmann, 1957) and the DP is a widely used Bayesian clustering algorithm outside of wildlife research, to our knowledge this is the first study to demonstrate that the performance of the DP model is dependent on the number of dimensions of the base distribution.

Not surprisingly, we found that the variability among clusters strongly affected the ability of the DP-COM to estimate the number of clusters in the community, as well as pairwise species clustering rates. While increasing ω resulted in higher true clustering rates, lower false clustering rates and lower bias in K, it also resulted in increased bias and CV in most $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_i^*$ and higher incidences of extreme bias and CVs in N_i. There appears to be a trade-off between improvements in estimation of community structure and species similarities as a function of cluster discrimination and the accuracy of other parameters of ecological interest. Regardless, coverage of these parameters was nominal or near nominal across scenarios.

At $\omega = 1$, the DP-COM was essentially unable to detect cluster structure and, in most iterations, estimated that all species belonged to the same cluster (regardless, estimates of $\boldsymbol{\beta}_i^*$ and N_i were largely unbiased). Further, in our simulation, the actual number of clusters was, on average, not estimated well (median bias was mostly >40%), and coverage of the true value was <90% for scenarios that estimated K with the lowest bias (i.e., $\omega = 5$ and $m \geq 2$). Both findings contradict results by Johnson and Sinclair (2017), whose proof of concept simulation for a community Poisson regression resulted in accurate estimates of K for various values of ω, as long as $\omega > 0.5$. We implemented the DP on parameters of the occupancy component of the DP-COM, which is

binary and partially latent (only for sites where the species is detected, do we observe occupancy state). It is conceivable that the differences between clusters need to be more pronounced, and/or that it is generally more difficult for the DP algorithm to retrieve community structure for a binary partially latent process. Based on these results, we suggest interpreting the absolute estimated number of clusters with caution and focus instead on estimates of species similarity.

We only explored two factors likely to affect the performance of the DP-COM, though many other factors may be influential. Particularly, we imagine that the total community size and the cluster-to-size ratio (i.e., whether communities consist of many small or few large clusters) may affect the estimation of community structure: we would expect that more clusters should improve estimation of parameters governing G_0, and more species per cluster should improve estimates of cluster-level parameters. Additional simulations with communities of 60 species distributed across 5 or 10 clusters (i.e., representing a scenario with a higher species-to-cluster ratio, and one with the same ratio as in our main simulation but with more data) somewhat support these expectations, with community β coefficients having slightly lower CVs in the scenario with more clusters, and species-level coefficients (which are derived from cluster-level estimates) having slightly lower CVs when there were more species per cluster (Appendix2: Table S12). Having a larger community reduced bias in community and species coefficients, regardless of the community structure. Neither scenario, however, suggested that using the DP over a normal COM led to greater improvements in either precision or bias of estimates when communities were larger (Appendix S2: Figure S4 and S5). Factors of study design, such as spatial and temporal repeats, as well as the amount of data available for each species have been shown to affect performance of occupancy models (MacKenzie and Royle 2005, Pacifici et al. 2014) and may affect the DP-COM as well. Due to the computational cost of the DP-COM, however, we were unable to explore these additional dimensions in more depth.

Accuracy of parameter estimates

Whereas estimates of typical parameters of interest (number of sites occupied, coefficients of the probit-linear predictor of occupancy) were, on average, unbiased under both modeling approaches, bias and CV were high in some individual species-iteration combinations, particularly in estimates of species-specific coefficients (Appendix2: Figure S1). Even though the DP-COM adequately reflected the clustered nature of the simulated communities, it did not

consistently improve bias and precision of estimates. We performed some exploratory post-hoc analyses (results not shown) that showed that specific species-iteration combinations had consistently high CV and bias across both modeling approaches, suggesting that some characteristic of the data was responsible for poor estimates. We investigated whether instances of large CVs and bias were associated with sparse data, but patterns were inconclusive. We do not have data on bias and precision of parameter estimates under the two modeling approaches fit to data generated under a normal COM (i.e., a non-clustered community), but we suspect that the incidences of high CVs and bias are related to the clustered structure of the community.

Sensitivity to prior

It has been shown that the estimate of the concentration parameter α, which determines the number of clusters, is sensitive to its prior (Dorazio 2009). Following the principle of preferring parsimonious models, we adopted the prior by Johnson and Sinclair (2017), which allows for a wide range of values of K but favors smaller values and did not appear to affect estimates of K in their simulation. Nonetheless, under most scenarios, we observed positive bias in $\hat{K}$. Because of the time-intensive nature of the DP model, thorough testing of sensitivity of $\hat{\alpha}$ and, by extension, $\hat{K}$, to priors was beyond the scope of this study. For a small subset of simulations, however, we explored whether a negative-exponential prior, which puts even more weight on fewer clusters, would reduce the positive bias in $\hat{K}$, but found no improvements. Dorazio (2009) suggested a *Gamma*(a, b) prior where a and b are chosen depending on n (the number of species in the data set), so that the prior on α reflects a *discrete-Uniform*(0, n) prior on K. On the other hand, West et al. (1994) suggest a static *Gamma*(3.5, 0.5) prior allowing for a wide range of possible values for K, with low probability at 0 and n. Studies employing the DP-COM should evaluate the influence of the choice of prior for α on main quantities of interest.

Structure and habitat associations in the MFNP. bird community

We found considerable structure within the MFNP bird community, with 166 species clustering into 27 groups regarding their associations with habitat type and distance to oil well. Some species pairs showed similarity scores >0.90, being in the same cluster virtually all the time.

Across the community, we found that occupancy of more species was significantly related to

habitat type (open versus woodland) than influenced by distance from oil wells. It is conceivable that the effect of oil drilling operations on bird occurrence may be temporally limited to when wells are active (Fuda et al. 2018). Our analysis of occupancy across multiple months may mask any such temporal effects and only show effects of this factor in cases of strong species responses. Coefficients of the two predictors of occupancy were positively correlated, indicating that an increasing preference of woodland habitat corresponded with greater avoidance of oil wells. No species that had strong negative associations with distance to oil wells had strong positive associations with woodland habitat; similarly, none of the species strongly preferring woodland habitat had significant negative associations with distance to oil well. This suggests that birds with a preference for more closed habitat tend to be more sensitive to habitat disturbance, a conclusion reached for birds and other taxa in a recent meta-analysis (Keinath et al. 2017).

The mechanisms determining how closely related species, possibly with similar morphologies and diets, can coexist has been an ongoing debate in ecology for decades (Hutchinson 1959, Gotelli 2000, Graham et al. 2004). Many studies have found that sister taxa commonly occupy different ecological niches and that co-occurring species are generally more distantly related (e.g., Silva et al. 2014), while others argue that phylogeny begets morphological similarity and, thus, higher likelihood of niche overlap (Gonçalves-Souza et al. 2014). These patterns are of interest in conservation biology as well, with recent studies exploring the effect of phylogeny and other traits on species susceptibility to disturbance (Nowakowski et al. 2017). When comparing within-family clustering rates – a measure of how similarly species in the present study use space – to average similarity of the community, only the sunbirds stood out as more similar than average. The tropical sunbirds are largely nectivorous but also consume fruit and insects and, thus, generally considered to be forest/woodland species where their specialized food sources are likely more plentiful (Cheke et al. 2019). The five species represented in our sample demonstrated significant niche conservativism, with consistently strong positive associations with woodland habitat and distance to oil (only one species, the Marico sunbird *Cinnyris mariquensis*, had 95BCI overlapping 0 for the latter). Thus, based on our findings the sunbirds represent an example of where "phylogeny begets niche overlap", and possibly of low response diversity with respect to anthropogenic influence (oil wells). Of course, habitat partitioning among these species may very well happen on scales other than the one measured in

this study. Even though species-specific coefficient estimates were generally very similar between the DP and the normal COM, sunbird coefficients were more similar to each other under the DP-COM than the normal COM, suggesting that the approach is better able to represent similarities among species (Appendix S1: Figure S2).

In contrast, members of the Cisticolidae and Meropidae (bee-eaters), with 12 and 5 representative species respectively, demonstrated clustering probabilities that were lower than average, and only slightly higher than expected under random clustering. Thus, they exemplify the "niche differentiation among closely related taxa" argument. Bee-eaters are considered habitat generalists, occupying both forests, edge and open habitat; their aerial behavior is generally independent of vegetation type (Fry 2019). Cisticolidae is a broad taxon that includes Old World warblers and other allies that occupy a range of habitats including forest, open woodland, scrub and grassland (Ryan 2019). This example illustrates the potential usefulness of the DP-COM for addressing ecological questions of species coexistence, estimating similarity of species while fully accounting for imperfect species detection and uncertainty in coefficient estimates.

Conclusion

Dirichlet process distributions provide a flexible tool to model latent structure in wildlife communities and populations. Our DP-COM is a straight-forward extension of popular community occupancy models (e.g., Zipkin et al. 2009, Ruiz-Gutiérrez et al. 2010, Sollmann et al. 2017) and can be implemented in JAGS, a software that has become increasingly popular among ecologists and wildlife researchers (Kéry 2010, Kéry and Schaub 2012). Implementing the DP-COM in JAGS was computationally much more expensive than the normal COM – for the bird data set, the difference was on the scale of hours (normal COM) versus >5 days (DP-COM). Based on our simulation study, run time increases non-linearly with the addition of species to the data set (from 15 minutes for a 30-species community to 1.5 hours for a 60-species community). Mixing of chains was slow, suggesting that longer chains, and thus more computation time, would be beneficial. Whereas implementation of these models can be accelerated by using a custom MCMC algorithm, and likely also by using the reversible jump MCMC capacities of NIMBLE (de Valpine et al. 2017), they still remain computationally involved (Johnson and Sinclair 2017). This complicates thorough evaluation of model

performance under different conditions via simulations and makes models less accessible to practitioners. Even though the DP model has fewer parameters than the normal COM, the improvement in precision of estimates was marginal or non-existent, and in spite of the distinctly clustered simulated community, the normal COM returned estimates of ecological parameters that were, for the most part, as accurate and precise as those of the DP-COM. For studies where estimates of occupancy and associated covariate coefficients are the main focus, our results thus suggest the much faster and better-mixing normal COM provides reliable results. We did not test whether joint prediction of community occupancy at new sites benefits from the DP-COM, and this warrants further investigation for studies where prediction is a key objective. The DP-COM may be the better approach in situations where researchers would otherwise resort to a priori grouping of species. Especially for sparse data species, inference on the species level is affected by how groups are defined (Pacifici et al. 2014); under a DP-COM, estimates of parameters for such species will represent the average over possible group associations and thus avoid subjectivity in choosing a certain grouping. The main advantage of the DP-COM is the information about community structure and species similarity with respect to occupancy predictors that the normal model cannot provide directly. We present an example of how this information can be used to address questions of ecological relevance with the Uganda bird example.

Our model development only considers the simplest case of a DP model, in which no information on species cluster membership is available. The DP model can be extended to include covariates that can inform the probability of cluster membership (Johnson et al. 2013). In the context of community occupancy models, inclusion of potential clustering covariates enables testing whether species attributes such as taxonomy or functional traits explain community structure. As such, in spite of its drawbacks, the semi-parametric DP-COM holds potential as a flexible modelling approach in situations where community structure and species similarities are of primary interest.

Acknowledgments

Data collection was made by the Wildlife Conservation Society and financed by the USAID/WILD Program. We are particularly grateful to the ornithologists who collected the data, notably Hamlet Mugabe, Dennis Tumuhamye, and Taban Bruhan. We are also grateful for

permission to undertake the research by the Uganda Wildlife Authority. We further thank Paul Conn (NOAA) for suggesting the conceptual DP-COM. Data and code storage on Dryad is sponsored by UC Davis. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Author contributions: PM, SA and SP oversaw data collection in the field, led the field teams or made regular visits to check on the data collection and to supply field teams with food and organize other logistical arrangements. AJP and SP conceived of the project, secured funding, and designed the study. RS, MJE and WAL conceived of the analytical approach for the paper, RS, DSJ and WAL led the model and simulation study development, RS performed the data analysis, RS and DSJ led simulation results interpretation, while RS and MJE led case study interpretation and writing of the manuscript, with input from all coauthors.

Supporting Information

Additional supporting information may be found online at: [link to be added in production]

Data Availability

Data and code can be accessed on the Dryad Digital Repository: https://doi.org/10.25338/B8GG8P

References

Azizyan, M., A. Singh, and L. Wasserman. 2013. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. Pages 2139–2147 Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Lake Tahoe, Nevada.

Bellman, R. 1957. Dynamic programming. Princeton University Press. Princeton, NJ.

Cheke, R., C. Mann, and A. Bonan. 2019. Sunbirds (Nectariniidae). Page *in* J. del Hoyo, A. Elliott, J. Sargatal, D. A. Christie, and E. de Juana, editors. Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona, Spain.

Dorazio, R. M. 2009. On selecting a prior for the precision parameter of Dirichlet process mixture models. Journal of Statistical Planning and Inference 139:3384–3390.

Dorazio, R. M., M. Kéry, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic

metacommunity systems. Ecology 91:2466–2475.
Dorazio, R. M., B. Mukherjee, L. Zhang, M. Ghosh, H. L. Jelks, and F. Jordan. 2008. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. Biometrics 64:635–644.
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.
Dorazio, R. M., J. A. Royle, B. Söderström, and A. Glimskär. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology 87:842–854.
Dunstan, P. K., S. D. Foster, and R. Darnell. 2011. Model based grouping of species across environmental gradients. Ecological Modelling 222:955–963.
Dunstan, P. K., S. D. Foster, F. K. Hui, and D. I. Warton. 2013. Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. Journal of agricultural, biological, and environmental statistics 18:357–375.
Escobar, M. D., and M. West. 1995. Bayesian density estimation and inference using mixtures. Journal of the american statistical association 90:577–588.
Fry, H. 2019. Bee-eaters (Meropidae). Page *in* J. del Hoyo, A. Elliott, J. Sargatal, D.A. Christie, and E. de Juana, editors. Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona, Spain.
Fuda, R. K., S. J. Ryan, J. B. Cohen, J. Hartter, and J. L. Frair. 2018. Assessing the impacts of oil exploration and restoration on mammals in Murchison Falls Conservation Area, Uganda. African Journal of Ecology 56:804–817.
Gelman, A., and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. First edition. Cambridge University Press, New York, USA.
Gonçalves-Souza, T., J. A. F. Diniz-Filho, and G. Q. Romero. 2014. Disentangling the phylogenetic and ecological components of spider phenotypic variation. PloS one 9:e89314.
Gotelli, N. J. 2000. Null model analysis of species co-occurrence patterns. Ecology 81:2606–2621.
Graham, C. H., S. R. Ron, J. C. Santos, C. J. Schneider, and C. Moritz. 2004. Integrating

phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58:1781–1793.

Houle, M. E., H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. 2010. Can shared-neighbor distances defeat the curse of dimensionality? Pages 482–500 International Conference on Scientific and Statistical Database Management. Springer.

Hutchinson, G. E. 1959. Homage to Santa Rosalia or why are there so many kinds of animals? The American Naturalist 93:145–159.

Johnson, D. S., R. R. Ream, R. G. Towell, M. T. Williams, and J. D. L. Guerrero. 2013. Bayesian clustering of animal abundance trends for inference and dimension reduction. Journal of Agricultural, Biological, and Environmental Statistics 18:299–313.

Johnson, D. S., and E. H. Sinclair. 2017. Modeling joint abundance of multiple species using Dirichlet process mixtures. Environmetrics 28:e2440.

Keinath, D. A., D. F. Doak, K. E. Hodges, L. R. Prugh, W. Fagan, C. H. Sekercioglu, S. H. Buchart, and M. Kauffman. 2017. A global analysis of traits predicting species sensitivity to habitat fragmentation. Global Ecology and Biogeography 26:115–127.

Kellner, K. 2019. jagsUI: A Wrapper Around "rjags" to Streamline "JAGS" Analyses. https://CRAN.R-project.org/package=jagsUI

Kéry, M. 2010. Introduction to WinBUGS for ecologists: A Bayesian approach to regression, ANOVA and related analyses. Academic Press, Burlington, MA.

Kéry, M., and J. A. Royle. 2008. Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. Journal of Applied Ecology 45:589–598.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press, Burlington, MA.

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

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2017. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. 2nd edition. Elsevier, Amsterdam.

MacKenzie, D. I., and J. A. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. Journal of Applied Ecology 42:1105–1114.

783 Mori, A. S., T. Furukawa, and T. Sasaki. 2013. Response diversity determines the resilience of
784 ecosystems to environmental change. Biological Reviews 88:349–364.
785 Nowakowski, A. J., J. I. Watling, S. M. Whitfield, B. D. Todd, D. J. Kurz, and M. A. Donnelly.
786 2017. Tropical amphibians in shifting thermal landscapes under land-use and climate
787 change. Conservation Biology 31:96–105.
788 Ohlssen, D. I., L. D. Sharples, and D. J. Spiegelhalter. 2007. Flexible random-effects models
789 using Bayesian semi-parametric models: applications to institutional comparisons.
790 Statistics in Medicine 26:2088–2112.
791 Pacifici, K., E. F. Zipkin, J. A. Collazo, J. I. Irizarry, and A. DeWan. 2014. Guidelines for a
792 priori grouping of species in hierarchical community models. Ecology and Evolution
793 4:877–888.
794 Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs
795 sampling. Pages 20–22 Proceedings of the 3rd International Workshop on Distributed
796 Statistical Computing (DSC 2003).
797 Plumptre, A. J., T. R. Davenport, M. Behangana, R. Kityo, G. Eilu, P. Ssegawa, C. Ewango, D.
798 Meirte, C. Kahindo, M. Herremans, and others. 2007. The biodiversity of the Albertine
799 Rift. Biological conservation 134:178–194.
800 R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for
801 Statistical Computing, Vienna, Austria. http://www.R-project.org /
802 Reich, B. J., and H. D. Bondell. 2011. A spatial Dirichlet process mixture model for clustering
803 population genetics data. Biometrics 67:381–390.
804 Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology.
805 Academic Press, London, UK.
806 Ruiz-Gutiérrez, V., E. F. Zipkin, and A. A. Dhondt. 2010. Occupancy dynamics in a tropical bird
807 community: unexpectedly high forest use by birds classified as non-forest species.
808 Journal of Applied Ecology 47:621–630.
809 Ryan, P. 2019. Cisticolas and allies (Cisticolidae). Page *in* J. del Hoyo, A. Elliott, J. Sargatal,
810 D.A. Christie, and E. de Juana, editors. Handbook of the Birds of the World Alive. Lynx
811 Edicions, Barcelona, Spain.
812 Sauer, J. R., and W. A. Link. 2002. Hierarchical modeling of population stability and species
813 group attributes from survey data. Ecology 83:1743–1751.

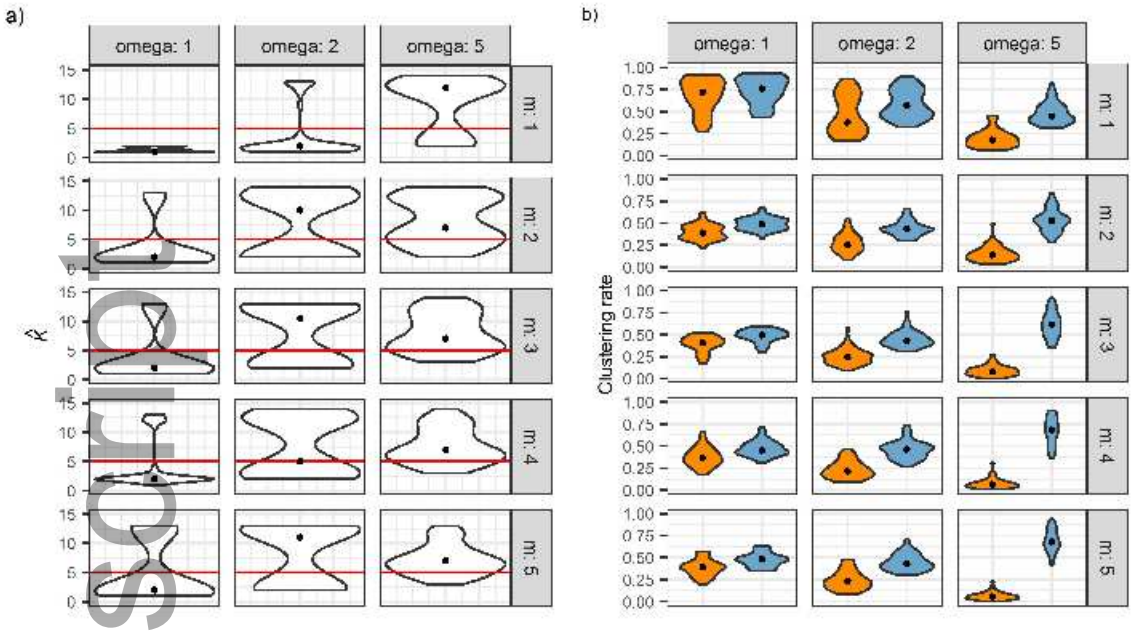
814 Sethuraman, J. 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4:639–650.
815 Silva, D. P., B. Vilela, P. De Marco Jr, and A. Nemesio. 2014. Using ecological niche models
816 and niche analyses to understand speciation patterns: the case of sister neotropical orchid
817 bees. PLoS One 9:e113246.
818 Sollmann, R., B. Gardner, K. A. Williams, A. T. Gilbert, and R. R. Veit. 2016. A hierarchical
819 distance sampling model to estimate abundance and covariate associations of species and
820 communities. Methods in Ecology and Evolution 7:529–537.
821 Sollmann, R., A. Mohamed, J. Niedballa, J. Bender, L. Ambu, P. Lagan, S. Mannan, R. C. Ong,
822 A. Langner, B. Gardner, and Wilting, Andreas. 2017. Quantifying mammal biodiversity
823 co-benefits in certified tropical forests. Diversity and Distributions 23:317–328.
824 Tiao, G. C., and A. Zellner. 1964. Bayes's theorem and the use of prior knowledge in regression
825 analysis. Biometrika 51:219–230.
826 de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. 2017.
827 Programming with models: writing statistical algorithms for general model structures
828 with NIMBLE. Journal of Computational and Graphical Statistics 26:403–413.
829 Yamaura, Y., M. Kery, and J. A. Royle. 2016. Study of biological communities subject to
830 imperfect detection: bias and precision of community N-mixture abundance models in
831 small-sample situations. Ecological Research 31:289–305.
832 Zipkin, E. F., A. DeWan, and A. J. Royle. 2009. Impacts of forest fragmentation on species
833 richness: a hierarchical approach to community modelling. Journal of Applied Ecology
834 46:815–822.
835 Figure legends
836 Figure 1: Estimated number of clusters *K* (a) and pairwise clustering rates (b) from a Dirichlet
837 Process (DP) community occupancy model used to analyze data simulated under different levels
838 of cluster distinctiveness (ω) and different number of coefficients in the probit-linear predictor of
839 occupancy (corresponding to dimensions of the multivariate DP), *m*. For a), violin plots depict
840 posterior modes of *K* across iterations; red line shows the data generating value. For b), violins
841 show the average number of MCMC iterations during which two species were estimated to be in
842 the same cluster when in the simulated data they were in the same cluster (blue) and when they
843 were in different clusters (orange). In both panels, dots represent the median across iterations.
844

Figure 2: Bias (a) and coefficient of variation, CV, (b) of estimated number of sites occupied by species from community occupancy models using either a Dirichlet Process (DP) or a normal species level random effect. Models were used to analyze data simulated under different levels of cluster distinctiveness (ω) and different number of coefficients in the probit-linear predictor of occupancy (corresponding to dimensions of the multivariate DP in the DP COM), *m*. Violins represent estimates across species and iterations in a given scenario. Plot y-axes capped at -1/1 (a) and 0/1 (b) for aesthetic reasons.

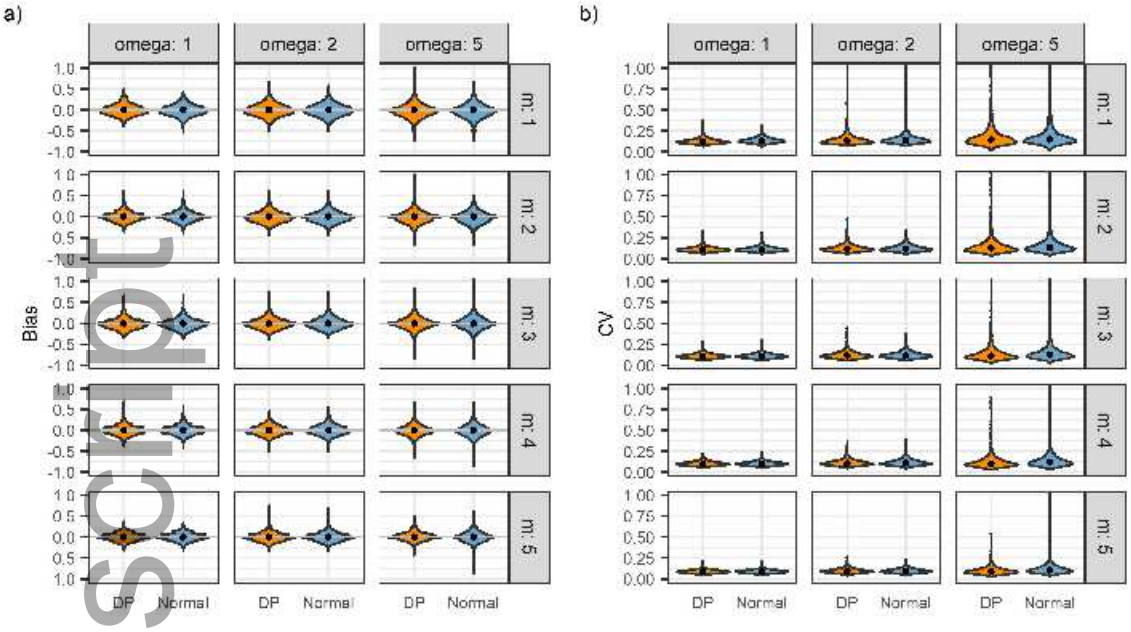
Figure 3: Probability of joint cluster membership for 166 birds in Murchison Falls National Park, Uganda, estimated from a Dirichlet Process community occupancy model, based on coefficients in the probit-linear predictor of occupancy probability, including the effect of habitat type (open versus woodland) and distance from oil well. Both axes represent species identity and color gradient expresses the probability of joint cluster membership.

Figure 4: Pairwise probabilities of joint cluster membership (similarity), estimated from a Dirichlet Process community occupancy model, for 10 bird families with at least 5 species observed during a survey in Murchison Falls National Park, Uganda (number of species given above error bars). Dots: average probabilities of joint cluster membership across species; error bars: 5^{th} and 95^{th} percentiles; black line/grey rectangle: mean and 5^{th} and 95^{th} percentile of probabilities of joint cluster membership for entire community; red line: maximum clustering probability observed when simulating random clustering.

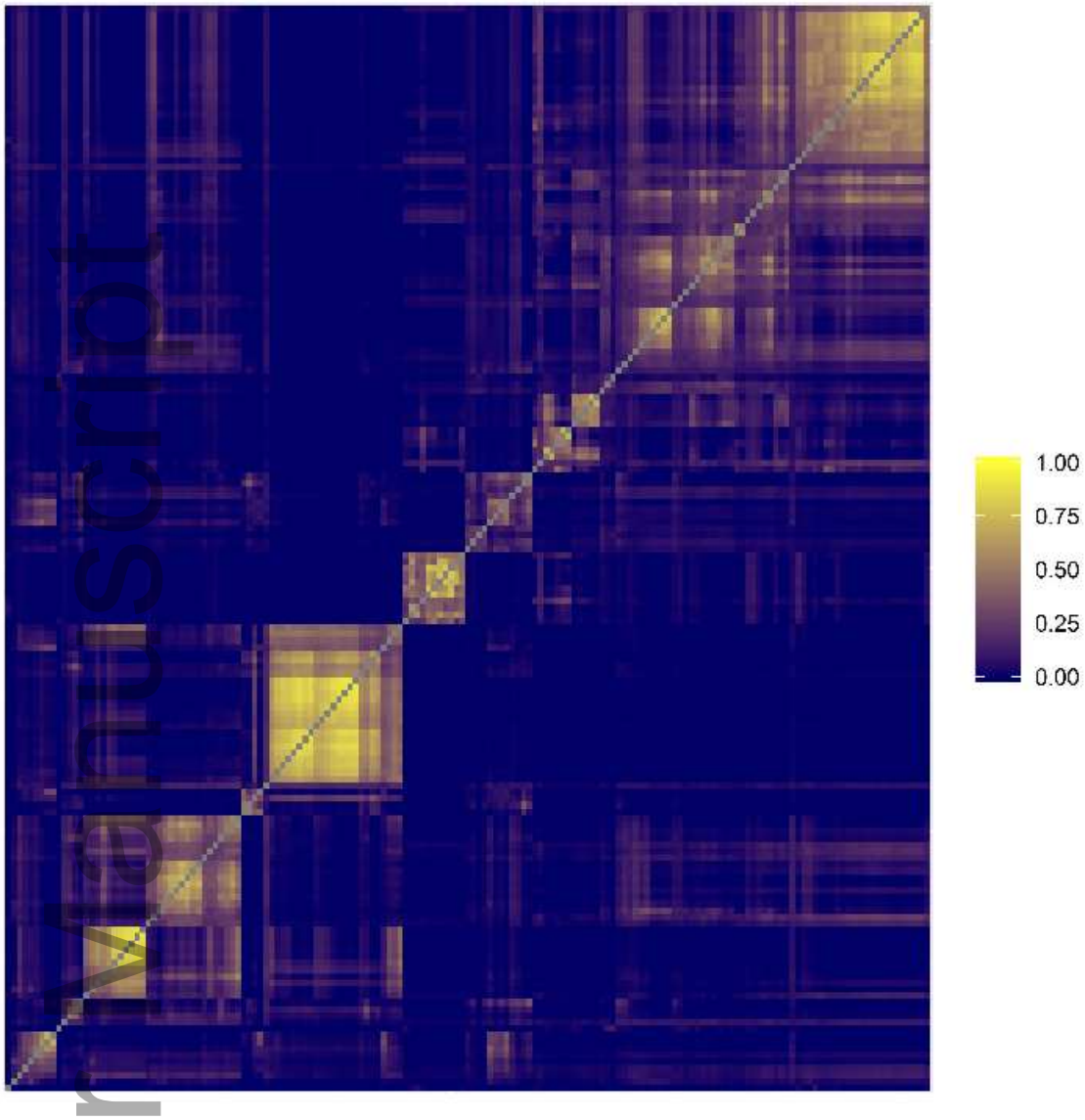
Figure 5: Beta coefficients for effect of woodland habitat, β(habitat), and distance to oil well, β(Oil), on occupancy probability for 166 birds surveyed in Murchison Falls National Park, Uganda, estimated with a Dirichlet Process community occupancy model. Effects considered strong when 95% Bayesian Credible Intervals did not overlap 0.



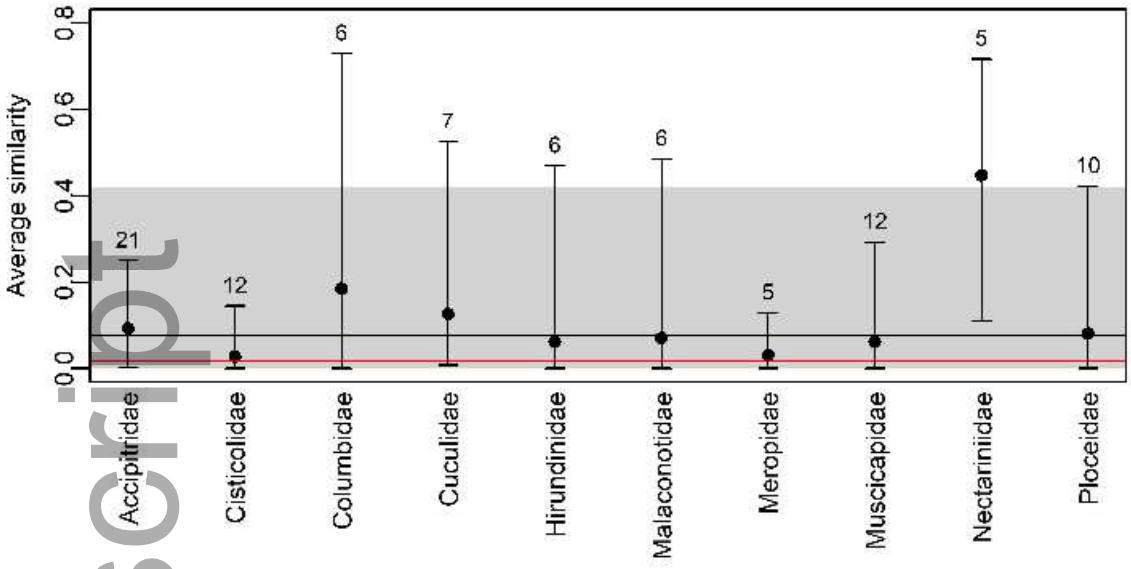
eap_2249_f1.jpg



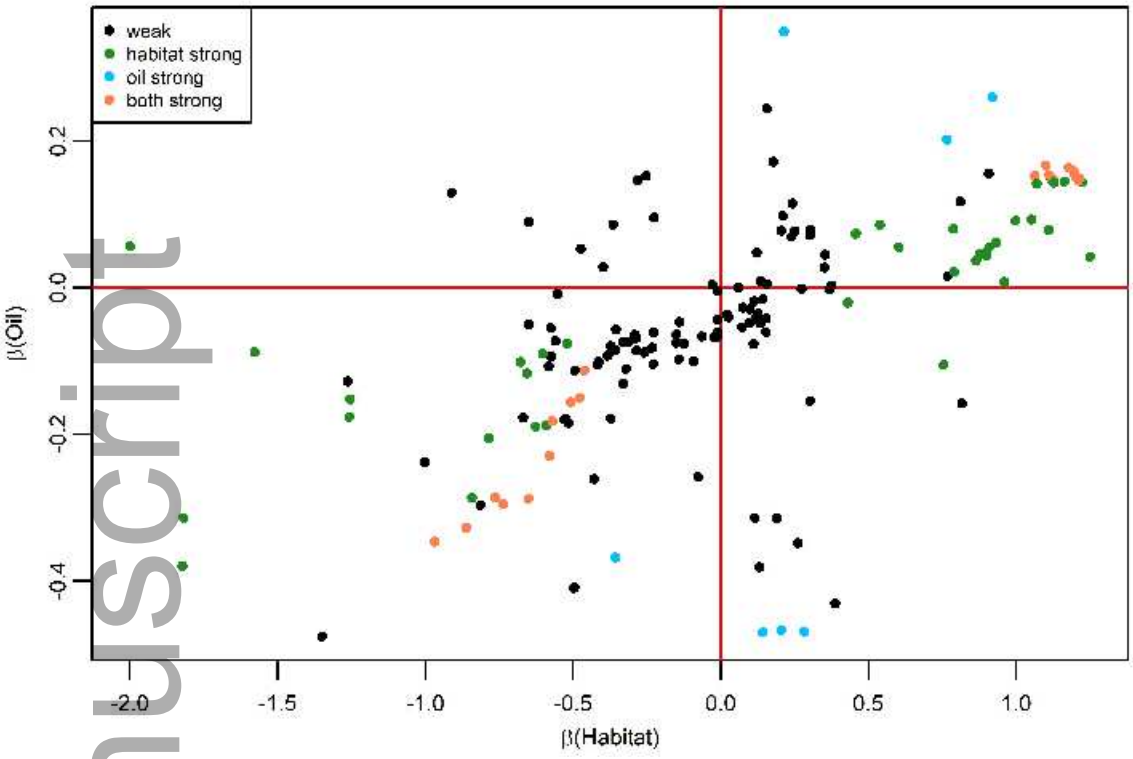
eap_2249_f2.jpg



eap_2249_f3.jpg



eap_2249_f4.jpg



eap_2249_f5.jpg