Modeling Joint Abundance of Multiple Species Using Dirichlet Process Random Effects

Devin S. Johnson[1] and Elizabeth H. Sinclair

Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA,
Seattle, Washington, U.S.A.

May 30, 2016

[1] Email: devin.johnson@noaa.gov

Abstract

We present a method for modeling multiple species distributions simultaneously using Dirichlet Process random effects to cluster species into guilds. Guilds are ecological groups of species that behave or react similarly to some environmental conditions. By modeling latent guild structure, we capture the cross-correlations in abundance or occurrence of species over surveys. In addition, ecological information about the community structure is obtained as a byproduct of the model. By clustering species into similar functional groups, prediction uncertainty of community structure at additional sites is reduced over treating each species separately. The method is illustrated with a small simulation demonstration, as well as an analysis of a mesopelagic fish survey from the eastern Bering Sea near Alaska. The simulation data analysis shows that guild membership can be extracted as the differences between groups become larger and if guild differences are small the model naturally collapses all the species into a small number of guilds which increases predictive efficiency by reducing the number of parameters to that which is supported by the data.

Key words: Abundance, Dirichlet Process, Joint species distribution model, Multivariate, occurence

1 Introduction

In recent years there has been considerable development of methodology for modeling and predicting abundance and occurrence of species of interest. Much of this development uses a hierarchical framework for developing models to fit the complexities of the observed data or natural abundance processes (Cressie et al., 2009; Royle and Dorazio, 2008; Hobbs and Hooten, 2015). Some of these complexities may include: spatial and temporal dependence (Carroll et al., 2010; Latimer et al., 2009; Johnson et al., 2013b; Thorson et al., 2015; Ward et al., 2010), nondetection of individuals at sampled sites (Dorazio and Connor, 2014; Royle, 2004), and zero-inflation (Johnson and Fritz, 2014). Many of these species distribution models (SDMs) were used to make inference to a single species or one-at-a-time modeling if community inference was desired. However, by not recognizing the fact that species interact, use of single species models for making inference for community abundance and structure can produce misleading results (Clark et al., 2014). Hence, new joint species distribution models (JSDMs), which explicitly model species interactions (or, cross-correlation) have recently been developed (e.g., Dorazio and Connor, 2014; Latimer et al., 2009; Thorson et al., 2015). Herein, we propose a novel JSDM approach which models species interactions through membership in a latent ecological guild (Simberloff and Dayan, 1991) or functional group within the sampled range of habitats.

Typically, description of an abundance model begins with a GLM structure for the abundance process using a discrete value distribution such as Poisson or negative-binomial. For example one might model the abundance as a Poisson observation with log-mean being a function of covariates that might include habitat variables or variables related to the sampling procedure which are thought to be related to the observed abundance. Alternatively, one might log transform the abundance and use Gaussian linear models (Johnson et al., 2013b; Johnson and Fritz, 2014; Ward et al., 2010), but the general mean structure is usually the same. Herein, we will focus on the GLM versions. The focus of the abundance modeling is related to either establishing an ecological relationship between (joint) abundance and the environmental covariates or predicting abundance at unsampled locations.

To extend the single species GLM oriented model to account for interactions of multiple species and improve prediction and inference of community structure and joint abundance, there have been several approaches which differ in the details of interaction modeling, but all fit the GLM framework by adding random effects which are either directly correlated between species (Clark et al., 2014; Dorazio and Connor, 2014; Latimer et al., 2009) or when marginalized from the (log-linear) model imply a cross-species correlation structure (Thorson et al., 2015). The direct approach of using a free parameter for every pair of species when modeling the species-level correlation has been successfully implemented (Clark et al., 2014; Latimer et al., 2009), however, in those studies there were a high number of sights sampled or a low number of species considered. In other studies, unstructured covariance did not produce reliable results (Dorazio and Connor, 2014). Thus, recent efforts to contribute novel methodology for JSDMs have focused on reducing the number of parameters used to model species interactions. Dorazio and Connor (2014) used a known species trait proximity matrix to model the species-level covariance matrix using a spatial correlation function. By using the known information on species similarity there are only two parameters necessary to model the cross-correlation. Another low complexity approach has been proposed by Thorson et al. (2015) using linear combinations of latent random effects. Specifically, the latent effects are spatial fields, but the same methodology could be applied using independent random effects. If the number of random effects is small relative to the number of species modeled, the number of parameters necessary for modeling species cross-correlation can be significantly reduced from the unstructured scenario.

As a novel alternative, we propose an JSDM that uses latent ecological guilds to model interactions among species and obtain joint abundance inference. Herein, we also consider joint species occurrence as well, where occurrence is defined as the binary presence (i.e., abundance > 0) or absence (abundance $= 0$) of a species. Dorazio and Connor (2014) used known guild membership of different species to model independence of some species in a cross-correlated JSDM. Simberloff and Dayan (1991) defines an ecological guild to be "a group of species that exploit the same class of environmental resources in a similar way." With this definition in mind, we seek to build a model where species are cross-correlated in abundance because there are unknown group effects for some set of covariates, i.e., if the group (guild) structure was known they could be represented by (group × covariate) interaction terms in the abundance GLM models. To accomplish this task we format the model as a latent class or mixture model (see McLachlan and Peel, 2004). Mixture models or latent class models are often used to model dependance between variables in a nonparametric fashion because for a sufficiently large number of groups, marginalizing over the random latent classes can approximate any dependence structure to whatever degree desired (McLachlan and Peel, 2004; Vermunt et al., 2008). It has been shown that this holds even when the conditional models are independent given group membership (Dunson and Xing, 2009). In an ecological abundance context, finite mixture models have been used in the past to model spatial heterogeneity and correlation in a nonparametric fashion (Dorazio et al., 2008; Johnson et al., 2013b). In this paper we take inspiration from nonparametric dependence methods used for spatial association and apply it to species interaction in abundance modeling.

In the following section we describe the general infinite mixture framework using latent classes and describe the Dirichlet Process (DP) for modeling class membership and the number of classes. There are several choices of models for number and assignment of latent classes, but we utilize the DP due to its long history and good clustering properties (Casella et al., 2014). Parameter estimation in the DP-JSDM is challenging due to the latent class process. We provide a reversible-jump MCMC (RJMCMC; Green 2003) algorithm for making Bayesian inference. Finally, we apply the method to few simulated data sets, as well as, a real data set on mesopelagic fish communities in the eastern Bering Sea, Alaska.

2 Methods

2.1 General model framework

We begin the description of the proposed methods with some notation. First we assume there are J surveys, for which abundance (or count index; hereafter we use the term "counts") of I different species is measured. Let n_{ij} be the observed count for ith species in survey j. We also use the vector notation $\mathbf{n}_i = (n_{i1}, \dots, n_{iJ})'$ and $\mathbf{n} = (\mathbf{n}_1', \dots, \mathbf{n}_I')'$ for the n_{ij}, as well as, other quantities described later. For occurrence modeling we denote occurrence as $y_{ij} = 1$ if $n_{ij} > 0$ otherwise $y_{ij} = 0$. In practice, n_{ij} need not necessarily be observed for occurrence modeling. The notation $\mathbf{y}_i$ and $\mathbf{y}$ are similar to the abundance counterparts.

For abundance modeling, there are several possible distributions that could be used to model the observed discrete counts, Poisson, negative binomial, zero-inflated Poisson, etc., so we will generically denote this observation model as $[n_{ij}|z_{ij}, \boldsymbol{\gamma}]$ where z_{ij} is a latent Gaussian variable controlling the level of expected abundance and $\boldsymbol{\gamma}$ is a set of, possibly nuisance, parameters. The notation "$[A|B]$" refers to the conditional distribution of A given B. For example, if a Poisson distribution is used

$$[n_{ij}|z_{ij}, \boldsymbol{\gamma}] = \text{Poisson}(n_{ij}|e^{z_{ij}}), \tag{1}$$

and $\boldsymbol{\gamma}$ is not necessary. In the example analysis of mesopelagic fish surveys we utilize a zero-inflated Poisson (ZIP) model, so,

$$[n_{ij}|z_{ij}, \boldsymbol{\gamma}] = \gamma_{ij} 1_{[n_{ij}=0]} + (1 - \gamma_{ij})\text{Poisson}(n_{ij}|e^{z_{ij}}), \tag{2}$$

the additional γ_{ij} parameter is the mixing probability for the extra zeros. For occurrence modeling we use

$$[y_{ij}|z_{ij}] = \text{Bernoulli}(\Phi^{-1}\{z_{ij}\}), \tag{3}$$

where $\Phi(\cdot)$ is the standard normal CDF, that is, a probit link function.

To account for unknown interspecies correlations we take a clustering approach inspired by the analysis of Johnson et al. (2013b) for incorporating spatial structure when there are no reasonable distance metrics or neighborhood groupings are unknown. The model is constructed by envisioning an unknown partition, p, of the species into κ_p groups such that species within groups (clusters) behave similarly with respect to the abundance process. For a given p, we model (in vector form) the latent $\mathbf{z}$ process with the linear model

$$[\mathbf{z}|p, \boldsymbol{\delta}_p, \boldsymbol{\beta}, \sigma] = N(\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p, \boldsymbol{\Sigma}), \tag{4}$$

where

- $\mathbf{X}$ is a design matrix of covariates for which there are no group-level effects,
- $\boldsymbol{\beta}$ is a vector of regression coefficients,
- $\mathbf{K}_p = \mathbf{C}_p \otimes \mathbf{H}$, where $\mathbf{C}_p$ is an $I \times \kappa_p$ binary matrix indicating which species belong to each group in p and $\mathbf{H}$ is a $J \times q$ matrix of q habitat covariates recorded at the jth survey,
- $\boldsymbol{\delta}_p = (\boldsymbol{\delta}'_1, \ldots, \boldsymbol{\delta}'_{\kappa_p})'$ is a vector of normally distributed random effects, where, $[\boldsymbol{\delta}_k|\boldsymbol{\Omega}] = \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, for $k = 1, \ldots, \kappa_p$.
- $\boldsymbol{\Sigma}$ is a diagonal matrix with entries σ^2_{ij} (for occurance modeling $\sigma_{ij} = 1$).

To reduce the parameter complexity of the proposed model we suggest the following for general practice:

(i) for abundance models, set $\boldsymbol{\sigma} = \text{diag}(\boldsymbol{\Sigma}^{1/2}) = \exp\{\mathbf{L}\boldsymbol{\theta}\}$, where $\mathbf{L}$ is a matrix of design covariates and

(ii) set $\boldsymbol{\Omega} = \omega^2(\mathbf{H}'\mathbf{H})^{-1}$, where $\omega = \exp(\xi)$.

With respect to (i), there are some useful special cases, namely, $\mathbf{L} = \mathbf{1}$ gives $\sigma_{ij} = \sigma$ and $\mathbf{L} = \mathbf{I}_I \otimes \mathbf{1}_J$ gives $\sigma_{ij} = \sigma_i$. However, the overdispersion parameters could also be modeled based on covariates associated with sampling methods, etc. Suggestion (ii) was formulated from the covariances of the g-prior (Tiao and Zellner, 1964). The g-prior, $N(\mathbf{0}, \omega^2(\mathbf{H}'\mathbf{H})^{-1})$, is an often used prior for regression coefficient parameters. It has the nice benefit that, with a single parameter, it automatically controls the scale of variance and covariance for each coefficient based on the scale of the covariates and their cross-correlation. The exponential reparameterization is used for ease of inference, that is ξ can be unconstrained.

The previous description assumed that the correct partitioning of the species is known, however, for most real data sets, the correct partition is unknown. Thus, we must also provide a probability model over partitions, $[p|\alpha]$, such that marginalization over the unknown partitions creates random coefficient vectors that are nonparametric in their distribution. A commonly used distribution over partitions is the Chinese Restaurant Process (CRP) a finite number of individuals to an unknown number of groups is described as follows, for a given parameter $\alpha > 0$,

1. A customer enters the restaurant and sits at one of an infinite number of tables.

2. The next customer enters and chooses to sit at the occupied table with probability $1/(1 + \alpha)$ or a new table with probability $\alpha/(1 + \alpha)$.

3. In general,the $i + 1$ customer sits at an occupied table with probability proportional to the number of customers already seated or chooses an unoccupied table with probability proportional to α.

Under the CRP model individuals are exchangeable, i.e., individuals join clusters based only on how many other individuals are in the cluster, not who else is in the cluster. This fact forms the basis for Bayesian inference for the CRP model via MCMC (Neal, 2000). The density function for the CRP cluster model is given by,

$$[p|\alpha] = \mathcal{CRP}(\alpha) \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha+I)}\alpha^{\kappa_p}\prod_{k=1}^{\kappa_p}(g_{pk}-1)!, \tag{5}$$

where g_{pk} is the size of the kth cluster (group) in p. Note, that the distribution of p is only a function of the number and sizes of the groups. Realizations of p with the same number of groups and groups sizes have the same probability regardless of which individuals fall in which cluster.

The Dirichlet process is connected to the CRP process because a DP process is constructed using the same procedure to seat the guests in the CRP model. Specifically, in terms of (4), let $\bar{\boldsymbol{\delta}}_i$ be the coefficient associated with the ith species, that is $\bar{\boldsymbol{\delta}}_i = \sum_{k=1}^{\kappa_p} C_{ik}\boldsymbol{\delta}_k$, where C_{ik} is the (i,k) entry of the $\mathbf{C}_p$ matrix. Now, if $\bar{\boldsymbol{\delta}}_i$ follows a DP then, conditionally,

$$[\bar{\boldsymbol{\delta}}_i|\bar{\boldsymbol{\delta}}_1,\ldots,\bar{\boldsymbol{\delta}}_{i-1},\alpha,\boldsymbol{\Omega}] = \frac{\alpha}{\alpha+i-1}\mathcal{N}(\mathbf{0},\boldsymbol{\Omega}) + \sum_{k=1}^{u_i}\frac{n_k}{\alpha+i-1}\boldsymbol{\delta}_k, \tag{6}$$

where u_i is the number of unique values, $\boldsymbol{\delta}_k$, of $\bar{\boldsymbol{\delta}}_{i'}$ $i'=1,\ldots,i-1$, and n_k is the number of species 1 through $i-1$ belonging to group k. In other words, a new table is represented by a new value of $\boldsymbol{\delta}_k$. Thus, the CRP partitioning combined with the $\boldsymbol{\delta}$ realizations for each group implies that $[\bar{\boldsymbol{\delta}}_i|\alpha,\Omega] = \mathcal{DP}(\alpha,\Omega)$.

Like the spatial covariance model use by Dorazio and Connor (2014), the DP-JSDM also marginally possesses generally positive cross-covariance structure. This makes intuitive sense as one is clustering similar species together or, if species are dissimilar, allowing them to be independent. The covariance structure of the DP-JSDM can be derived by forming an intercept random effect, $\boldsymbol{\eta} = \mathbf{K}_p\boldsymbol{\delta}_p$, such that $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\eta}+\boldsymbol{\epsilon}$, where $[\boldsymbol{\epsilon}] = N(\mathbf{0},\boldsymbol{\Sigma})$. Then, conditioning on the cluster assignment, the covariance matrix of the random effect $\boldsymbol{\eta}$ is,

$$\text{Var}(\boldsymbol{\eta}|p) = \mathbf{C}_p\mathbf{C}_p' \otimes \mathbf{H}\boldsymbol{\Omega}\mathbf{H}', \tag{7}$$

and the marginal variance is given by the mixture,

$$\mathrm{Var}(\boldsymbol{\eta}) = \left\{\sum_p \mathbf{C}_p\mathbf{C}_p'[p|\alpha]\right\} \otimes \mathbf{H}\boldsymbol{\Omega}\mathbf{H}' = \boldsymbol{\Psi} \otimes \mathbf{H}\boldsymbol{\Omega}\mathbf{H}', \tag{8}$$

where $\boldsymbol{\Psi}$ is a matrix with (i, i') entries equal to the probabilities that species i shares a guild with species i'. We term the $\boldsymbol{\Psi}$ matrix to be the species proximity matrix due to the fact that is forms a distance, of sorts, in the guild space of the species. Although, the covariance is never negative between any two species, it can be zero, thus those species that occupy different guilds will have uncorrelated η random effects, i.e., if $\psi_{ii'} \approx 0$, then $\mathrm{Cov}(\eta_{ij}, \eta_{i'j})$ ≈ 0.

It should be noted, however, that although the covariance of the $\boldsymbol{\eta}$ random effect is generally, positive, that does not mean that there are only 'positive' (or zero) relationships between species. The clustering is based on the relationship each species has with the chosen covariates. For example, one species may react positively along a covariate gradient ($\delta_i > 0$) and another reacts negatively along that same gradient ($\delta_i < 0$), therefore if a new site has a high level of this covariate, the first species will be predicted to be relatively abundant, while the other species prediction will be lower.

2.2 Bayesian inference

Because of the hierarchical and variable dimensional nature of the parameter space of the DP-JSDM model we employ a Bayesian approach via MCMC (Markov Chain Monte Carlo) for model fitting and inference. The posterior distribution of interest is given by

$$\begin{aligned}[\mathbf{z}, p, \boldsymbol{\delta}_p, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}|\mathbf{n}] \propto [\mathbf{n}|\mathbf{z}]\ [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\delta}_p, \boldsymbol{\sigma}] \\ \times\ [\boldsymbol{\delta}_p|\omega, p]\ [p|\alpha]\ [\omega]\ [\boldsymbol{\sigma}]\ [\boldsymbol{\beta}]\ [\alpha],\end{aligned} \tag{9}$$

where $[\omega]$, $[\boldsymbol{\sigma}]$, $[\boldsymbol{\beta}]$, and $[\alpha]$ are the prior distributions for the parameters.

There are several derived parameters which may be of interest for making desired ecological inference. First, are predictions of community abundance rates at new locations or times. Second, one may be interested in the overall effect of the environmental covariates for a particular species represented by $\bar{\boldsymbol{\delta}}_i$. Finally, the matrix $\mathbf{C}_p\mathbf{C}_p'$ is an $I \times I$ indicator that a species is in the same guild (associated with) another species. The posterior mean of $\mathbf{C}_p\mathbf{C}_p'$ provides estimated guild proximity matrix, $\boldsymbol{\Psi}$. Finally, the number of guilds, κ_p (number of columns in $\mathbf{C}_p$) may be of interest.

The most direct way to make inferences on the proposed hierarchical clustering model is through a reversible-jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 2003) to sample the posterior distribution of the parameters and clustering assignment. Here, we provide an overview of the RJMCMC, additional details of the sampler are given in Appendix A.

In our description, we will assume the following prior distributions for the parameters:

$$[\boldsymbol{\beta}] = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\ [\boldsymbol{\delta}_p|\omega, p] = \mathcal{N}(\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes \omega^2\mathbf{Q}),$$

$$[\omega] = \mathcal{HT}(\phi_\omega, d_\omega),\ [\sigma] = \mathcal{HT}(\phi_\sigma, d_\sigma)$$

$$[p|\alpha] = \mathcal{CRP}(\alpha), \text{ and } [\alpha] = \mathcal{G}(a, b),$$

where $\mathbf{I}_{\kappa_p}$ is an identity matrix of size κ_p, $\mathbf{Q}$ is a known positive-definite matrix, $\mathcal{HT}(\phi, d)$ represents a scaled half-t distribution with scale parameter ϕ and d degrees of freedom, and $\mathcal{G}$ represents a gamma distribution with parameters a and b. For most of these parameters, the priors can be adjusted to whatever distribution the user would like, the trade-off being a Metropolis-Hastings (MH) update instead of a Gibbs step (e.g., for $\boldsymbol{\beta}$) or no difference at all if the parameter has to be updated with an MH step to begin with (ω, σ, and α). However, the normal $[\boldsymbol{\delta}_p|\omega, p]$ prior is necessary to the proposed RJMCMC algorithm. Although, the known $\mathbf{Q}$ is not necessary. This is not as critical as it sounds as the marginal distribution is still a nonparametric DP process we just require that the base distribution be a multivariate normal.

The majority of the proposed RJMCMC algorithm is a standard Metropolis-within-Gibbs (hybrid) sampler for a GLM-like model (e.g., zero-inflated models might also be considered for the abundance distributions). Conditioned on a realization of p, all the other parameters can be updated with a traditional MH step or a Gibbs step. Hence, we do not focus on their updates here (see Appendix A). However, to update p, the dimension of the $\boldsymbol{\delta}_p$ vector will potentially change, necessitating the trans-dimensional aspect of the RJMCMC. Naively, the trans-dimensional moves require a joint $(p, \boldsymbol{\delta}_p)$ proposal which can be rejected often if one of those quantities is a bad fit for the current state of the remaining parameters even though the other is acceptable. Second, proposing new p such that the MCMC chain will mix well over the space of partitions is itself challenging. Because we are assuming that $[\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\delta}_p, \boldsymbol{\sigma}]$ and $[\boldsymbol{\delta}_p|\omega, p]$ are multivariate normal, the first problem can be handled with the partial-analytic RJMCMC method proposed by Godsill (2001) and utilized by Johnson and Hoeting (2011) and Johnson et al. (2013b) in similar trans-dimensional MCMC applications. The partial-analytic method allows proposing new model (p in this case) without jointly proposing the associated model specific parameters ($\boldsymbol{\delta}_p$) because they can be analytically marginalized. This is a special case of a collapsed Gibbs sampler (Van Dyk and Park, 2008).

To produce efficient moves through cluster (guild) space we use the the "individual links" definition of the the CRP process proposed by Blei and Frazier (2011) and subsequently used by Johnson et al. (2013b) for clustering spatial abundance trends. The links version of the CRP process is constructed as follows:

1. A customer enters the restaurant and sits at one of an infinite number of tables.

2. The next customer enters and chooses to sit with the first customer with probability $1/(1+\alpha)$ or a new table with probability $\alpha/(1+\alpha)$.

3. In general, upon entering the restaurant, the $i+1$ customer sits with a previous *customer* (not a table) with probability proportional to 1 or the new customer sit by himself (self-links) with probability proportional to α.

4. Groups are constructed by collecting all cliques of the mathematical graph formed by the links between customers.

Blei and Frazier (2011) show that this definition of the CRP process is equivalent to the traditional definition presented previously. However, MCMC sampling is now based on sampling independent links between individuals. In terms of the multiple species model, let $\ell_i \in \{1, \ldots, i\}$ be the link for the ith species. The full conditional distribution of ℓ_i is,

$$[\ell_i|\cdot] \propto [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\delta}_p, \boldsymbol{\sigma}]\ [\boldsymbol{\delta}_p|\omega, p]\ [\ell_i|\alpha], \tag{10}$$

where p is the partition constructed from all ℓ_i and ,

$$[\ell_i|\alpha] = \frac{\alpha 1_{\{\ell_i = i\}} + 1_{\{\ell_i < i\}}}{1+\alpha}, \tag{11}$$

and $1_{\{.\}}$ is an indicator function for the condition in the brackets. It would be tempting to sample from this discrete distribution in Gibbs fashion, however, note that it depends on $\boldsymbol{\delta}_p$ which may be of different dimension under a different value of ℓ_i. We can collapse over $\boldsymbol{\delta}_p$ and use the marginal distribution

$$\begin{aligned}[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha] &= \int [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\delta}_p, \boldsymbol{\sigma}]\ [\boldsymbol{\delta}_p|\omega, p]\ [\ell_i|\alpha]\ d\boldsymbol{\delta}_p \\ &= [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, p]\ [\ell_i|\alpha] \\ &\propto \mathcal{N}(\mathbf{z}|\mathbf{X}\boldsymbol{\beta}, \mathbf{K}_p(\mathbf{I}_{\kappa_p} \otimes \omega^2 \mathbf{Q})\mathbf{K}_p' + \boldsymbol{\Sigma})\ [\ell_i|\alpha],\end{aligned} \tag{12}$$

which does not depend on $\boldsymbol{\delta}_p$. This approach was used by Johnson and Hoeting (2011) and Johnson et al. (2013b), however, we found that for a large number of species and samples, the covariance matrix $\mathbf{K}_p(\mathbf{I}_{\kappa_p} \otimes \omega^2 \mathbf{Q})\mathbf{K}_p' + \boldsymbol{\Sigma}$ may be quite large and the inversion necessary to evaluate the $[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha]$ for each species and potential link would make the chain prohibitively slow in practice. So, we sought an alternative formulation of the marginal distribution that did not require inversion of such a large covariance matrix. Using Laplace's method (see Kass and Raftery 1995, Section 4.1) we can write

$$\begin{aligned}[\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, p] &= \int [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\delta}_p, \boldsymbol{\sigma}]\ [\boldsymbol{\delta}_p|\omega, p]\ d\boldsymbol{\delta}_p \\ &= (2\pi)^{\kappa_p/2}|\widehat{\mathbf{V}}_p|^{-1/2} \cdot \mathcal{N}(\hat{\boldsymbol{\delta}}_p|\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes \omega^2\mathbf{Q}) \cdot \mathcal{N}(\mathbf{z}|\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\hat{\boldsymbol{\delta}}_p, \boldsymbol{\Sigma}),\end{aligned} \tag{13}$$

where $\widehat{\mathbf{V}}_p = \mathbf{K}_p'\boldsymbol{\Sigma}\mathbf{K}_p + (\mathbf{I}_{\kappa_p} \otimes \omega^{-2}\mathbf{Q}^{-1})$ and $\hat{\boldsymbol{\delta}}_p = \mathbf{V}_p^{-1}(\mathbf{K}_p'\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}))$, which are respectively the inverse covariance and mean for the Gaussian full conditional distribution $[\boldsymbol{\delta}_p|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, p]$. This is the same distribution used to update $\boldsymbol{\delta}_p$ with a Gibbs step following an update of p. Normally, Laplace's method produces an approximation to the integral, but in this case the approximation is exact because the log integrand is quadratic in $\boldsymbol{\delta}_p$ (Goutis and Casella, 1999). By writing the integral in this way we need only invert $\boldsymbol{\Sigma}$, which is diagonal, and $\mathbf{Q}$ because $(\mathbf{I}_{\kappa_p} \otimes \omega^2\mathbf{Q})^{-1} = \mathbf{I}_{\kappa_p} \otimes \omega^{-2}\mathbf{Q}^{-1}$. If we use $\mathbf{Q} = (\mathbf{H}'\mathbf{H})^{-1}$ as previously suggested, then the inverse is trivial. Because, $[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha]$ is relatively cheap to evaluate for each $\ell_i = 1, \ldots, i$ we can use a Gibbs step and draw from the discrete distribution $[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha]$ for each $i = 1, \ldots, I$, with $[\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, p]$ evaluated using (13) instead of (12).

3 A Simulation Proof-of-Concept

To examine the ability of the CRP cluster model to make inference to species interaction, as well as, to make joint community abundance predictions, we tested the model and RJMCMC sampler with a small group of simulated data sets. In analyzing the simulated data our objective was to assess whether the DP-JDSM model would, in practice, produce generally correct estimates of the guild structure. Second, would the DP-JSDM exhibit the expected behavior that as ω becomes small, the number of guilds (groups) estimated will go to one as the functional differences between the guilds (with respect to the variables in $\mathbf{H}$) becomes insignificant.

3.1 Simulation and Analysis

Data were simulated for $I = 20$ species, $J = 35$ samples, and $\kappa_p = 5$ groups. Six data sets were simulated corresponding to ω equal to 0.25, 0.5, 0.75, 1, 1.5, and 2. While the true number of groups is always technically equal to five, the practical differences between the groups tends to zero as ω becomes smaller. The group sizes were $g_{pk} = 7, 5, 4, 3$, and 1. Three environmental variables composing the guild design matrix $\mathbf{H}$ were generated from a standard normal distribution. In addition, a single survey effort variable, $\mathbf{x}$ was generated to adjust overall abundance measurement. The global design matrix was set to $\mathbf{X} = [\mathbf{1}, \mathbf{x}, \mathbf{H}_x]$, where $\mathbf{H}_x = [\mathbf{H}'| \dots |\mathbf{H}']'$, that is, $\mathbf{H}$ matrix is concatenated I times over species. Thus, $\boldsymbol{\delta}_p$ denotes guild differences from the overall global effect of the environmental variables, $\mathbf{H}$. In order to maintain identifiability, we imposed the constraint that $\sum_{k=1}^{\kappa_p} \boldsymbol{\delta}_k = \mathbf{0}$. The global coefficient was set to $\boldsymbol{\beta} = (2, 1, 0, -1, 0.5)'$ and each $\boldsymbol{\delta}_k$; $k = 1, \dots, 5$, was drawn from $N(\mathbf{0}, \omega^2 \mathbf{H}'\mathbf{H})$. In these simulations all $\sigma_{ij} = 0$, therefore, $\mathbf{z} \equiv \mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p$. However, a common σ was estimated in each analysis using a Poisson observation model, that is, $[n_{ij}|z_{ij}] = \text{Poisson}(e^{z_{ij}})$.

The prior distributions used were the same as specified in Section 2.2, specifically,

- $[\boldsymbol{\beta}]$: $\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\hat{\mu}_0, 0, 0, 0)'$ and $\hat{\mu}_0$ is the log of the mean observed count and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 100(\mathbf{X}'\mathbf{X})^{-1}$.
- $[\omega]$: $\phi_\omega = 1$ and $d_\omega = 1$ which implies a half-Cauchy prior distribution.
- $[\sigma]$: $\phi_\sigma = 1$ and $d_\sigma \to \infty$ which implies a half-normal prior distribution.
- $[\alpha]$: $a = 0.258$ and $b = 0.038$.

The prior distribution parameters for the gamma distribution $[\alpha]$ were chosen based upon the method of Dorazio (2009) with one alteration. Dorazio (2009) used the method to choose a and b such that the prior distribution over the number of groups was approximately uniform, that is, $[\kappa_p] \approx 1/I$, $\kappa_p = 1, \dots, I$. However, we agree with the philosophy of Casella et al. (2014) that *a priori* we should prefer fewer groups, therefore, using the same optimization approach as Dorazio (2009), we chose a and b such that, approximately, $[\kappa_p] \propto 1/\kappa_p$. So, all else being equal, a smaller number of groups is *a priori* preferred.

For each of the six simulated datasets, we sampled the posterior distribution (9) using the RJMCMC algorithm detailed in Appendix A. Each sample consisted of 50,000 iterations following a burnin of 10,000 iterations. We created the `multAbund`[2] package for the `R` statistical environment (R Development Core Team, 2015) which contains the code to run the RJMCMC algorithm described in Appendix A.

3.2 Simulation results

As expected, when ω became small the DP-JSDM model was not able to distinguish guild differences between the species and essentially estimated one single group (Figure 1 $\omega = 0.25$). As ω increased and guild differences became apparent the model was able to separate the species into their respective guilds reasonably well (Figure 1). In addition, as ω became large the precision with which the number of guilds was estimated increased as well (Figure 2). There may be some bias as a few of the simulation runs produced $\hat{\kappa}_p = 6$ (Figure 2; $\omega = 1$ and 2), however, a full simulation experiment would be necessary to assess that fact.

[2] Available from github at: `https://github.com/dsjohnson/multAbund`. The package can be installed from within an R session using the `devtools` package, but users need to be able to compile source code on their platform as the `multAbund` package uses `C++` code in its routines.

4 Example: Mesopelagic fish abundance

4.1 Data

In our next demonstration of the DP-JSDM we analyze community structure and abundance of fishes that migrate diurnally between three mesopelagic depths in the eastern Bering Sea near Alaska. The tendency for most mesopelagic species to vertically migrate makes them an important trophic link between the deep scattering layer and upper surface waters (Sinclair et al., 2015) yet, fundamental aspects of multi-species distributions and relative abundances have not been previously described.

The field effort identified three primary sample stations over highly productive areas of the eastern Bering Sea pelagic (Figure 3). In the summers of 1999 and 2000 a total of 29 daytime and 16 nighttime trawls were conducted at three depths (250, 500, and 1000 m) during a narrow sampling period. Four of these trawls were not analyzed due to technical difficulties in the field and we discarded them, resulting in $J = 41$ samples. Trawls were run at-depth for 15–90 minutes resulting in collections of over 50,000 individuals representing 55 species of fish and squid. Essentially, each individual trawl sample represents a community as influenced by depth and time of day. Here we will demonstrate the DP-JSDM using $I = 20$ of the relatively most common fish species (as opposed to squids, etc.).

The variables we put in the **H** design matrix reflect the belief that the species segregate into guilds based on diurnal vertical migration characteristics. So, the guild covariates recorded for each trawl are daylight cycle (day or night) and depth category (250m, 500m, or 1000m). Here we used the full interaction model to define the **H** design matrix (i.e., '$\sim$ `cycle*depth`' in `R` language model syntax). Because the duration of the trawl varied from survey to survey, the duration was included in the **X** matrix to model the overall abundance of fish caught in the trawl.

4.2 Model and analysis

Initial attempts at fitting a DP-JSDM proceeded in the same manner as the analysis of the simulation data in the previous section. Namely, we used the same Poisson model for the observed abundance counts. However, after initial fittings it became evident that the trawl data set possessed a significant level of zero-inflation relative to the Poisson distribution. This is likely due to the spatial patchiness of pelagic fish occurrence distributions (Benoit-Bird and Au, 2003). In addition, there may also be detection issues in the survey such that a zero count in the trawl does not necessarily mean absence of the species. However, unlike Dorazio and Connor (2014) we do not have replicated surveys in which to separate detection and absence. Therefore, we utilized a zero-inflated Poisson (ZIP) model in place of a Poisson GLM. The ZIP model used for this analysis is

$$[n_{ij}|z_{ij}, \gamma_i] = \gamma_i 1_{[n_{ij}=0]} + (1-\gamma_i)\text{Poisson}(n_{ij}|e^{z_{ij}}), \tag{14}$$

where $1_{[n_{ij}=0]}$ is an indicator of a zero count and γ_i is a species-specific zero-inflation mixture

$$[\text{logit } \gamma_i] = \mathcal{T}(\phi_\gamma, d_\gamma), \tag{15}$$

with scale parameter $\phi_\gamma = 1.5$ and degrees of freedom $d_\gamma = 6$. This prior results in the translated prior for γ_i that is approximately uniform in (0,1). For the remaining parameters we used the same prior specification as the simulated data analysis of Section 3.1.

To assess if there is any improvement gained by using the DP-JSDM we also fitted the 'independent species' JSDM, that is $\kappa_p = I$, to the data. The JSDM we fitted was did not truly treat each species independently because there are shared terms in the $\mathbf{X}$ design matrix (trawl duration) but it allows us to assess improvement in classifying animals into functional guilds relative to cycle and depth over treating them separately. To ascertain the magnitude of improvement we would have liked to be able to use the 'leave one out' Bayesian predictive information criterion (BPIC) given by

$$\begin{aligned} -2\ \text{BPIC} &= -2\sum_{i,j} E\{\log[n_{ij}|\mathbf{n}_{-(i,j)}, \mathbf{z}_{-(ij)}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}_p, p, \boldsymbol{\sigma}, \omega, \alpha]\} \\ &= -2\sum_{i,j} E\{\log[n_{ij}|\mathbf{n}_{-(i,j)}, \mathbf{z}_{-(ij)}, \boldsymbol{\gamma}]\} \end{aligned} \tag{16}$$

where $\mathbf{n}_{-(i,j)}$ is a vector of all observed data except n_{ij} and $\log[n_{ij}|\mathbf{n}_{-(i,j)}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}_p, p, \boldsymbol{\sigma}, \omega, \alpha]$ is the log posterior predictive density for the (i, j)th observation. However, it would be computationally infeasible to rerun the RJMCMC for every left out (i, j) entry. So, we used the 'Widely Applicable Information Criterion' (WAIC; Watanabe (2013)) as an approximation (Watanabe, 2010; Link and Sauer, 2016) to -2 BPIC, where

$$\begin{aligned} \text{WAIC} = -2\sum_{i,j} & E\{\log[n_{ij}|\mathbf{n}, \mathbf{z}, \boldsymbol{\gamma}]\} \\ & + 2\sum_{i,j} Var\{\log[n_{ij}|\mathbf{n}, \mathbf{z}, \boldsymbol{\gamma}]\} \end{aligned} \tag{17}$$

The WAIC requires only one run of the RJMCMC with the full data set. There are also other selection methods applicable, see Hooten and Hobbs (2015) for others.

The model was fitted using the `R` package `multAbund`. The RJMCMC algorithm was run for 100,000 iterations following a burnin of 10,000 iterations. The package contains code to fit the Poisson abundance data model as well as the ZIP and Bernoulli probit model for occurrence. In addition to the joint analysis of abundance, we also analyzed the trawl survey data as an occurrence data set where $y_{ij} = 1$ if $n_{ij} > 0$, else $y_{ij} = 0$.

4.3 Results

After fitting the ZIP version of the DP-JSDM and the independent species JSDM there was a substantial improvement in WAIC under the DP-JSDM. WAIC for the DP-JSM model was 3052.071 and WAIC = 3078.992 for the independence model. The posterior mode of the number of guilds was $\hat{\kappa}_p = 8$ with 95% of the posterior probability mass falling on $\kappa_p = 8$

or 9 guilds. Figure 4 illustrates the estimated posterior matrix, $\widehat{\boldsymbol{\Psi}} = E[\mathbf{C}_p\mathbf{C}_p']$ which defines the probability that any two species share the same vertical migration guild. Using $1 - \widehat{\boldsymbol{\Psi}}$ as a measure of distance between species, we plotted the species according to the associated dendogram (Figure 5), which gives a better visualization of the groupings. The predicted abundance for each species was calculated as $\hat{\mathbf{n}}^* = E[\mathbf{n}^*|\mathbf{n}]$ where $\mathbf{n}^* = (n_1^*, \ldots, n_I^*)'$ is an observation under the observed environmental conditions (Figure 6). Results for the $\boldsymbol{\gamma}$ parameters are presented in Table B.1 of Appendix B along with estimates of the $\bar{\boldsymbol{\delta}}_i$ values (Figure B.1). Appendix C provides similar figures and results for the DP-JSDM model using binary occurrence data instead of the observed abundance.

The model profiled a wide range in behavior among species from the two dominant fish families in the Bering Sea, Myctophidae and Bathylagidae. All but one of the 8 guilds described by the model (Figures 5 and C.2) include a single species from one or both of these families, implying that they partition the water column based on a characteristic response to physical factors and foraging requirements.

The accuracy and predictive capability of the model was confirmed by the correct cluster assignment of individual species with known relative abundance and depth distribution profiles in the Bering Sea (i.e., bathylagids, *Leuroglossus schmidti* and *Lipolagus ochotensis*). Then by virtue of guild membership, the model described distribution patterns in species for which there is little reported data (i.e., myctophids, *Stenobrachias leucopsarus* and *Diaphus theta*).

For instance, *L. schmidti* and *S. leucopsarus* formed the tightest cluster in both abundance and occurrence dendograms (Figures 5 and C.2). Each is the most abundant species within their respective families in the Bering Sea (Brodeur et al., 1999; Sinclair et al., 1999) and both were highly represented throughout the water column day and night in this study. Guild identity with *L. schmidti* suggests that *S. leucopsarus* shares a similar life history and foraging strategy wherein juveniles and adults have indistinct vertical migration and are stratified in the water column according to age (size) with adults remaining below 240 m (Beamish et al., 1999; Mecklenburg et al., 2002).

The bathylagid *L. ochotensis* and myctophid *D. theta* also form a guild in abundance (Figure 5) along with *Stenobrachias nannochir* in occurrence guilds (Figure C.2). *Lipolagus ochotensis* and *S. nannochir* are among the most abundant mesopelagic species in the Bering Sea Sinclair et al. (1999); Mecklenburg et al. (2002). Both are size-stratified by depth with adults residing in the deepest layers and especially present between 500-1000 m (Mecklenburg et al., 2002). As a strong vertical migrator, *L. ochotensis* is also abundant between 200-500 m Sinclair et al. (1999); Mecklenburg et al. (2002). Little is known about *D. theta* from directed catch in the Bering Sea, however guild identity with *S. nannochir* and especially with *L. ochotensis* suggests they share similar patterns of behavior. The model implication that *D. theta* is an age-stratified strong vertical migrator available at upper mesopelagic depths (Figure 6, B.1, and C.3) is supported by observations that it is a primary prey item of Dall's porpoise (*Phocoenoides dalli*) in the top 250 m of water column (Crawford, 1981).

The best example of the degree of fine detail captured by the model was demonstrated by *Bathylagus pacificus*, a common and abundant species of Bathylagidae that formed its own cluster (Figure 5). Like other members of its family *B. pacificus* demonstrates a bimodal pattern in body size at depth (Peden et al., 1985; Mecklenburg et al., 2002). In our study, juvenile fish were concentrated at mid-layer levels during the day (500 meters) rising to 250 meters at night, while adults concentrate at deeper daytime layers (1000 m) rising to 500 m at night (Sinclair and Stabeno, 2002). This vertical migratory movement is apparent in the log abundance plots (Figure 6; and $\bar{\boldsymbol{\delta}}_i$ values in Figure B.1) that together with known age distribution suggest *B. pacificus* may form its own guild based on abundances at depth driven by varying foraging requirements of juvenile and adults.

5 Discussion

We presented a new methodology for modeling joint species distributions based on Dirichlet process random effects to model species associations through a latent guild structure. Instead of trying to directly parameterize cross-correlation in a species-specific random effect, we used latent membership in an ecological guild. Species belonging to the same guild followed the same response to environmental conditions through random coefficients effects in a GLM-like setting. Unlike simple cross-correlated species random intercepts, the DP-JSDM provides some valuable information on which species belong to guilds together and for the species within a guild, how they respond to the selected environmental conditions together.

A fundamental aspect of mesopelagic ecology is diel vertical migration. The DP-JSDM successfully identified community structure among 20 species of fish from the eastern Bering Sea within this framework. The selected model parameters of depth and light describe real-time clusters of species that move together similarly through the water column on a 24 hour cycle, presumably in relation to foraging. Based on studies conducted in the North Pacific Ocean, the diets of many of these same species collected from different depths match vertical distribution patterns of the variety of copepods and euphausiids that they consume (Beamish et al., 1999).

Although the DP-JSDM model was initially desired to model species association, it has the added benefit that it automatically adjusts to the necessary complexity because the number of guilds is also simultaneously being estimated as well. In the simulation experiment it was demonstrated that if there is apparently little difference between the species in their response to the recorded environmental conditions the DP-JSDM will collapse to one guild, that is, no statistical difference between the species. This reduction in model complexity was noted by Johnson et al. (2013b) in reference to spatially clustering abundance trends.

In our description of the model and our examples, we have provided a relatively straightforward demonstration of the model and associated RJMCMC algorithm. However, there are several extensions that would be useful in other ecological settings. Here we did not have repeated observations at each site, so, we could not add an identifiable detection model to the observation process, although, we illustrated that covariates (i.e., trawl duration) could be added as a quasi-detection model as Ver Hoef and Frost (2003) used. However, if multiple observations are available for each site, then a detection process could be added to the observation model. Dorazio and Connor (2014) made use of an N-mixture model and the DP-JSDM could use that as well. Instead of the ZIP model, one could add a another observation model,

$$[\tilde{n}_{ijk}, n_{ij}|...] = \text{Binomial}(\tilde{n}_{ijk}|n_{ij}, \gamma_{ijk})\text{Poisson}(n_{ij}|z_{ij}), \tag{18}$$

as the observation portion of the model, where $\tilde{n}_{ij}$ is the observed abundance of species i at site j during survey k and γ_{ijk} is the probability of each of the n_{ij} individuals being observed. If one marginalizes over the true abundances, the Poisson observation model results,

$$[\tilde{n}_{ijk}|\gamma_{ijk}, z_{ij}] = \text{Poisson}(\tilde{n}_{ijk}|\log\gamma_{ijk} + z_{ij}), \tag{19}$$

where $E[n_{ijk}] = \exp\{\log\gamma_{ijk} + z_{ij}\}$. The same approach could also be used for occurrence modeling, in which case, it becomes occupancy modeling, that is, for the observed presence $\tilde{y}_{ijk}$, we use the hierarchical observation model,

$$[\tilde{y}_{ijk}, y_{ij}|...] = \text{Bernoulli}(\tilde{y}_{ijk}|y_{ij}\gamma_{ijk})\text{Bernoulli}(y_{ij}|z_{ij}), \tag{20}$$

where the probability that $\tilde{y}_{ijk} = 1$ is $y_{ij}\gamma_{ijk}$. The main point being that the process model does not change in either of these two settings, so, the DP-JSDM can easily be adapted to these situations.

There is also an alteration that can be made when many sites are visited and spatial correlation between sights might also be a consideration. We are not calling this an extension, because spatial correlation can be added without making additions to the basic structure presented. All that needs to be changed to add random spatial effects is to use the basis function approach of Ver Hoef and Jansen (2014), Johnson et al. (2013a), or Hefley et al. (2016). In under a spatial basis function model, the random spatial field is modeled as $\boldsymbol{\eta} = \mathbf{H}\boldsymbol{\delta}$ where the columns of the matrix $\mathbf{H}$ contain the spatial basis functions evaluated at each of the modeled sites (rows). Each basis column represents a different frequency. In the notation just presented it should be fairly obvious how the DP-JSDM can be changed to contain spatial correlation, one simply needs to use a basis function matrix for the environmental design matrix. In that case, it might be appropriate to use $[\boldsymbol{\delta}|\omega] = \mathcal{N}(\mathbf{0}, \omega^2\mathbf{I})$ for the DP baseline distribution to match prior specifications that are usually used in spatial analysis. And, of course, one could combine the spatial model with the previously mentioned detection model extensions to form mutivariate spatial models for occupancy and abundance modeling.

Acknowledgments

The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

References

Albert, J. and Chib, S. (1993). Bayesian-analysis of binary and polychotomous reponse data. *Journal of the American Statistical Association*, 88(422):669–679.

Beamish, R., Leask, K., Ivanov, O., Balanov, A., Orlov, A., and Sinclair, B. (1999). The ecology, distribution, and abundance of midwater fishes of the subarctic pacific gyres. *Progress in Oceanography*, 43(2):399–442.

Benoit-Bird, K. J. and Au, W. W. (2003). Spatial dynamics of a nearshore, micronekton sound-scattering layer. *ICES Journal of Marine Science: Journal du Conseil*, 60(4):899–913.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.

Brodeur, R. D., Wilson, M. T., Walters, G. E., and Melnikov, I. V. (1999). Forage fishes in the bering sea: distribution, species associations, and biomass trends. *Dynamics of the Bering Sea*, pages 509–536.

Carroll, C., Johnson, D. S., Dunk, J. R., and Zielinski, W. J. (2010). Hierarchical bayesian spatial models for multispecies conservation planning and monitoring. *Conservation Biology*, 24(6):1538–1548.

Casella, G., Moreno, E., and Girón, F. J. (2014). Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9:613–658.

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Crawford, T. W. (1981). Vertebrate prey of phocoenoides dalli,(dall's porpoise): associated with the japanese high seas salmon fishery in the north pacific ocean. Master's thesis, University of Washington.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570.

Dorazio, R. M. (2009). On selecting a prior for the precision parameter of dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 139(9):3384–3390.

Dorazio, R. M. and Connor, E. F. (2014). Estimating abundances of interacting species using morphological traits, foraging guilds, and habitat. *PloS one*, 9(4):e94323.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., and Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a dirichlet process prior. *Biometrics*, 64(2):635–644.

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Godsill, S. (2001). On the relationship between Markov Chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10:230–248.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224.

Green, P. J. (2003). Trans-dimensional markov chain monte carlo. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press, Inc., New York.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2016). The basis function approach for modeling autocorrelation in ecological data. *Ecography*, In press.

Hobbs, N. T. and Hooten, M. B. (2015). *Bayesian models: a statistical primer for ecologists.* Princeton University Press.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., and Pond, B. A. (2013a). Spatial occupancy models for large data sets. *Ecology*, 94(4):801–808.

Johnson, D. S. and Fritz, L. (2014). agtrend: A bayesian approach for estimating trends of aggregated abundance. *Methods in Ecology and Evolution*, 5:1110–1115.

Johnson, D. S. and Hoeting, J. A. (2011). Bayesian multimodel inference for geostatistical regression models. *Plos One*, 6(11):e25677.

Johnson, D. S., Ream, R. R., Towell, R. G., Williams, M. T., and Guerrero, J. D. L. (2013b). Bayesian clustering of animal abundance trends for inference and dimension reduction. *Journal of Agricultural Biological and Environmental Statistics*, 18(3):299–313.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Latimer, A., Banerjee, S., Sang Jr, H., Mosher, E., and Silander Jr, J. (2009). Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern united states. *Ecology Letters*, 12(2):144–154.

Link, W. A. and Sauer, J. R. (2016). Bayesian cross-validation for model evaluation and selection, with application to the north american breeding survey. *Ecology*, In press.

McLachlan, G. and Peel, D. (2004). *Finite mixture models.* John Wiley &amp; Sons.

Mecklenburg, C. W., Mecklenburg, T. A., and Thorsteinson, L. K. (2002). *Fishes of Alaska.*

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Peden, A. E., Ostermann, W., and Pozar, L. J. (1985). *Fishes observed at Canadian Weathership Station Papa (500N, 1450W): with notes on the transpacific cruise of the CSS Endeavor*. Number 18. British Columbia Provincial Museum.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical Modeling and Inference in Ecology*. Academic Press- Elsevier Ltd.

Shaby, B. and Wells, M. T. (2011). Exploring an adaptive metropolis algorithm. Technical Report 2011-14, Department of Statistical Science, Duke University.

Simberloff, D. and Dayan, T. (1991). The guild concept and the structure of ecological communities. *Annual review of ecology and systematics*, pages 115–143.

Sinclair, E., Balanov, A., Kubodera, T., Radchenko, V., and Fedorets, Y. A. (1999). Distribution and ecology of mesopelagic fishes and cephalopods. *Dynamics of the Bering Sea (TR Loughlin and K Ohtani, eds.), Alaska Sea Grant College Program AK-SG-99-03, University of Alaska Fairbanks*, pages 485–508.

Sinclair, E. and Stabeno, P. (2002). Mesopelagic nekton and associated physics of the southeastern bering sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(26):6127–6145.

Sinclair, E., Walker, W., and Thomason, J. (2015). Body size regression formulae, proximate composition and energy density of eastern Bering Sea mesopelagic fish and squid. *PloS ONE*, In press.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. (2015). Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*.

Tiao, G. C. and Zellner, A. (1964). Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika*, pages 219–230.

Van Dyk, D. A. and Park, T. (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796.

Ver Hoef, J. M. and Frost, K. J. (2003). A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics*, 10:201–219.

Ver Hoef, J. M. and Jansen, J. K. (2014). Estimating abundance from counts in large data sets of irregularly-spaced plots using spatial basis functions. *arXiv preprint arXiv:1410.3163*.

Vermunt, J. K., Van Ginkel, J. R., Der Ark, V., Andries, L., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1):369–397.

Ward, E. J., Chirakkal, H., Gonzalez-Suarez, M., Aurioles-Gamboa, D., Holmes, E. E., and Gerber, L. (2010). Inferring spatial structure from time-series data: using multivariate state-space models to detect metapopulation structure of california sea lions in the gulf of california, mexico. *Journal of Applied Ecology*, 47(1):47–56.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897.

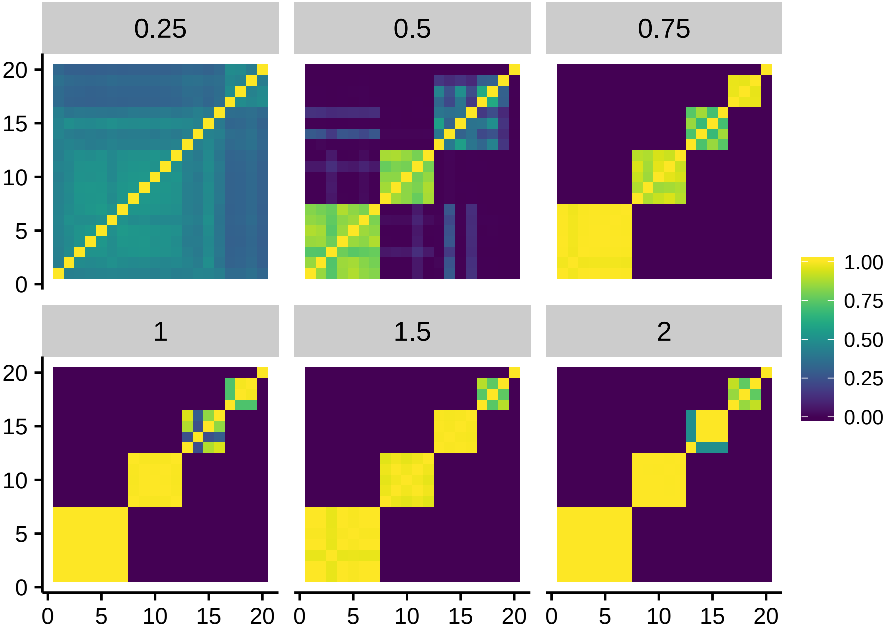

Figure 1: Estimated probabilities of joint guild membership between each species.

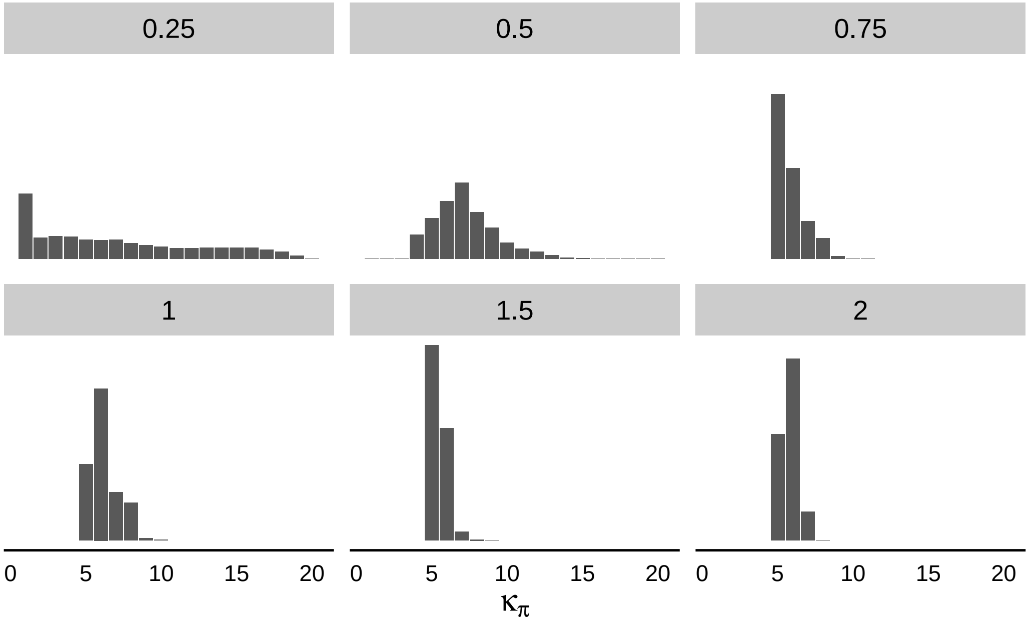

Figure 2: Estimated number of guilds, κ_p, for simulated Poisson data sets with ω ranging from 0.25 to 2.

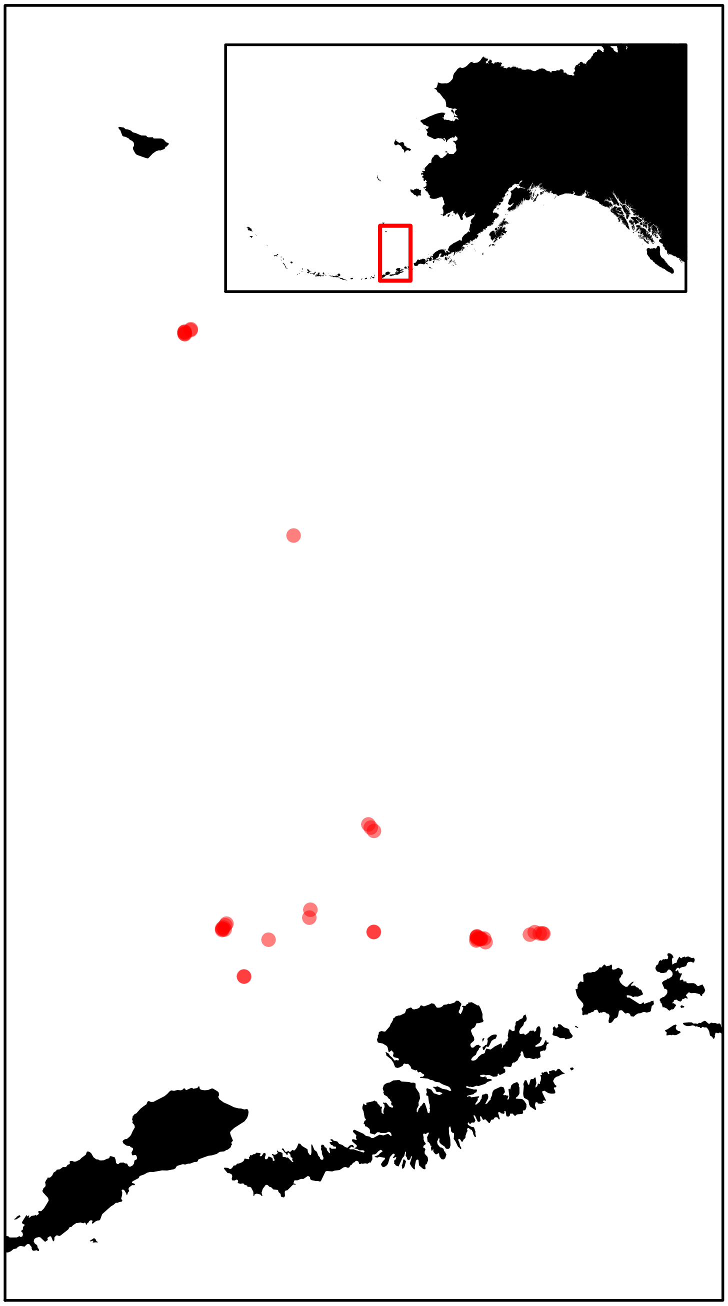

Figure 3: Locations of the mesopelagic trawl surveys. There were $J = 41$ separate trawl surveys used the analysis of Section 4, however, some surveys were attempted geographically near other surveys, so, they are somewhat obscured in the figure.

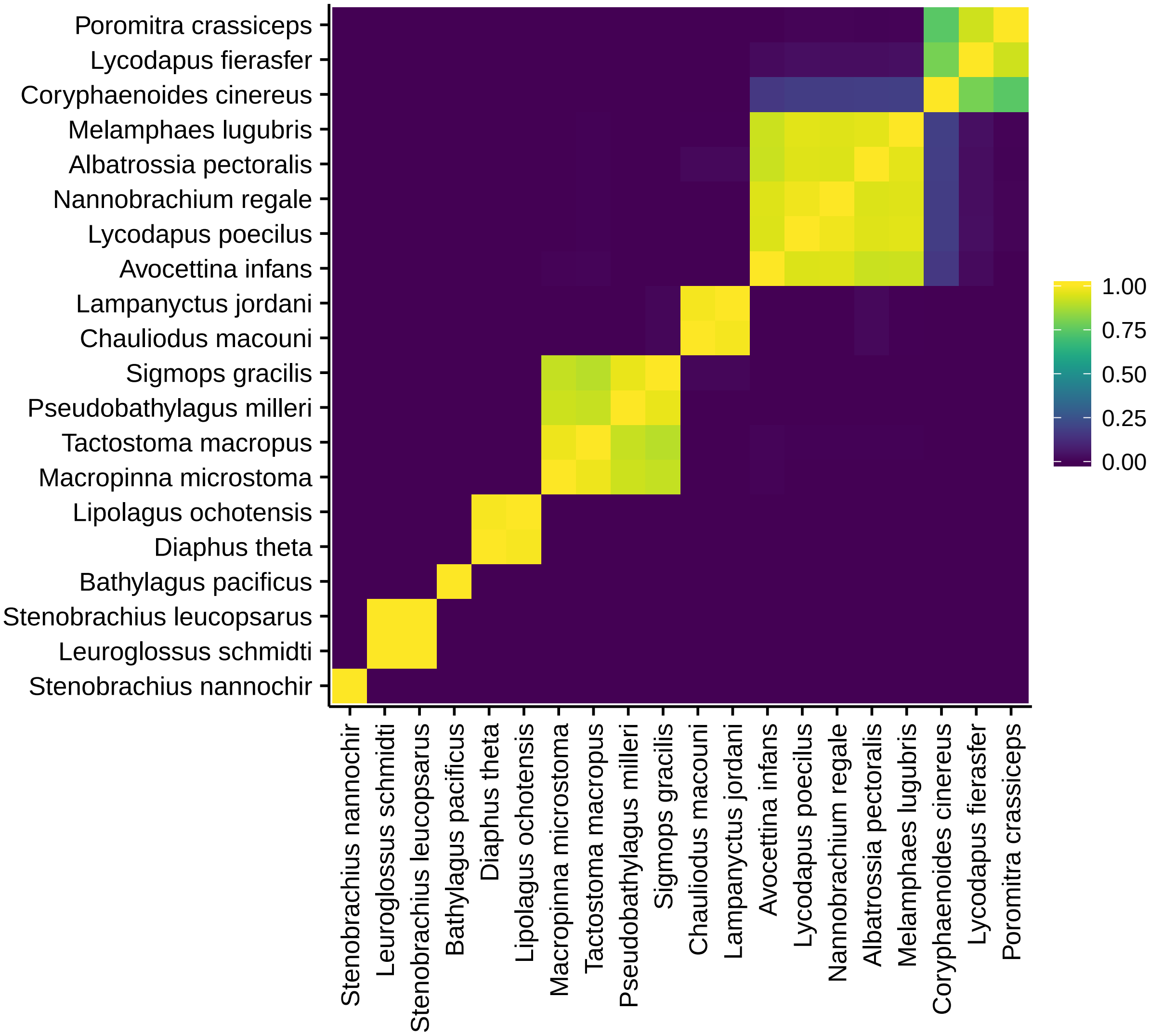

Figure 4: Estimated probability of joint guild membership for each of the fish species in the trawl survey with respect to abundance

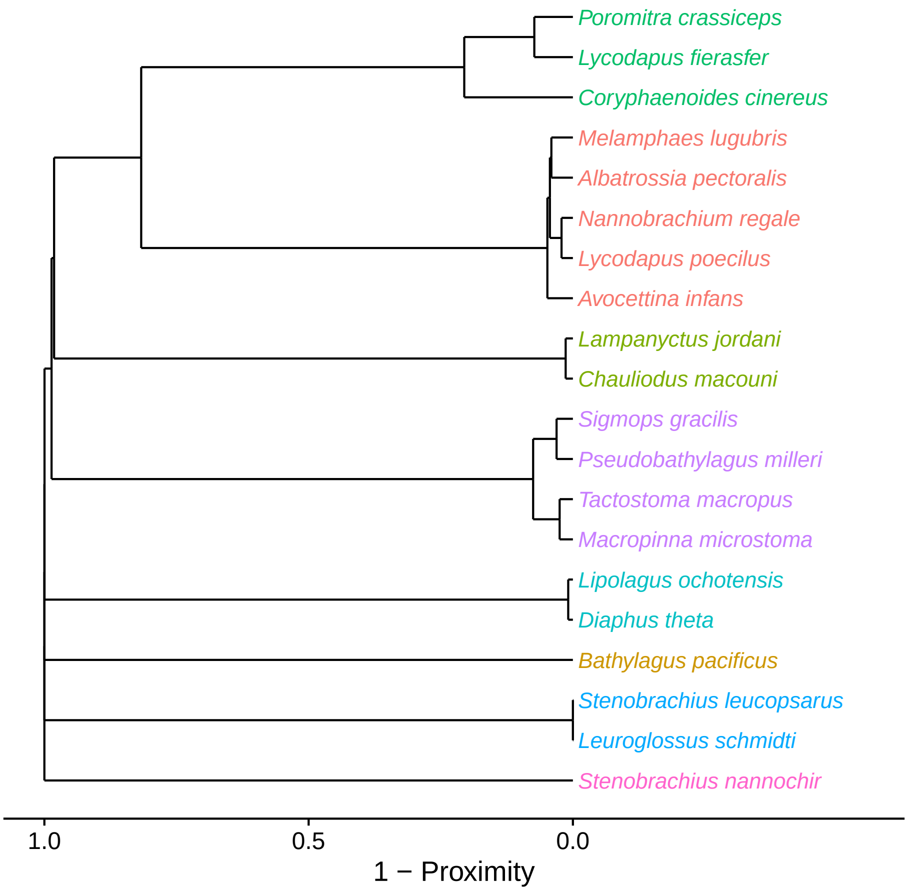

Figure 5: Clustering of trawl survey fish species based on the estimated probability of joint guild membership. The matrix $1 - \widehat{\boldsymbol{\Psi}}$ was used as a distance matrix for forming the dendogram. The colored labels reflect guild groupings based on the posterior mode number of guilds, $\hat{\kappa}_p = 8$

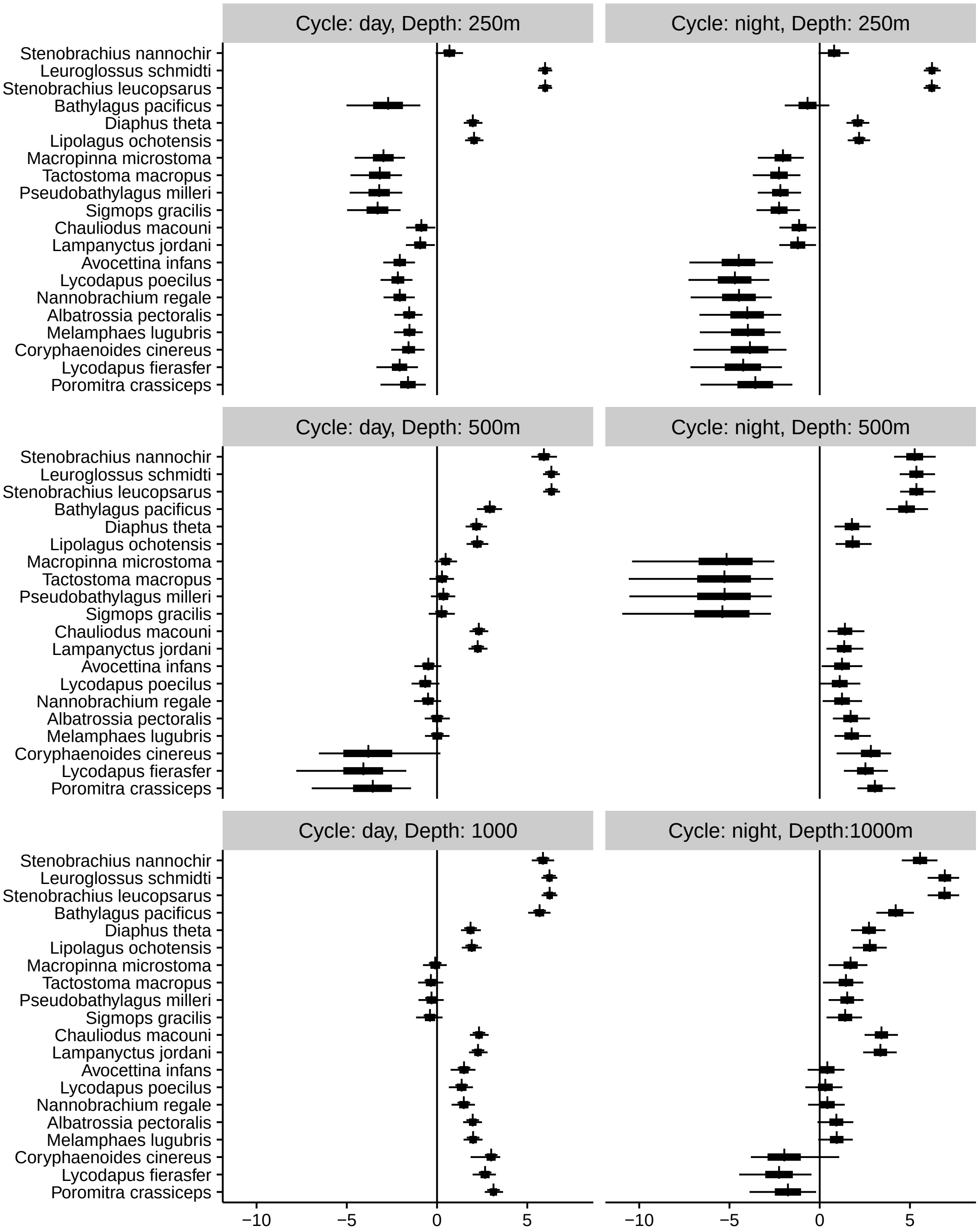

Figure 6: Species-specific predictions of log-abundance for each level of cycle (day or night), and depth (250, 500, or 1000 m).

Appendix A: RJMCMC details

A.1 Prior distributions

Here we describe the details for performing the necessary parameter updates in the RJMCMC algorithm. To facilitate the description the reader should recall we use the following prior distributions in full vector form (where appropriate):

- $[\text{logit } \gamma_i] = \mathcal{T}(\phi_\gamma, d_\gamma)$ for $i = 1, \ldots, I$
- $[\boldsymbol{\beta}] = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$,
- $[\boldsymbol{\delta}_p|\omega] = \mathcal{N}(\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes \omega^2(\mathbf{H}'\mathbf{H})^{-1})$,
- $[\omega] = \mathcal{HT}(\phi_\omega, d_\omega)$
- $[\sigma] = \mathcal{HT}(\phi_\sigma, d_\sigma)$
- $[p|\alpha] = \mathcal{CRP}(\alpha)$
- $[\alpha] = \mathcal{G}(a, b)$,

where $\mathcal{T}$ denotes a t distribution, $\mathcal{N}$ is a (multivariate) normal distribution, $\mathcal{HT}$ is a half-t distribution, $\mathcal{CRP}$ is the Chinese restaurant process, and $\mathcal{G}$ is a gamma distribution. Now, we can describe the Markov Chain Monte Carlo (MCMC) sampler. The sampler is constructed from repeated draws from the full conditional posterior distributions. We use the notation $[x|\cdot]$ to represent the conditional distribution of the variable 'x' given all of the other model components.

A.2 Updating z

We will first describe the updating of **z** for the abundance models. Unfortunately, for the abundance models used in this paper (e.g., $[n_{ij}|z_{ij}, \boldsymbol{\gamma}]$ = ZIP or Poisson), the full conditional distribution does not exist in a nice closed form and we suspect this is the case for every abundance model one may want to use. The full conditional distribution required for the update is,

$$[\mathbf{z}|\cdot] \propto [\mathbf{n}|\mathbf{z},\boldsymbol{\gamma}] \cdot \mathcal{N}(\mathbf{z}|\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p, \boldsymbol{\Sigma}), \tag{A.1}$$

for which a Metropolis-Hastings (MH) step is used with a random walk proposal distribution $[\mathbf{z}^*|\mathbf{z}] = N(\mathbf{z}, \mathbf{R}_z)$, where $\mathbf{R}_z$ is a diagonal matrix that is tuned for optimal sampling. In the `R` package `multAbund` we use the adaptive random walk proposal described by Shaby and Wells (2011) that continually adjusts proposal distribution throughout the MCMC run. Once the new $\mathbf{z}^*$ is drawn, each z^*_{ij} is accepted with probability

$$\max\left\{1, \frac{[z^*_{ij}|\cdot]}{[z_{ij}|\cdot]}\right\}. \tag{A.2}$$

Note, that even though $\mathbf{z}^*$ is proposed as a vector, the independence of each element implies that each z^*_{ij} can be accepted or rejected independently.

If one is analyzing occurrence data with a probit link as described in the main text of the paper, then the full conditional distribution,

$$[\mathbf{z}|\cdot] \propto [\mathbf{y}|\mathbf{z}] \cdot \mathcal{N}(\mathbf{z}|\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p, \boldsymbol{\Sigma}), \tag{A.3}$$

is available in closed form. For each (i, j), the necessary full conditional distribution is

$$[z_{ij}|\cdot] = \mathcal{N}^{b_{ij}}_{a_{ij}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p, \boldsymbol{\Sigma}), \tag{A.4}$$

where $\mathcal{N}^{b_{ij}}_{a_{ij}}$ is a truncated normal distribution with lower bound

$$a_{ij} = \begin{cases} -\infty & \text{for } y_{ij} = 0 \\ 0 & \text{for } y_{ij} = 1 \end{cases} \tag{A.5}$$

and upper bound

$$b_{ij} = \begin{cases} 0 & \text{for } y_{ij} = 0 \\ \infty & \text{for } y_{ij} = 1 \end{cases} \tag{A.6}$$

(Albert and Chib, 1993). If another link function is used, then the same procedure as the abundance model updates is used with a MH acceptance step.

A.3 Udating γ

Here, the only model used where $\boldsymbol{\gamma}$ was present is the ZIP model used in the analysis of the fish survey data. Therefore, we only describe updating of this parameter with respect to the ZIP model with species-specific ZIP parameters, γ_i. The full conditional distribution of logit γ_i is

$$[\text{logit } \gamma_i|\cdot] = [\mathbf{n}_i|\mathbf{z}_i, \gamma_i] \cdot \mathcal{T}(\text{logit } \gamma_i|, \phi_\gamma, d_\gamma). \tag{A.7}$$

As with the $\mathbf{z}$ updates, the adaptive random walk MH update $\mathcal{N}(\{\text{logit } \gamma_i, R_\gamma)$ was used were R_γ is continually adapted through the RJMCMC.

A.4 Updating $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_p$

All of the coefficient vectors in the model have a normal prior distribution, thus the full conditional distributions $[\boldsymbol{\beta}|\cdot]$ and $[\boldsymbol{\delta}_p|\cdot]$ are normal distributions where each is given in Table A.1.

Table A.1: Means and variances for sampling of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_p$. Each parameter has a full conditional distribution of the form $\mathcal{N}(\mathbf{V}^{-1}\mathbf{m}, \mathbf{V}^{-1})$.

Distribution	$\mathbf{V}$	$\mathbf{m}$
$[\boldsymbol{\beta}\|\cdot]$	$\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$	$\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{K}\boldsymbol{\delta}_p) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$
$[\boldsymbol{\delta}_p\|\cdot]$	$\mathbf{K}_p'\boldsymbol{\Sigma}^{-1}\mathbf{K}_p + (\mathbf{I}_{\kappa_p} \otimes \boldsymbol{\Omega})^{-1}$	$\mathbf{K}_p'\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$

A.5 Updating ω and σ

Using an $\mathcal{HT}$ family of priors is not directly conjugate, therefore, a MH step is used here as well. Recall that here we are using $\mathbf{\Omega} = \omega^2(\mathbf{H}'\mathbf{H})^{-1}$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, where $\omega = \exp(\xi)$ and $\sigma = \exp(\theta)$. These choices could be easily modified if desired. For ω, the full conditional distribution is given by

$$[\omega|\cdot] \propto \mathcal{N}(\boldsymbol{\delta}_p|\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes \mathbf{\Omega}) \cdot \mathcal{HT}(\omega|\phi_\omega, d_\omega). \tag{A.8}$$

when converting to the log parameterization, we obtain the full conditional for ξ,

$$[\xi|\cdot] \propto \mathcal{N}(\boldsymbol{\delta}_p|\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes e^{2\xi}(\mathbf{H}'\mathbf{H})^{-1}) \cdot \mathcal{HT}(e^{\xi}|\phi_\omega, d_\omega) \cdot \xi \tag{A.9}$$

As in the z updates, we use a normal random-walk proposal $[\xi^*|\cdot] = \mathcal{N}(\xi, R_\xi)$, where R_ξ is adaptively tuned throughout the MCMC run in the way as the $\mathbf{z}$ updates. With regards to σ, the θ parameter is updated in an identical fashion with the full conditional distribution given by

$$[\theta|\cdot] \propto \mathcal{N}(\mathbf{z}|\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\boldsymbol{\delta}_p, e^{\theta}\mathbf{I}) \cdot \mathcal{HT}(e^{\theta}|\phi_\sigma, d_\sigma) \cdot \theta \tag{A.10}$$

and adaptive random walk proposal distribution $\mathcal{N}(\theta^*|\theta, R_\theta)$.

A.6 Updating p and α

The update of p was described in the main portion of the paper, therefore we omit it here and refer the reader to Section 2.2 for details.

The CRP parameter α is updated through an MH step with the previously described adaptive random walk proposal on $\log\alpha$. The full conditional distribution is given by

$$[\alpha|\cdot] \propto \mathcal{CRP}(p|\alpha) \cdot \mathcal{G}(\alpha|a, b). \tag{A.11}$$

However, as with all of the positive valued parameters, we choose to reparameterize to the log scale to make use of the adaptive random walk proposal distribution. So, the full conditional distribution for $\log \alpha$ is

$$[\log \alpha|\cdot] \propto \mathcal{CRP}(p|\alpha) \cdot \mathcal{G}(\alpha|a, b) \cdot \log \alpha. \tag{A.12}$$

The same adaptive procedure was used with an MH acceptance step to sample the full conditional distribution.

Appendix B: Additional results for fish survey abundance model

Table A.2: Results for species-specific Zero-inflated Poisson (ZIP) mixture parameters, γ_i. The 'Estimate' column is the posterior mode estimate and the 'CI' columns are the upper and lower 0.95 highest probability density interval values. The mixture probabilities represent the probability that a given species is unavailable for surveying in a particular survey.

	Estimate	Lower CI	Upper CI
Albatrossia pectoralis	0.20	0.03	0.44
Avocettina infans	0.52	0.27	0.74
Bathylagus pacificus	0.04	0.00	0.20
Chauliodus macouni	0.03	0.00	0.17
Coryphaenoides cinereus	0.14	0.00	0.46
Diaphus theta	0.18	0.04	0.33
Lampanyctus jordani	0.08	0.00	0.24
Leuroglossus schmidti	0.01	0.00	0.07
Lipolagus ochotensis	0.13	0.02	0.28
Lycodapus fierasfer	0.43	0.20	0.69
Lycodapus poecilus	0.58	0.35	0.79
Macropinna microstoma	0.07	0.00	0.36
Melamphaes lugubris	0.18	0.00	0.40
Nannobrachium regale	0.54	0.28	0.75
Poromitra crassiceps	0.03	0.00	0.28
Pseudobathylagus milleri	0.28	0.00	0.56
Sigmops gracilis	0.37	0.03	0.66
Stenobrachius leucopsarus	0.01	0.00	0.07
Stenobrachius nannochir	0.04	0.00	0.15
Tactostoma macropus	0.32	0.00	0.57

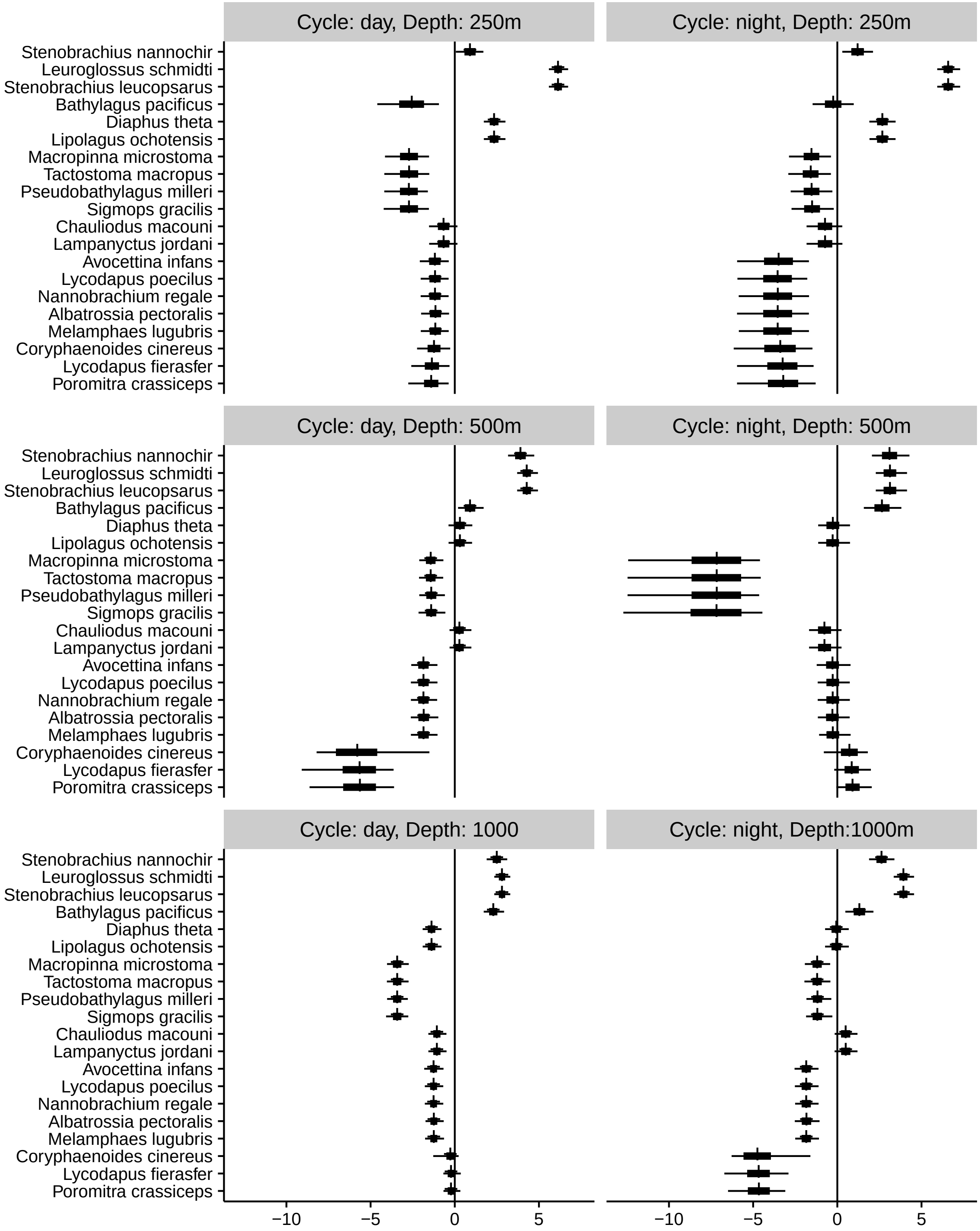

Figure B.1: Species-specific δ estimates, $\bar{\delta}_i$, for each level of cycle (day or night), and depth (250, 500, or 1000 m).

Appendix C: Mesopeleagic fish survey occurrence modeling

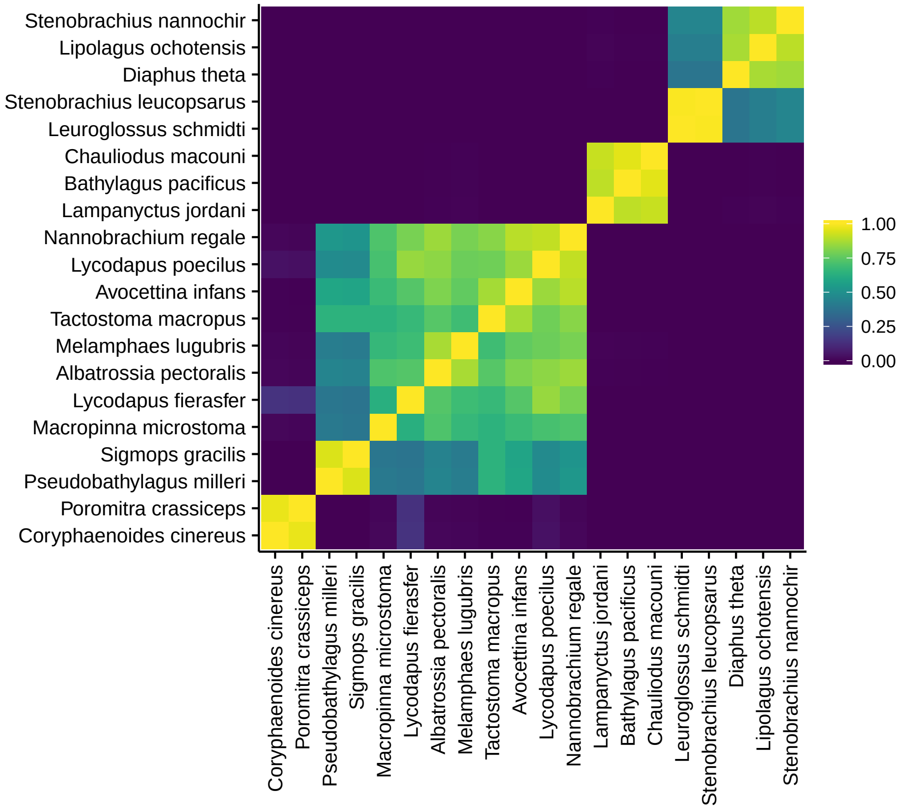

Figure C.1: Estimated probability of joint guild membership for each of the fish species in the trawl survey.

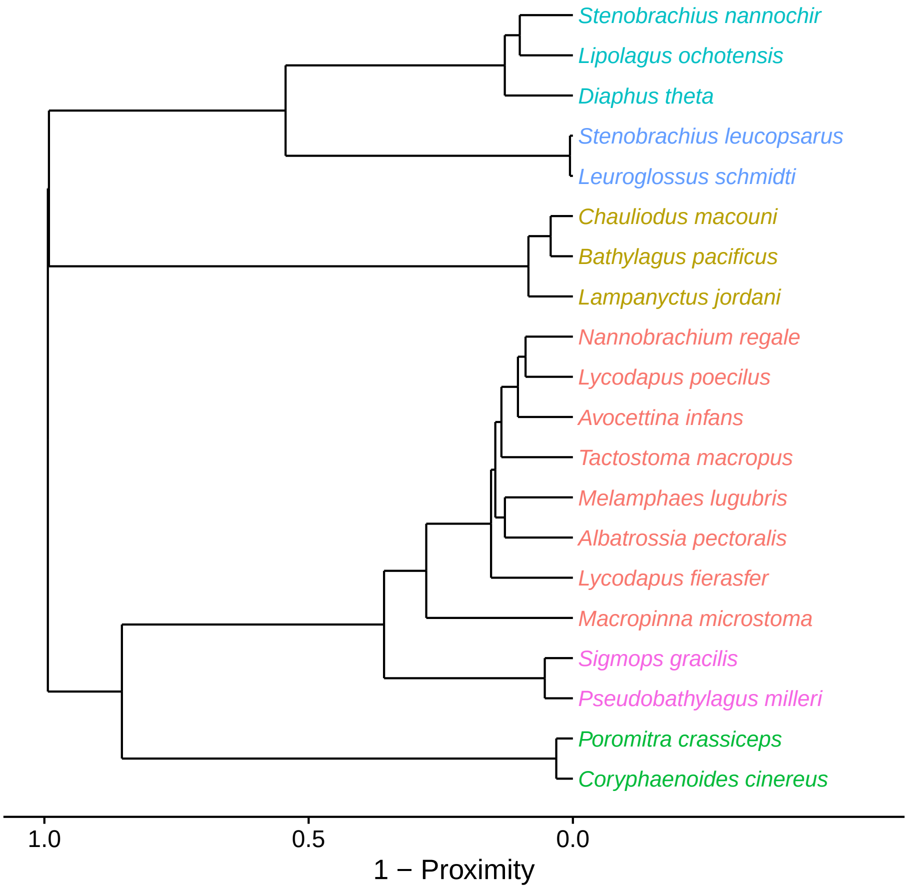

Figure C.2: Clustering of trawl survey fish species based on the estimated probability of joint guild membership.

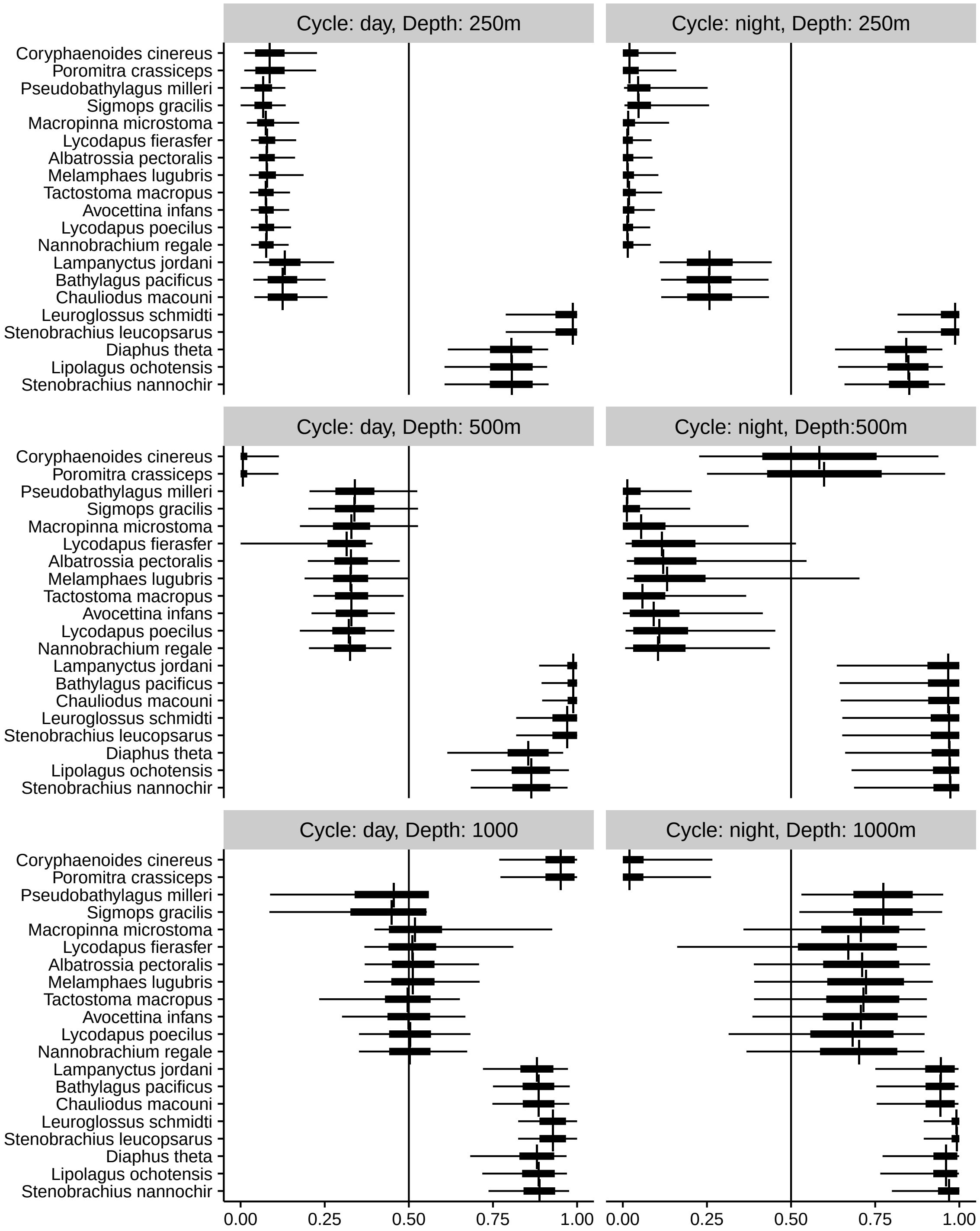

Figure C.3: Species-specific predictions of occurrence for each level of cycle (day or night), and depth (250m, 500m, or 1000m).