MCMC samplers for models fit in `spOccupancy`

Jeffrey W. Doser

October 01, 2021

Contents

1 Introduction

This vignette provides statistical details on the MCMC algorithms used to fit each occupancy model in `spOccupancy`. We provide detailed descriptions of the joint posterior distributions for each model, how each parameter is updated in the model fitting process, and provide relevant citations to more specific documentation of the approaches where necessary. We also provide information on the composition sampling algorithms used for each model to predict at out-of-sample locations.

2 Single species occupancy model (`PGOcc`)

2.1 Model description

Let z_j be the true presence (1) or absence (0) of a species at site j, with $j = 1, \ldots, J$. We assume this latent occupancy process arises from a Bernoulli process following

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j), \\ \text{logit}(\psi_j) &= \boldsymbol{x}_j' \boldsymbol{\beta}, \end{aligned} \tag{1}$$

where ψ_j is the probability of occurrence at site j, which is a function of site-specific covariates $\boldsymbol{X}$ and a vector of regression coefficients ($\boldsymbol{\beta}$).

We do not directly observe z_j and rather we observe an imperfect representation of the latent occurrence process. Let $y_{j,k}$ be the observed detection (1) or nondetection (0) of a species of interest at site j during replicate k for each of $k = 1, \ldots, K_j$ replicates at each site j. We envision the detection-nondetection data as arising from a Bernoulli process conditional on the true latent occurrence process:

$$\begin{aligned} y_{j,k} &\sim \text{Bernoulli}(p_{j,k} z_j), \\ \text{logit}(p_{j,k}) &= \boldsymbol{v}_{j,k}' \boldsymbol{\alpha}, \end{aligned} \tag{2}$$

where $p_{j,k}$ is the probability of detecting a species at site j during replicate k (given it is present at site j), which is a function of site and replicate specific covariates $\boldsymbol{V}$ and a vector of regression coefficients ($\boldsymbol{\alpha}$).

We ssume multivariate normal priors for the occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) regression coefficients to complete the Bayesian specification of a single species occupancy model. Traditionally, when estimation occurs in a Bayesian framework, the regression coefficients for occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) must be updated using Metropolis updates, which can lead to slow convergence times and bad mixing of MCMC chains (Clark and Altwegg 2019). Instead, we introduce Polya-Gamma latent variables (Polson, Scott, and Windle 2013) for both the occurrence and detection portions of the model, which induces Gibbs updates for all parameters in the single species occupancy model.

More specifically, let $\omega_{j,\beta}$ follow a Polya-Gamma distribution with parameters 1 and 0, i.e., $\omega_{j,\beta} \sim \text{PG}(1, 0)$. Given this latent variable, we can express the Bernoulli process of z_j as

$$\begin{aligned}\psi_j^{z_j}(1-\psi_j)^{1-z_j} &= \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})^{z_i}}{1+\exp(\boldsymbol{x}_j'\boldsymbol{\beta})} \\ &= \exp(\kappa_j \boldsymbol{x}_j'\boldsymbol{\beta}) \int \exp(-\frac{\omega_{j,\beta}}{2}(\boldsymbol{x}_j'\boldsymbol{\beta})^2) p(\omega_{j,\beta} \mid 1,0) d\omega_{j,\beta},\end{aligned} \tag{3}$$

where $\kappa_j = z_j - 0.5$ and $p(\omega_{j,\beta})$ is the probability density function of a Polya-Gamma distribution with parameters 1 and 0 (Polson, Scott, and Windle 2013). Similarly, we define $\omega_{j,k,\alpha} \sim \text{PG}(1,0)$ as a Polya-Gamma latent variable for the detection portion of the occupancy model, which results in a simliar re-expression of the Bernoulli likelihood for $y_{j,k}$ as we showed above for z_j. These re-expressions of the Bernoulli processes result in Gibbs updates for both the occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) regression coefficients when they are assigned normal priors (Polson, Scott, and Windle 2013; Clark and Altwegg 2019).

Our full joint posterior for a single species occupancy model thus takes the following form:

$$\begin{aligned}[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha \mid \boldsymbol{Y}] \propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} &\text{Bernoulli}(y_{j,k} \mid p_{j,k} z_j) \times \\ &\text{Bernoulli}(z_j \mid \psi_j) \times \\ &\text{PG}(\omega_{j,\beta} \mid 1,0) \times \\ &\text{PG}(\omega_{j,k,\alpha} \mid 1,0) \times \\ &\text{Normal}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \\ &\text{Normal}(\boldsymbol{\alpha} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)\end{aligned}$$

2.2 MCMC sampler

The Polya-Gamma data augmentation induces a Gibbs update for all parameters in the single species occupancy model. We first sample the occurrence and detection auxiliary variables from

$$\begin{aligned}\omega_{j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j'\boldsymbol{\beta}), \\ \omega_{j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}'\boldsymbol{\alpha}),\end{aligned} \tag{4}$$

respectively. We next sample the occurrence regression coefficients $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \mid \cdot \sim \text{Normal}\Big([\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}[\boldsymbol{X}'(\boldsymbol{z} - 0.5\boldsymbol{1}_J) + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta], [\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}\Big), \tag{5}$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,\beta}, \ldots, \omega_{J,\beta})$. Similarly, we sample the detection regression coefficients $\boldsymbol{\alpha}$ from

$$\boldsymbol{\alpha} \mid \cdot \sim \text{Normal}\Big([\boldsymbol{\Sigma}_\alpha^{-1} + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}[\tilde{\boldsymbol{V}}'(\tilde{\boldsymbol{y}} - 0.5\boldsymbol{1}_{J^*}) + \boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha], [\boldsymbol{\Sigma}_\alpha^{-1} + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}\Big). \tag{6}$$

The detection regression coefficients $\boldsymbol{\alpha}$ are only informed by the locations where $z_j = 1$, since we assume no false positive detections in the standard occupancy model. We define J^* as the total number of sites at the current iteration of the MCMC with $z_j = 1$. $\boldsymbol{S}_\alpha$ is a diagonal matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,1,\alpha}, \ldots, \omega_{J^*,K_{J^*},\alpha})$. The matrix $\tilde{\boldsymbol{V}}$ is the matrix of detection covariates associated with the sites where $z_j = 1$. Similarly, $\tilde{\boldsymbol{y}}$ is a vector of stacked detection-nondetection data values at the entries associated with $z_j = 1$.

Finally, z_j is set to 1 for all sites where there is at least one detection, and thus we only need to sample z_j at sites where there are no detections. Thus, for all locations with no detections, we sample z_j according to

$$z_j \mid \cdot \sim \text{Bernoulli}\left(\frac{\psi_j \prod_{k=1}^{K_j}(1-p_{j,k})}{1-\psi_j+\psi_j \prod_{k=1}^{K_j}(1-p_{j,k})}\right). \tag{7}$$

2.3 Prediction

Prediction for a nonspatial single species occupancy model is a simple composition sampling problem. Given a set of occurrence covariates at a set of non-sampled locations ($\boldsymbol{X}_0$), we can derive the latent occurrence probability and the latent occurrence state at each non-sampled site $j = 1, \ldots, J_0$ for each posterior sample q of the MCMC sampler following

$$\begin{aligned} \text{logit}(\psi_j^{(q)}) &= \boldsymbol{x}_{0,j}'\boldsymbol{\beta}^{(q)}, \\ z_j^{(q)} &\sim \text{Bernoulli}(\psi_j^{(q)}). \end{aligned} \tag{8}$$

3 Single species spatial occupancy model (`spPGOcc`)

3.1 Gaussian Process formulation

3.1.1 Model description

We extend the previous single species occupancy model to incorporate a spatial Gaussian Process that accounts for unexplained spatial variation in species occurrence across a region of interest. The species-specific occurrence probability at site j, ψ_j, now takes the form

$$\text{logit}(\psi_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \text{w}_j, \tag{9}$$

where w_j is a realization from a zero-mean spatial Gaussian Process, i.e.,

$$\mathbf{w} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})). \tag{10}$$

We define $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})$ as a $J \times J$ covaraince matrix that is a function of the distances between any pair of site coordinates $\boldsymbol{s}$ and $\boldsymbol{s}'$ and a set of parameters ($\boldsymbol{\theta}$) that govern the spatial process. The vector $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta} = \{\sigma^2, \phi, \nu\}$, where σ^2 is a spatial variance parameter, ϕ is a spatial decay parameter, and ν is a spatial smoothness parameter. ν is only specified when using a Matern correlation function.

The detection portion of the occupancy model remains unchanged from the non-spatial occupancy model and follows Equation (2). Formulation of Polya-Gamma latent variables is also exactly analogous to the nonspatial model (Equation (3)), with all references to ψ_j now including the latent spatial random effects in addition to the site-level covariates.

Following standard recommendations for point-referenced spatial data (Banerjee, Carlin, and Gelfand 2003), we assign an inverse-Gamma prior to the spatial variance parameter and uniform priors to the spatial decay and spatial smoothness parameters. Our full joint posterior distribution takes the following form, where IG stands for the inverse-Gamma distribution:

$$
\begin{aligned}
[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha, \mathbf{w}, \boldsymbol{\theta} \mid \boldsymbol{Y}] \propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} & \text{Bernoulli}(y_{j,k} \mid p_{j,k} z_j) \times \\
& \text{Bernoulli}(z_j \mid \psi_j) \times \\
& \text{Normal}(\mathbf{w} \mid \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})) \times \\
& \text{PG}(\omega_{j,\beta} \mid 1, 0) \times \\
& \text{PG}(\omega_{j,k,\alpha} \mid 1, 0) \times \\
& \text{Normal}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \\
& \text{Normal}(\boldsymbol{\alpha} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \times \\
& \text{IG}(\sigma^2 \mid a_{\sigma^2}, b_{\sigma^2}) \times \\
& \text{Uniform}(\phi \mid a_\phi, b_\phi) \times \\
& \text{Uniform}(\nu \mid a_\nu, b_\nu)
\end{aligned}
$$

3.1.2 MCMC sampler

We first sample the occurrence and detection auxiliary variables from

$$
\begin{aligned}
\omega_{j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j' \boldsymbol{\beta} + \text{w}_j), \\
\omega_{j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}' \boldsymbol{\alpha}),
\end{aligned} \tag{11}
$$

The Polya-Gamma scheme induces a Gibbs update for the occurrence regression coefficients, which are updated at each iteration according to

$$
\boldsymbol{\beta} \mid \cdot \sim \text{Normal}\Big([\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}' \boldsymbol{S}_\beta \boldsymbol{X}]^{-1} [\boldsymbol{X}' (\boldsymbol{z} - 0.5 \mathbf{1}_J - \boldsymbol{S}_\beta \mathbf{w}) + \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta], [\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}' \boldsymbol{S}_\beta \boldsymbol{X}]^{-1} \Big), \tag{12}
$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,\beta}, \ldots, \omega_{J,\beta})$.

The full conditional for the detection regression coefficients is the same as in the non-spatial model shown in Equation (6).

The spatial variance parameter, σ^2, is sampled via a Gibbs update of the form

$$
\sigma^2 \mid \cdot \sim \text{IG}\Big(\frac{J}{2} + a_{\sigma^2}, \frac{\mathbf{w}' \boldsymbol{R}^{-1} \mathbf{w}}{2} + b_{\sigma^2} \Big), \tag{13}
$$

where $\boldsymbol{R}$ is a $J \times J$ spatial correlation matrix.

The full conditional distributions for the spatial range parameter, ϕ, and spatial smoothness parameter, ν, are not available in closed form, and thus we use random walk Metropolis updates (e.g., Robert and Casella (2013)) to update the parameters. We use a random-walk Metroplis step with a multivariate normal proposal distribution (either of dimension 1 or of dimension 2 if Matern covariance function is used). To use the normal distribution as a proposal distribution, we transform the parameters to have a support spanning the entire real line, including a Jacobian adjustment for the Metropolis step. Tuning parameters are adaptively updated using Adaptive MCMC following Roberts and Rosenthal (2009).

The Polya-Gamma data augmentation scheme also enables a Gibbs update for the latent spatial Gaussian process ($\mathbf{w}$), as opposed to a traditional spatial occupancy model that requires a Metropolis update for the latent spatial process. The spatial process is updated according to

$$\mathbf{w} \mid \cdot \sim \text{Normal}\Big([\boldsymbol{S}_\beta + \boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{z} - 0.5\mathbf{1}_J - \boldsymbol{S}_\beta \boldsymbol{X}\boldsymbol{\beta}], [\boldsymbol{S}_\beta + \boldsymbol{\Sigma}^{-1}]^{-1}\Big). \tag{14}$$

Finally, for all sites with no detections, the latent occurrence values z_j are updated following Equation (7).

3.1.3 Prediction

Prediction for spatial occupancy models requires use of standard results for conditional multivariate normal distributions (Banerjee, Carlin, and Gelfand 2003). To predict latent occurrence and occurrence probability at non-sampled sites, we first need to predict the spatial process at the unobserved locations. Let $\mathbf{w}_0$ denote the spatial process at J_0 non-sampled locations. We assume that $\mathbf{w}_0$ and $\mathbf{w}$ (the spatial process at observed locations) arise from a multivariate normal distribution following

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{w}_0 \end{bmatrix} \mid \cdot \sim \text{Normal}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right), \tag{15}$$

where $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{12}$ is the $J \times J_0$ cross-covariance matrix between $\mathbf{w}$ and $\mathbf{w}_0$, and $\boldsymbol{\Sigma}_{22}$ is the variance-covariance matrix for $\mathbf{w}_0$. Using conditional multivariate normal theory, this results in the following posterior predictive distribution for the spatial process at nonsampled locations

$$\mathbf{w}_0 \sim \text{Normal}(\boldsymbol{\Sigma}'_{12}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{w}, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}'_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \tag{16}$$

We can use composition sampling to sample from this posterior predictive distribution by using the values for $\mathbf{w}$ at each sample q of the posterior distribution. This will generate a full predictive posterior sample which we can summarize with full uncertainty quantification.

Predicting all J_0 locations jointly can be expensive when J_0 is large. Thus, we perform independent individual predictions of the spatial process at each non-sampled location $j = 1, \ldots, J_0$. Finally, to predict the latent occurrence and latent occurrence probability at each non-sampled site j, we perform the following steps for each posterior sample q.

1. Sample $\text{w}_{0,j}^{(q)}$ from Equation (16), substituting in the current values at sample q of the spatial parameters and latent spatial process at the observed locations.
2. Compute the latent occurrence probability $\psi_j^{(q)}$ as $\text{logit}^{-1}(\boldsymbol{x}_{0,j}\boldsymbol{\beta}^{(q)} + \text{w}_{0,j}^{(q)})$.
3. Sample the latent occurrence from $z_j^{(q)} \sim \text{Bernoulli}(\psi_j^{(q)})$.

3.2 Nearest Neighbor Gaussian Process formulation

When the number of sites is moderately large, say 1000, the above described spatial Gaussian process model can be drastically slow as a result of needing to take the inverse of the spatial covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})$ at each MCMC iteration. Numerous approximation methods exist to reduce this computational cost (Heaton et al. 2019). One attractive approach is the Nearest Neighbor Gaussian Process (NNGP, Datta et al. (2016)). Instead of modeling the spatial process using a full GP as shown in Equation (10), we replace the GP prior specification with a NNGP, which leads to drastic increases in run time with nearly identicial inference and prediction as the full GP specification. See Datta et al. (2016) for theoretical details on the NNGP and its relationship to the full GP.

3.2.1 Model description

The joint posterior distribution for the NNGP model is exactly the same as that of the full GP spatial occupancy model except the Gausian Process specification assigned to the latent spatial random effects $\mathbf{w}$ is replaced with an NNGP prior (Datta et al. 2016).

3.2.2 MCMC sampler

Full conditionals for the Polya-Gamma latent variables, occurrence regression coefficients, detection regression coefficients, spatial range (and smoothness if applicable) parameter, and the latent occurrence values are sampled in the same manner as done for the full GP spatial occupancy model. The full conditional for the spatial variance parameter σ^2 similarly takes the form of an inverse-Gamma distribution. The Polya-Gamma data augmentation scheme induces the following full conditional for the latent spatial process when using an NNGP prior:

$$\mathbf{w} \mid \cdot \sim \text{Normal}(\mathbf{B}[\boldsymbol{S}_{\beta}^{-1}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\boldsymbol{\beta})], \mathbf{B}) \tag{17}$$

where $\boldsymbol{S}_{\beta}$ is a diagonal $J \times J$ matrix with diagonal entries equal to the latent Polya-Gamma variable values, $\tilde{z}_j = \frac{z_j - 0.5}{\omega_{j,\beta}}$, and $\mathbf{B} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})^{-1} + \boldsymbol{S}_{\beta}$. $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})$ is the sparse NNGP covariance matrix. For details of the NNGP covariance matrix, see Datta et al. (2016) and Finley et al. (2019).

As described by Finley, Datta, and Banerjee (2020) and Datta et al. (2016), the above block update of $\mathbf{w}$ is not computationally practical. Instead, we sequentially update the full conditionals individual for each $j = 1, \ldots, J$ element of $\mathbf{w}$ following the algorithm in Datta et al. (2016). This ensures each update of the full latent spatial random efffects vector occurs in $O(n)$ floating point operations (FLOPs).

3.2.3 Prediction

Prediction for the NNGP occupancy model follows a similar algorithm to that of the full GP spatial occupancy model. We first sample the observed spatial random effects when fitting the model, use these random effects to generate the spatial random effects at new locations, and subsequently use the predicted spatial random effects to generate predictions of latent occurrence and occurrence probability. More specifically, our approach for prediction follows exactly Algorithm 2 of Finley et al. (2019), which we reproduce in the context of occupancy models in the following.

The posterior predictive distribution for the spatial process at unobserved locations ($\mathbf{w}_0$) is

$$\mathbf{w}_0 \mid \cdot \sim \text{Normal}(\mathbf{B}^{-1}[\boldsymbol{S}_{\beta}^{-1}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\boldsymbol{\beta})], \mathbf{B}^{-1}). \tag{18}$$

Our predictive algorithm thus takes the following steps for each posterior sample q:

1. Sample $\text{w}_{0,j}^{(q)}$ from Equation (18), substituting in the current values at sample q of the spatal parameters and latent spatial process at the observed locations.
2. Compute the latent occurrence probability $\psi_j^{(q)}$ as $\text{logit}^{-1}(\boldsymbol{x}_{0,j}\boldsymbol{\beta}^{(q)} + \text{w}_{0,j}^{(q)})$.
3. Sample the latent occurrence from $z_j^{(q)} \sim \text{Bernoulli}(\psi_j^{(q)})$.

4 Multispecies occupancy model (`msPGOcc`)

4.1 Model description

Let $z_{i,j}$ be the true presence (1) or absence (0) of a species i at site j, with $j = 1, \ldots, J$ and $i = 1, \ldots, N$. We assume the latent occurrence process arises from a Bernoulli process following

$$\begin{aligned} z_{i,j} &\sim \text{Bernoulli}(\psi_{i,j}), \\ \text{logit}(\psi_{i,j}) &= \boldsymbol{x}_j' \cdot \boldsymbol{\beta}_i, \end{aligned} \tag{19}$$

where $\psi_{i,j}$ is the probability of occurrence of species i at site j, which is a function of site-specific covariates $\boldsymbol{X}$ and a vector of species-specific regression coefficients ($\boldsymbol{\beta}_i$). The regression coefficients in multispecies occupancy models are envisioned as random effects arising from a common community level distribution:

$$\boldsymbol{\beta}_i \sim \text{Normal}(\bar{\boldsymbol{\beta}}, \boldsymbol{T}_\beta), \tag{20}$$

where $\bar{\boldsymbol{\beta}}$ is a vector of community level mean effects for each occurrence covariate effect (including the intercept) and $\boldsymbol{T}_\beta$ is a diagonal matrix with diagonal elements $\boldsymbol{\tau}_\beta^2$ that represent the variability of each occurrence covariate effect among species in the community.

We do not directly observe $z_{i,j}$ and rather we observe an imperfect representation of the latent occurrence process. Let $y_{i,j,k}$ be the observed detection (1) or nondetection (0) of a species i of interest at site j during replicate k for each of $k = 1, \ldots, K_j$ replicates at each site j. We envision the detection-nondetection data as arising from a Bernoulli process conditional on the true latent occurrence process:

$$\begin{aligned} y_{i,j,k} &\sim \text{Bernoulli}(p_{i,j,k} \cdot z_{i,j}), \\ \text{logit}(p_{i,j,k}) &= \boldsymbol{v}'_{i,j,k} \cdot \boldsymbol{\alpha}_i, \end{aligned} \tag{21}$$

where $p_{i,j,k}$ is the probability of detecting species i at site j during replicate k (given it is present at site j), which is a function of site and replicate specific covariates $\boldsymbol{V}$ and a vector of species-specific regression coefficients ($\boldsymbol{\alpha}_i$). Similarly to the occurrence regression coefficients, the species specific detection coefficients are envisioned as random effects arising from a common community level distribution:

$$\boldsymbol{\alpha}_i \sim \text{Normal}(\bar{\boldsymbol{\alpha}}, \boldsymbol{T}_\alpha), \tag{22}$$

where $\bar{\boldsymbol{\alpha}}$ is a vector of community level mean effects for each detection covariate effect (including the intercept) and $\boldsymbol{T}_\alpha$ is a diagonal matrix with diagonal elements $\boldsymbol{\tau}_\alpha^2$ that represent the variability of each detection covariate effect among species in the community.

We assign multivariate normal priors for the community level occurrence ($\boldsymbol{bar\beta}$) and detection ($\bar{\boldsymbol{\beta}}$) means, and assign indepdent inverse-Gamma priors on the community level occurrence (τ_β^2) and detection (τ_α^2) variance parameters. Analogous to the single species occupancy model, we implement the model using Polya-Gamma data augmentation which induces fully Gibbs updates for all parameters. We specify Polya-Gamma data augmented variables for each species ($\omega_{i,j,\beta}$, $\omega_{i,j,k,\alpha}$, which follow the same scheme as that for the single species model in Equation (3), except all parameters in the equation are now indexed by i for each species. The full joint posterior distribution thus takes the following form:

$$
\begin{aligned}
[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha, \boldsymbol{\mu}_\beta, \boldsymbol{\mu}_\alpha, \boldsymbol{\tau}^2_\beta, \boldsymbol{\tau}^2_\alpha \mid \boldsymbol{y}] \propto \prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K_j} & \text{Bernoulli}(y_{i,j,k} \mid p_{i,j,k} z_{i,j}) \times \\
& \text{Bernoulli}(z_{i,j} \mid \psi_{i,j}) \times \\
& \text{PG}(\omega_{i,j,\beta} \mid 1, 0) \times \\
& \text{PG}(\omega_{i,j,k,\alpha} \mid 1, 0) \times \\
& \text{Normal}(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \boldsymbol{T}_\beta) \times \\
& \text{Normal}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\alpha}}, \boldsymbol{T}_\alpha) \\
& \text{Normal}(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
& \text{Normal}(\bar{\boldsymbol{\alpha}} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \\
& \prod_{r=1}^{n_\beta} \text{IG}(\tau^2_{r,\beta} \mid a_{r,\beta}, b_{r,\beta}) \\
& \prod_{t=1}^{n_\alpha} \text{IG}(\tau^2_{t,\alpha} \mid a_{t,\alpha}, b_{t,\alpha}),
\end{aligned}
$$

where r and t index across the number of occurrence and detection regression parameters, respectively.

4.2 MCMC sampler

The Polya-Gamma data augmentation induces a Gibbs update for all parameters in the multispecies occupancy model. For each iteration, we first sample all community level parameters followed by species level parameters. We first sample the community level regression coefficients $\bar{\boldsymbol{\beta}}$ from

$$
\bar{\boldsymbol{\beta}} \mid \cdot \sim \text{Normal}([\boldsymbol{\Sigma}_\beta^{-1} + N\boldsymbol{T}_\beta^{-1}]^{-1}\Big[\sum_{i=1}^{N}(\boldsymbol{T}_\beta \boldsymbol{\beta}_i) + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\Big], [\boldsymbol{\Sigma}_\beta^{-1} + N\boldsymbol{T}_\beta^{-1}]^{-1}). \tag{23}
$$

Similarly, we next sample the community level regression coefficients $\bar{\boldsymbol{\alpha}}$ from

$$
\bar{\boldsymbol{\alpha}} \mid \cdot \sim \text{Normal}([\boldsymbol{\Sigma}_\alpha^{-1} + N\boldsymbol{T}_\alpha^{-1}]^{-1}\Big[\sum_{i=1}^{N}(\boldsymbol{T}_\alpha \boldsymbol{\alpha}_i) + \boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha\Big], [\boldsymbol{\Sigma}_\alpha^{-1} + N\boldsymbol{T}_\alpha^{-1}]^{-1}). \tag{24}
$$

Next, we sample the community level occurrence variance parameter for each regression coefficient, $\tau^2_{r,\beta}$, from the following inverse-Gamma full conditional:

$$
\tau^2_{r,\beta} \mid \cdot \sim \text{IG}(a_{r,\beta} + \frac{N}{2}, a_{r,\beta} + \frac{\sum_{i=1}^{N}(\beta_{i,r} - \bar{\beta}_r)^2}{2}). \tag{25}
$$

Similarly, we next sample the community level detection variance parameter for each regression coefficient:

$$
\tau^2_{t,\alpha} \mid \cdot \sim \text{IG}(a_{t,\alpha} + \frac{N}{2}, a_{t,\alpha} + \frac{\sum_{i=1}^{N}(\alpha_{i,t} - \bar{\alpha}_t)^2}{2}). \tag{26}
$$

We now sample all species level coefficients. The coefficients are sampled one at a time for each species. First, we sample the occurrence and detection auxiliary variables for species i from

$$\begin{aligned}\omega_{i,j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j'\boldsymbol{\beta}_i),\\ \omega_{i,j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}'\boldsymbol{\alpha}_i).\end{aligned} \tag{27}$$

The occurrence regression coefficients for species i are subsequently drawn from the following multivariate Normal full conditional distribution

$$\boldsymbol{\beta}_i \mid \cdot \sim \text{Normal}\Big([\boldsymbol{T}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}[\boldsymbol{X}'(\boldsymbol{z}_i - 0.5\boldsymbol{1}_J) + \boldsymbol{T}_\beta^{-1}\bar{\boldsymbol{\beta}}], [\boldsymbol{T}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}\Big), \tag{28}$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal etnries equal to the latent Polya-Gamma variable values for species i. Similarly, we sample the detection regression coefficients for species i from

$$\boldsymbol{\alpha}_i \mid \cdot \sim \text{Normal}\Big([\boldsymbol{T}_\alpha^{-1} + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}[\tilde{\boldsymbol{V}}'(\tilde{\boldsymbol{y}}_i - 0.5\boldsymbol{1}_{J_i^*}) + \boldsymbol{T}_\alpha^{-1}\bar{\boldsymbol{\alpha}}], [\boldsymbol{T}_\alpha^{-1} + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}\Big). \tag{29}$$

The species-level detection regression coefficients $\boldsymbol{\alpha}_i$ are only informed by the locations where $z_{i,j} = 1$, since we assume no false positive detections in the standard occupancy model. We define J_i^* as the total number of sites at the current iteration of the MCMC with $z_{i,j} = 1$. $\boldsymbol{S}_\alpha$ is a diagonal matrix with diagonal entries equal to the latent Polya-gamma variable values $(\omega_{i,1,1,\alpha}, \ldots, \omega_{i,J_i^*,K_{J_i^*},\alpha})$. The matrix $\tilde{\boldsymbol{V}}$ is the matrix of detection covariates associated with the sites where $z_{i,j} = 1$. Similarly, $\tilde{\boldsymbol{y}}_i$ is a vector of stacked detection-nondetection data values at the entries associated with $z_{i,j} = 1$.

Finally, we sample the latent occurrence states for each species. $z_{i,j}$ is set to 1 for all sites where there is at least one detection of species i, and so we only need to sample $z_{i,j}$ at sites where there are no detections. Thus, for all locations with no detections of the species i, we sample $z_{i,j}$ according to

$$z_{i,j} \mid \cdot \sim \text{Bernoulli}\left(\frac{\psi_{i,j}\prod_{k=1}^{K_j}(1 - p_{i,j,k})}{1 - \psi_{i,j} + \psi_{i,j}\prod_{k=1}^{K_j}(1 - p_{i,j,k})}\right). \tag{30}$$

Note the full conditional for $z_{i,j}$ is exactly the same as that for the single species occupancy model in Equation (7), except all values are now additionally indexed by species (i).

4.3 Prediction

Prediction for a nonspatial multispecies occupancy model is a simple composition sampling problem exactly analogous to the single species model. Given a set of occurrence covariates at a set of non-sampled locations $(\boldsymbol{X}_0)$, we can derive the latent occurrence probability and the latent occurrence state at each non-sampled site $j = 1, \ldots, J_0$ for each species i for each posterior sample q of the MCMC sampler following

$$\begin{aligned}\text{logit}(\psi_{i,j}^{(q)}) &= \boldsymbol{x}_{0,j}'\boldsymbol{\beta}_i^{(q)},\\ z_{i,j}^{(q)} &\sim \text{Bernoulli}(\psi_{i,j}^{(q)}).\end{aligned} \tag{31}$$

5 Multispecies spatial occupancy model (`spMsPGOcc`)

5.1 Gaussian Process formulation

5.1.1 Model description

We extend the previous multispecies occupancy model to incorporate a distinct spatial Gaussian Process (GP) for each species that accounts for unexplained spatial variation in each individual species occurrence

across a spatial region. Occurrence probability for species i at site j, $\psi_{i,j}$, now takes the form

$$\text{logit}(\psi_{i,j}) = \boldsymbol{x}_j'\boldsymbol{\beta}_i + \text{w}_{i,j}, \tag{32}$$

where the species-specific regression coefficients $\boldsymbol{\beta}_i$ follow the community level distribution in Equation (20), and $\text{w}_{i,j}$ is a realization from a zero-mean spatial GP, i.e.,

$$\mathbf{w}_i \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_i(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta}_i)). \tag{33}$$

We define $\boldsymbol{\Sigma}_i(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta}_i)$ as a $J \times J$ covariance matrix that is a function of the distances between any pair of site coordinates $\boldsymbol{s}$ and $\boldsymbol{s}'$ and a set of parameters ($\boldsymbol{\theta}_i$) that govern the spatial process. The vector $\boldsymbol{\theta}_i$ is equal to $\boldsymbol{\theta}_i = \{\sigma_i^2, \phi_i, \nu_i\}$, where σ_i^2 is a spatial variance parameter for species i, ϕ_i is a spatial decay parameter for species i, and ν_i is a spatial smoothness parameter for species i. ν_i is only specified when using a Matern correlation function.

Note that we estimate a distinct parameter vector $\boldsymbol{\theta}_i$ for each species and assume the spatial processes are independent of each other. This is a naive approach for incorporating spatial processes in a multispecies occupancy model, as we do not leverage the potential correlation in spatial processes among species in a linear model of coregionalization approach (Gelfand et al. 2004). Despite the simplicity of the approach, such models have been shown to yield improved insight in species distributions across broad locations (Wright et al. 2021), and the Bayesian shrinkage component of the multispecies model (Equation (20)) will make estimates in a multispecies spatial occupancy model more precise than a single species spatial occupancy model, in particular for rare species. In future implementations of `spOccupancy` we plan to implement multispecies occupancy models in a more rich inferential framework that leverages between species correlations in spatial processes and non-spatial components, similar to the models of Tobler et al. (2019) and Taylor-Rodriguez et al. (2019).

The detection portion of the multispecies spatial occupancy model remains unchanged from the non-spatial multispecies occupancy model and follows Equations (21) and (22). Formulation of Polya-Gamma latent variables is also exactly analogous to the nonspatial model (Equation (3)), with all parameters including an index for species (i) and all references to $\psi_{i,j}$ now including the latent spatial random effects in addition to the site-level covariates.

Following standard recommendations for point-referenced spatial data (Banerjee, Carlin, and Gelfand 2003), we assign an inverse-Gamma prior to the spatial variance parameter for each species and uniform priors to the spatial decay and spatial smoothness parameters for each species. Our full joint posterior distribution takes the following form, where IG stands for the inverse-Gamma distribution:

$$
\begin{aligned}
[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha, \boldsymbol{\mu}_\beta, \boldsymbol{\mu}_\alpha, \boldsymbol{\tau}_\beta^2, \boldsymbol{\tau}_\alpha^2 \mid \boldsymbol{y}] \propto & \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K_j} \text{Bernoulli}(y_{i,j,k} \mid p_{i,j,k} z_{i,j}) \times \\
& \text{Bernoulli}(z_{i,j} \mid \psi_{i,j}) \times \\
& \text{PG}(\omega_{i,j,\beta} \mid 1, 0) \times \\
& \text{PG}(\omega_{i,j,k,\alpha} \mid 1, 0) \times \\
& \text{Normal}(\mathbf{w} \mid \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})) \times \\
& \text{Normal}(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \boldsymbol{T}_\beta) \times \\
& \text{Normal}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\alpha}}, \boldsymbol{T}_\alpha) \\
& \text{Normal}(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
& \text{Normal}(\bar{\boldsymbol{\alpha}} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \\
& \prod_{r=1}^{n_\beta} \text{IG}(\tau_{r,\beta}^2 \mid a_{r,\beta}, b_{r,\beta}) \\
& \prod_{t=1}^{n_\alpha} \text{IG}(\tau_{t,\alpha}^2 \mid a_{t,\alpha}, b_{t,\alpha}) \\
& \text{IG}(\sigma_i^2 \mid a_{\sigma^2,i}, b_{\sigma^2,i}) \times \\
& \text{Uniform}(\phi_i \mid a_{\phi,i}, b_{\phi,i}) \times \\
& \text{Uniform}(\nu_i \mid a_{\nu,i}, b_{\nu,i})
\end{aligned}
$$

5.1.2 MCMC sampler

The Polya-Gamma data augmentation induces a Gibbs update for all parameters in the multispecies spatial occupancy model except the spatial range parameters (ϕ_i) and the spatial smoothness parameters ν_i if specified. For each iteration, we first sample all community level parameters followed by species level parameters. Full conditional distributions for all community level parameters are exactly the same as those for the nonspatial multispecies model. See Equations (23)-(26).

The species-level coefficients are sampled one at a time for each species. First, we sample the occurrence and detection auxiliary variables for species i from

$$
\begin{aligned}
\omega_{i,j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j' \boldsymbol{\beta}_i + \text{w}_{i,j}), \\
\omega_{i,j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}' \boldsymbol{\alpha}_i).
\end{aligned} \tag{34}
$$

The occurrence regression coefficients for species i are subsequently drawn from the following multivariate normal full conditional distribution

$$
\boldsymbol{\beta}_i \mid \cdot \sim \text{Normal}\Big([\boldsymbol{T}_\beta^{-1} + \boldsymbol{X}' \boldsymbol{S}_\beta \boldsymbol{X}]^{-1} [\boldsymbol{X}' (\boldsymbol{z}_i - 0.5 \mathbf{1}_J - \boldsymbol{S}_\beta \mathbf{w}_i) + \boldsymbol{T}_\beta^{-1} \bar{\boldsymbol{\beta}}], [\boldsymbol{T}_\beta^{-1} + \boldsymbol{X}' \boldsymbol{S}_\beta \boldsymbol{X}]^{-1}\Big), \tag{35}
$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal etnries equal to the latent Polya-Gamma variable values for species i. The full conditioanl for the species-level detection regression coefficients is the same as in the non-spatial model shown in Equation (29).

The spatial variance parameter for species i, σ_i^2, is sampled via a Gibbs update of the form

$$
\sigma_i^2 \mid \cdot \sim \text{IG}\Big(\frac{J}{2} + a_{\sigma^2,i}, \frac{\mathbf{w}_i' \boldsymbol{R}_i^{-1} \mathbf{w}_i}{2} + b_{\sigma^2,i}\Big), \tag{36}
$$

where $\boldsymbol{R}$ is a $J \times J$ spatial correlation matrix. The full conditional distributions for the species-specific spatial range parameters, ϕ_i, and spatial smoothness parameters, ν_i are not available in closed form, and so they are updated using random walk Metropolis updates following the same procedure as described for the single species spatial occupancy models.

The Polya-Gamma data augmentation scheme induces a Gibbs update for the latent spatial Gaussian process for each species, which is updated according to

$$\mathbf{w}_i \mid \cdot \sim \text{Normal}\Big([\boldsymbol{S}_\beta + \boldsymbol{\Sigma}_i^{-1}]^{-1}[\boldsymbol{z}_i - 0.5\mathbf{1}_J - \boldsymbol{S}_\beta \boldsymbol{X}\boldsymbol{\beta}_i], [\boldsymbol{S}_\beta + \boldsymbol{\Sigma}_i^{-1}]^{-1}\Big). \tag{37}$$

Finally, for all sites with no detections for a given species, the latent occurrence values $z_{i,j}$ are updated following Equation (30).

5.1.3 Prediction

Because we assume independence between the spatial processes of the different species, prediction for the multispecies spatial occupancy model is exactly analogous to prediction for the single species spatial occupancy model described in Section 3.1.3, except prediction is done for each species i using the species specific values for that species. See Section 3.1.3 for the algorithm, noting that for the multispecies model all values are additionally indexed by species (i).

5.2 Nearest Neighbor Gaussian Process formulation

As with the single species model, we also implement `spMsPGOcc` with an NNGP. Use of the NNGP leads to even larger computational advances for the multispecies occupancy models, as we now replace each of the independent GPs for each of the N species with an independent NNGP.

5.2.1 Model description

The joint posterior distribtuion for the multispecies NNGP occupancy model is exactly the same as that of the full GP multispecies model except the GP prior assigned to the latent spatial random effects $\mathbf{w}_i$ for each species is replaced with an NNGP prior (Datta et al. 2016).

5.2.2 MCMC sampler and prediction

The full conditionals for all variables except the spatial variance parameters σ_i^2 and the latent spatial process $\mathbf{w}_i$ follow the same full conditionals as described in the full GP multispecies spatial occupancy model. The full conditionals for σ_i^2 and $\mathbf{w}_i$ follow exactly those described for the single species NNGP model in Section 3.2.2, where all parameters are now indexed by each species i. Prediction is also exactly analogous to that described for the single species NNGP model in Section 3.2.3.

6 Single species integrated occupancy model (`intPGOcc`)

Data integration is a model-based approach that leverages multiple data sources to provide inference and prediction on some latent process of interest. Data integration is particularly relevant in ecology as many data sources are often collected to study a single ecological phenomenon, with each data source having pros and cons. Often, multiple detection-nondetection data sources are available to study the occurrence and distribution of some species of interest. For example, both human point count surveys and autonomous recording units could be used to monitor a bird species of conservation concern. Different types of data

have different sources of observation error, which should be explicitly incorporated into a model to avoid attributing any variation in detection probability to the true ecological process. Here we describe single species integrated occupancy models, which combine multiple sources of detection-nondetection data (which may or may not be replicated) in a single hierarchical modeling framework.

6.1 Model description

The biological process model is exactly the same as single species occupancy models, which we now describe again for clarity. Let z_j be the presence or absence of a species at site j, with $j = 1, \ldots, J$. We assume this latent occurrence process arises from a Bernoulli process following

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j), \\ \text{logit}(\psi_j) &= \boldsymbol{x}_j'\boldsymbol{\beta}, \end{aligned} \tag{38}$$

where ψ_j is the probability of occurrence at site j, which is a function of site-specific covariates $\boldsymbol{X}$ and a vector of regression coefficients ($\boldsymbol{\beta}$).

We do not directly observe z_j and rather we observe an imperfect representation of the latent occurrence process. In integrated models, we have $r = 1, \ldots, R$ distinct sources of data that are all imperfect representations of a single, shared occurrence process. Let $y_{r,a,k}$ be the observed detection (1) or nondetection (0) of a species of interest in data set r at site a during replicate k. Because different data sources have different variables influencing the observation process, we envision a separate detection model for each data source that is conditional on a single, shared ecological process described by Equation (38). We envision the detection-nondetection data from source r as arising from a Bernoulli process conditional on the true latent occurrence process:

$$\begin{aligned} y_{r,a,k} &\sim \text{Bernoulli}(p_{r,a,k} z_{j[a]}), \\ \text{logit}(p_{r,a,k}) &= \boldsymbol{v}_{r,a,k}'\boldsymbol{\alpha}_r, \end{aligned} \tag{39}$$

where $p_{r,a,k}$ is the probability of detecting a species at site a during replicate k (given it is present at site a) for data source r, which is a function of site, replicate, and data source specific covariates $\boldsymbol{V}_r$ and a vector of regression coefficients specific to each data source ($\boldsymbol{\alpha}_r$). Note that $z_{j[a]}$ is the true occurrence status at site j corresponding to the ath data source site in the given data set r. Each data source may be available at all J sites in the region of interest or at a subset of the J sites. Additionally, data sources can overlap in the sites they sample, or they can be obtained at distinct sites within all J sites of interest in the overall region.

We assume multivariate normal priors for the occurrence ($\boldsymbol{\beta}$) and data-set specific detection ($\boldsymbol{\alpha}$) regression coefficients to complete the Bayesian specification of a single species occupancy model. Polya-Gamma data augmentation is implemented analgoous to previous models, where there is a single set of occurrence auxiliary variables ($\boldsymbol{\omega}_\beta$) and a distinct set of detection auxiliary variables for each data source ($\boldsymbol{\omega}_{r,\alpha}$).

In short, the integrated occupancy model has an identical process model to the single species occupancy model, and has a distinct detection model for each data source that are all conditional on the same shared ecological process (species occurrence). Our full joint posterior takes the same form as that of a single species occupancy model, except a separate conditional likelihood is specified for each data source which is dependent on its own unique set of detection regression cofficients and Polya-Gamma auxiliary variables.

6.2 MCMC sampler

The Polya-Gamma data augmentation induces a Gibbs update for all parameters in the single species integrated occupancy model. We first sample the occurrence and detection auxiliary variables from

$$\begin{aligned} \omega_{j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j'\boldsymbol{\beta}), \\ \omega_{r,j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{r,j,k}'\boldsymbol{\alpha}_r), \end{aligned} \tag{40}$$

The occurrence regression coefficients are sampled from the same full conditional as that in the single species occupancy model in Equation (5).

The detection regression coefficients for a given data source r follows Equation (6), with all parameters now indexed by r.

Finally, z_j is set to 1 for all sites where there is at least one detection from one more more data sources, and thus we only need to sample z_j at sites where there are no detections. Thus, for all locations with no detections, we sample z_j according to

$$z_j \mid \cdot \sim \text{Bernoulli}\left(\frac{\psi_j \prod_{\forall a=j}(1 - p_{r,a,j,k})}{1 - \psi_j + \psi_j \prod_{\forall a=j}(1 - p_{r,a,j,k})}\right), \tag{41}$$

where the product occurs over all the sites in the R data sources that correspond the jth location.

6.3 Prediction

Integrated occupancy models have an identical ecological process model to single species occupancy models, and so out-of-sample prediction follows the same approach. See Section 2.3 for details.

7 Single species integrated spatial occupancy model (`spIntPGOcc`)

Single species integrated spatial occupancy models are identical to integrated occupancy models except the ecological process model now incorporates a spatially-structured random effect following the discussion in Section 3. All details for the single species integrated spatial occupancy model have already been presented. Here we present the sections to consult for necessary details for each portion of the single species integrated spatial occupancy model.

7.1 Gaussian Process formulation

7.1.1 Model description

For the ecological process model, see Section 3.1.1. For the observation model for each data source, see Section 6.1.

7.1.2 MCMC sampler

For the ecological process parameters, see Section 3.1.2. For the observation process parameters, see Section 6.2.

7.1.3 Prediction

See Section 3.1.3.

7.2 Nearest Neighbor Gaussian Process formulation

7.2.1 Model description

For the ecological process model, see Section 3.2.1. For the observation model for each data source, see Section 6.1.

7.2.2 MCMC sampler

For the ecological process parameters, see Section 3.2.2. For the observation process parameters, see Section 6.2.

7.2.3 Prediction

See Section 3.2.3.

References

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2003. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman; Hall/CRC.

Clark, Allan E, and Res Altwegg. 2019. "Efficient Bayesian Analysis of Occupancy Models with Logit Link Functions." *Ecology and Evolution* 9 (2): 756–68.

Datta, Abhirup, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. 2016. "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets." *Journal of the American Statistical Association* 111 (514): 800–812.

Finley, Andrew O, Abhirup Datta, and Sudipto Banerjee. 2020. "SpNNGP R Package for Nearest Neighbor Gaussian Process Models." *arXiv Preprint arXiv:2001.09111*.

Finley, Andrew O, Abhirup Datta, Bruce D Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. 2019. "Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes." *Journal of Computational and Graphical Statistics* 28 (2): 401–14.

Gelfand, Alan E, Alexandra M Schmidt, Sudipto Banerjee, and CF Sirmans. 2004. "Nonstationary Multivariate Process Modeling Through Spatially Varying Coregionalization." *Test* 13 (2): 263–312.

Heaton, Matthew J, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, et al. 2019. "A Case Study Competition Among Methods for Analyzing Large Spatial Data." *Journal of Agricultural, Biological and Environmental Statistics* 24 (3): 398–425.

Polson, Nicholas G, James G Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–49.

Robert, Christian, and George Casella. 2013. *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Roberts, Gareth O, and Jeffrey S Rosenthal. 2009. "Examples of Adaptive Mcmc." *Journal of Computational and Graphical Statistics* 18 (2): 349–67.

Taylor-Rodriguez, Daniel, Andrew O Finley, Abhirup Datta, Chad Babcock, Hans-Erik Andersen, Bruce D Cook, Douglas C Morton, and Sudipto Banerjee. 2019. "Spatial Factor Models for High-Dimensional and Large Spatial Data: An Application in Forest Variable Mapping." *Statistica Sinica* 29: 1155.

Tobler, Mathias W, Marc Kéry, Francis KC Hui, Gurutzeta Guillera-Arroita, Peter Knaus, and Thomas Sattler. 2019. "Joint Species Distribution Models with Species Correlations and Imperfect Detection." *Ecology* 100 (8): e02754.

Wright, Wilson J, Kathryn M Irvine, Thomas J Rodhouse, and Andrea R Litt. 2021. "Spatial Gaussian Processes Improve Multi-Species Occupancy Models When Range Boundaries Are Uncertain and Nonoverlapping." *Ecology and Evolution*.