

multiocc: An R Package for Spatio-Temporal Occupancy Models for Multiple Species

by Staci Hepler and Robert Erhardt

Abstract Spatio-temporal occupancy models are used to model the presence or absence of a species at particular locations and times, while accounting for dependence in both space and time. Multivariate extensions can be used to simultaneously model multiple species, which introduces another dimension to the dependence structure in the data. In this paper we introduce `multiocc`, an R package for fitting multivariate spatio-temporal occupancy models. We demonstrate the use of this package fitting the multi-species spatio-temporal occupancy model to data on six species of birds from the Swiss MHB Breeding Bird Survey.

1 Introduction

Occupancy models are commonly used in ecological applications to model presence or absence of a species of interest across a geographical region over a period of study, or *season*. The goal is often to determine the likelihood that a species was present, and to study how that likelihood relates to environmental features. Occupancy models account for imperfect detection in a binary response variable Y by treating the true presence as a latent binary process Z . These methods have been used across a variety of applications including estimating the geographical distribution of species such as caribou in northern Ontario, Canada (Johnson et al., 2013) and of red fox and coyote in the eastern United States (Rota et al., 2016). Rahman et al. (2021) and Guillera-Arroita et al. (2019) used multispecies occupancy models to estimate the richness of mammal species in Bangladesh and breeding birds in Switzerland, respectively. Recently, occupancy models have been utilized in public health applications to estimate the risk or prevalence of imperfectly detected diseases such as Histoplasmosis (Hepler et al., 2022) and SARS-CoV-2 (Sanderlin et al., 2021).

A number of R packages exist for fitting specific types of occupancy models. Here, we highlight a few of the existing packages, and refer the reader to Doser et al. (2022) for a more comprehensive list. The package `unmarked` fits a variety of models using likelihood-based methods; however these models do not account for spatial dependence in the data. `stocc` fits single-species spatial occupancy models in the Bayesian framework (Johnson et al., 2013). Recent attention on joint species distribution models have highlighted the importance of jointly analyzing species that may be related. Recently, `spOccupancy` was developed to fit single-species and multi-species spatial occupancy models (Doser et al., 2022). However, the multi-species model implemented in `spOccupancy` accounts for cross-species dependence by assuming regression coefficients for different species arise from a shared community prior. While this model is reasonable when studying groups of species that are expected to respond to the environment similarly, it might not be suitable to all applications. Additionally, this package does not apply to multi-season occupancy models. `gjam` can analyze multivariate presence/absence data and assumes an unstructured covariance matrix to capture species dependence; however, it does not account for any spatio-temporal dependence (Clark et al., 2017). `Hmsc` includes a variety of joint species distribution models for analyzing presence/absence data of species within a community (Tikhonov et al., 2020; Ovaskainen et al., 2017). However, it does not account for imperfect detection in the binary observations.

In this paper, we present `multiocc`, an R package that implements the multivariate spatio-temporal occupancy model developed by Hepler and Erhardt (2021). This model can be used to analyze presence/absence data for $S \geq 1$ species across $T \geq 1$ seasons. Briefly, this model accounts for dependence across space, across seasons, and also between the species being jointly modeled. Thus, this R package overcomes limitations of the existing packages mentioned above and can be applied to more general study designs. As noted by Taylor-Rodriguez et al. (2017), modeling species independently ignores residual dependence between species and can yield misleading results. Guisan and Rahbek (2011) found independently modeling species predicted too many species per location, and Clark et al. (2014) found that prediction was improved by using a multivariate approach that exploits information from other species. Pollock et al. (2014) proposed a multivariate model for presence/absence data that captures residual correlation through a covariance matrix, Σ , and applied their method to data on frog and eucalyptus species, but the model does not account for spatio-temporal dependence or imperfect detection. The multivariate, spatio-temporal occupancy model of Hepler and Erhardt (2021) that is implemented in this package similarly accounts for residual species correlation through a matrix, Σ ,

but also accounts for spatio-temporal dependence and imperfect detection. In addition, this model has been shown to yield more accurate estimates of occupancy model parameters as compared to single-species models.

Additionally, this model and package can be used in the single-survey setting for occupancy models provided the number of seasons is “large enough”, regardless of the number of species of interest. This has important implications for study design, since previous research has stated that parameter estimation in standard occupancy models requires multiple surveys for each period of time where the occupancy status is unchanged (MacKenzie et al., 2002). The assumption of an unchanging occupancy field across surveys can be unreasonable for many applications. Further, inference can be biased if this assumption is violated, but standard occupancy models are often fit regardless (Rota et al., 2009). The model implemented here in **multiocc** allows researchers to assume that multiple, dependent seasons are observed with even just a single survey per season. This is less restrictive, as occupancy status does not need to remain static from one observation to the next, and inference is reliable provided there are enough seasons.

In the next section, we review the statistical details of the model and discuss the study designs under which this model is recommended. Then, we describe the core functions of the **multiocc** package. Lastly, we illustrate the implementation of the package’s functionality by fitting the model to data on six species of birds from the Swiss MHB Breeding Bird Survey.

2 Scientific Background

This package implements the multivariate spatio-temporal occupancy model proposed in Hepler and Erhardt (2021), which is briefly summarized here. The aim is to determine the likelihood that species s was *present* at location i during time period t , and how environmental features relate to the likelihood. However, imperfect detection implies that the observed data only confirms *detection*, and not necessarily true *presence*. Occupancy models differentiate between observed binary detection and latent true occupancy. Let $Z_{it}^{(s)}$ denote the true occupancy ($Z_{it}^{(s)} = 1$ indicates presence and $Z_{it}^{(s)} = 0$ absence) of species $s = 1, \dots, S$ at location $i \in I_t$ during season $t = 1, \dots, T$, where I_t denotes the set of locations of interest during season t with cardinality $|I_t| = n_t$. A *season* is defined to be a time period over which the true occupancy is unchanged. One feature of this R package is the ability to incorporate multiple independent observations, or *surveys* per season, but further to allow different number of surveys between sites.

Let V_{it} be the number of surveys of the occupancy field at location i during a season t . Let the observed binary detection be denoted with Y such that $Y_{itv}^{(s)} = 1$ indicates species s was detected at location i during the v th survey ($v = 1, \dots, V_{it}$) in season t and $Y_{itv}^{(s)} = 0$ indicates it was not detected. The proposed occupancy model assumes there are no false positives, so $Z_{it}^{(s)} = 0$ implies $Y_{itv}^{(s)} = 0$ for all v ; if the species is actually present ($Z_{it}^{(s)} = 1$), it may be detected during a survey ($Y_{itv}^{(s)} = 1$) or not ($Y_{itv}^{(s)} = 0$).

Royle and Dorazio (2008) introduced the hierarchical model

$$\begin{aligned} [Y_{itv}^{(s)} | Z_{it}^{(s)}, p_{itv}^{(s)}] &= \text{Bernoulli} \left(Z_{it}^{(s)} p_{itv}^{(s)} \right) \\ [Z_{it}^{(s)} | \psi_{it}^{(s)}] &= \text{Bernoulli} \left(\psi_{it}^{(s)} \right), \end{aligned} \quad (1)$$

where $\psi_{it}^{(s)}$ denotes the probability of occupancy for species s at location i during time t and $p_{itv}^{(s)}$ the probability of detection given occupancy during survey v . Hepler and Erhardt (2021) specified the detection probability as

$$p_{itv}^{(s)} = \Phi \left(\mathbf{W}_{itv}' \boldsymbol{\beta}^{(s)} \right), \quad (2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, \mathbf{W}_{itv} is a p_D -dimensional vector of covariates related to detectability, and $\boldsymbol{\beta}^{(s)}$ is a vector of regression coefficients. Note that the vector of covariates is assumed to be the same for all S species, but the regression coefficients are species-specific. For occupancy, Hepler and Erhardt (2021) specified the probability as

$$\psi_{it}^{(s)} = \Phi \left(\mathbf{X}_{it}' \boldsymbol{\alpha}^{(s)} + \eta_{it}^{(s)} \right), \quad (3)$$

where \mathbf{X}_{it} is a p_O -dimensional vector of covariates related to occupancy and is assumed to be the same for all species, $\boldsymbol{\alpha}^{(s)}$ is a species-specific vector of regression coefficients, and $\eta_{it}^{(s)}$ is a multivariate spatio-temporal random effect.

To reduce dimensionality and alleviate confounding with fixed effects, we use a restricted spatial regression model as in [Hughes and Haran \(2013\)](#) and [Bradley et al. \(2015\)](#). More specifically, if $\eta_t^{(s)} \equiv \{\eta_{it}^{(s)}, i \in I_t\}$ is the vector of random effects for species s and time t , then we assume $\eta_t^{(s)} = \mathbf{K}_t \gamma_t^{(s)}$, where \mathbf{K}_t is a $n_t \times q$ matrix of known spatial basis functions such that $\mathbf{K}_t^* \mathbf{K}_t = \mathbf{I}_q$ and $\gamma_t^{(s)}$ is a q -dimensional multivariate random effect. Letting $\eta_t \equiv (\eta_t^{(1)}, \dots, \eta_t^{(S)})$ be the $n_t S$ -dimensional vector of random effects, we have $\eta_t = \mathbf{K}_t^* \gamma_t$, where $\mathbf{K}_t^* = \mathbf{I}_S \otimes \mathbf{K}_t$ is the $n_t S \times qS$ block diagonal matrix whose (s, s) block is \mathbf{K}_t , and $\gamma_t = (\gamma_t^{(1)}, \dots, \gamma_t^{(S)})$. An intrinsic multivariate conditional autoregressive (MCAR) structure is used to capture multivariate and spatial dependence, with temporal dependence captured through an autoregressive model of order 1 ([Mardia, 1988](#)). More specifically, this dependence framework is captured within the restricted spatial regression model by assuming

$$\gamma_t \sim \begin{cases} N(\mathbf{0}, \mathbf{K}_t^* \mathbf{Q}_t \mathbf{K}_t^*) & \text{for } t = 1 \\ N(\mathbf{M} \gamma_{t-1}, \mathbf{K}_t^* \mathbf{Q}_t \mathbf{K}_t^*) & \text{for } t = 2, \dots, T, \end{cases} \quad (4)$$

where $\mathbf{Q}_t = \Sigma^{-1} \otimes (\mathbf{D}_t - \mathbf{A}_t)$ is the intrinsic MCAR precision matrix, Σ is an $S \times S$ matrix that captures conditional dependence (positive or negative) among the species, \mathbf{A}_t is an $n_t \times n_t$ adjacency matrix, \mathbf{D}_t is a diagonal matrix whose (i, i) th element is the number of neighbors of location i , and $\mathbf{M} = \text{diag}(\rho_1, \dots, \rho_S) \otimes \mathbf{I}_q$ is the propagator matrix such that ρ_s accounts for temporal autocorrelation for species s . Following the work of [Bradley et al. \(2015\)](#), [Hepler and Erhardt \(2021\)](#) used spatio-temporal Moran's I (MI) basis functions for \mathbf{K}_t . More specifically, the MI operator for time t is defined as

$$\mathcal{I}(\mathbf{X}_t, \mathbf{A}_t) = \left(\mathbf{I}_{n_t} - \mathbf{X}_t (\mathbf{X}_t' \mathbf{X}_t)^{-1} \mathbf{X}_t' \right) \mathbf{A}_t \left(\mathbf{I}_{n_t} - \mathbf{X}_t (\mathbf{X}_t' \mathbf{X}_t)^{-1} \mathbf{X}_t' \right). \quad (5)$$

The q columns in \mathbf{K}_t are chosen to be the q eigenvectors of $\mathcal{I}(\mathbf{X}_t, \mathbf{A}_t)$ that correspond to the largest eigenvalues.

Since the model is fit within the Bayesian paradigm, prior distributions are chosen for all remaining hyperparameters. The prior distributions for the regression coefficients α and β are improper and uniform over the real line. The temporal autocorrelation parameters ρ_1, \dots, ρ_S are assumed to be uniform over $(0, 1)$. The cross-species covariance matrix Σ has an inverse Wishart prior distribution with $S + 2$ degrees of freedom and identity scale matrix. To assist in computation, the data augmentation strategy of [Albert and Chib \(1993\)](#) is implemented in the Markov chain Monte Carlo (MCMC) algorithm. The model specification results in all known full conditional distributions, and thus a Gibbs sampling algorithm is used to simulate from the posterior distribution. Derivations of the full conditional distributions needed to implement the Gibbs sampling algorithm can be found in the supplementary material of [Hepler and Erhardt \(2021\)](#).

3 The multiocc package

The entry point for the package is the function `multiocbuild(detection = ..., occupancy = ..., coords = ..., DataNames = ..., threshold = ...)`. This function accepts raw data as inputs, and it outputs all ordered matrices and the adjacency information needed to run the MCMC algorithm and sample from the posterior distribution. This is intentionally split off as a separate function, rather than written internally as a precursor step in the MCMC algorithm. This structure allows the user to specify subsets of covariates, seasons, surveys, and/or locations for different model runs and receive informative errors (or confirmation) that the specification will lead to a valid model fit. The single output list produced by `multiocbuild()` contains an internally consistent set of ordered data frames as well as species and covariates names used, and helps the user avoid confusion by collecting all model inputs into a single list rather than requiring multiple specifications across objects which may be internally inconsistent. The five required arguments are:

- 'detection' is a data frame with one row for each site/season/survey combination for a total of $\sum_{t=1}^T \sum_{i \in I_t} V_{it}$ rows. The first three columns should identify the 'site', 'season', and 'survey', and be named as such. Site can be a character vector. For each of the S species there is a column of binary detection (1=yes, 0=no) for that particular species during the site/season/survey combination. Titles of these columns are chosen by the user, but are commonly the names of the S species. Next are the set of possible detection covariates used in \mathbf{W} . Titles of these are chosen by the user, and there is no need to include an intercept column in this data frame as one is automatically added by `multiocbuild()`. It is strongly encouraged that the quantitative explanatory variables be standardized to assist with chain mixing. Not every variable included in detection will necessarily be included in the model, as the `DataNames` argument described below is what defines the covariates to be used in \mathbf{W} .

- ‘occupancy’ is a data frame with one row for each site/season combination (note there is no survey for occupancy, as the species either is or is not present for all surveys). The number of rows is therefore $\sum_{t=1}^T n_t$, where T is the number of seasons, and n_t is the number of locations of interest (either observed or where predictions are desired) during season t . The first two columns are ‘site’ and ‘season’ and must be named as such. Site can be a character vector but must be consistent with how ‘site’ is specified in detection. Next are the set of possible occupancy covariates used in \mathbf{X} . Titles of these are chosen by the user, and there is no need to include an intercept column in this data frame as one is automatically added by `multioccbuild()`. It is strongly encouraged that the quantitative explanatory variables be standardized to assist with chain mixing. Not every variable included in occupancy will necessarily be included in the model, as the `DataNames` argument described below is what defines the covariates to be used in \mathbf{X} .
- ‘coords’ is a data frame with one column for the site name — which must be consistent with the naming of ‘site’ in occupancy and detection — and two columns for the numeric coordinates of the site location. Column names must be ‘site’, ‘x’, and ‘y’. The number of rows is equal to the number of unique locations in the study. Units are arbitrary, but must be the same in both location columns ‘x’ and ‘y’ for a valid Euclidean distance measure.
- ‘DataNames’ is a list with three elements, always titled ‘species’, ‘detection’, and ‘occupancy’. Each element is a vector specifying the precise titles of the S species, and the names of the covariates for detection and occupancy. The names in these vectors must correspond to the names of the corresponding columns in the detection and occupancy data frames. An intercept is added by default. `DataNames` is required because it allows the user to model a subset of species and/or a subset of all variables in the objects occupancy and detection, and therefore it is very easy to run multiple versions of the model with distinct covariates on the same data. If the user desires to have no covariates for detection, they can either write `detection=c()` or leave the argument `detection` out as an input. Similarly, if the user does not want any covariates to be included in occupancy, they can either write `occupancy=c()` or leave the argument `occupancy` out as an input.
- ‘threshold’ is a number which determines whether or not two sites are considered neighbors in the adjacency structure of the model. If the (Euclidean) distance between two sites is less than threshold, they are neighbors; if not, they are not neighbors. This is used to identify the non-zero entries in the adjacency matrices, A_t .

A few points of clarification are needed. First, the data frame occupancy should *only* have columns named ‘site’, ‘season’, and names of covariates which may appear in ‘DataNames\$occupancy’. The data frame ‘detection’ should *only* have columns named ‘site’, ‘season’, ‘survey’, names for species detection data which also appear in ‘DataNames\$species’, and names of detection covariates which may appear in ‘DataNames\$detection’. The ordering of ‘DataNames\$species’ will determine the ordering of all results shown by species. Internal checks in `multioccbuild()` include a scan for duplicated site/season combinations in occupancy, or site/season/survey combinations in detection. The presence of either type of duplication halts the function. The function also scans to ensure all entries for detections y are 0, 1, or NA. Also, if there are any site/season combinations in detection but not occupancy, they are removed from detection with a warning shown, but the function still runs. If there are any site/season combinations in occupancy but not detection, corresponding rows are added to detection with NAs for all variables, which enables prediction of occupancy for those site/season combinations.

There are also internal checks to identify missing covariates in either \mathbf{X} or \mathbf{W} . There are two cases of missing covariates which our model cannot accommodate. The first is any missingness in \mathbf{X} . The second is when y is observed (as either 0 or 1), but covariates for detection \mathbf{W} for the corresponding site/season/survey are missing. Both situations lead to a deletion of rows. If there are missing covariates in \mathbf{X} , then the row(s) are removed from \mathbf{X} and the V_{it} corresponding rows for the same site/season combination are removed from \mathbf{W} . If there are missing covariates for any site/season/survey combination of \mathbf{W} for cases with non-missing y , then the value of y is changed to NA which effectively removes the detection from the observed data but retains predictions of occupancy for the corresponding site/season combination.

It is possible to use this package to predict occupancy for site/season combinations where no observations were taken (meaning there are either no rows in ‘detection’ for that site/season, or the rows in ‘detection’ have y with NA), provided the values of the occupancy covariates are available. The site/season combinations of interest just need to be included as rows in the ‘occupancy’ and ‘coords’ data frames that are provided as input to `multioccbuild()`. The site/season combinations that are in ‘occupancy’ but not ‘detection’ are identified, and the Gibbs Sampling algorithm will simulate values of z from the posterior predictive distribution. We note that in other packages, prediction is a two-step procedure where the site/seasons to be predicted are provided after the model

fit. That is not the case here because computation of the Moran's I basis functions requires the full set of site/seasons at all locations - both observed and unobserved (Bradley et al., 2015).

From a modeling perspective, the primary decisions the researcher needs to make when executing the `multioccbuild()` function are: (1) Which specific covariates belong in \mathbf{X} and \mathbf{W} to model occupancy and detection? (2) Which species should be included? and (3) What threshold should be used to determine the definition of neighbor in the adjacency matrix? This R package does not restrict the type or number of covariates included in \mathbf{X} and \mathbf{W} , but it is worth noting the literature includes discussions of possible difficulty identifying regression coefficients when there is only a single survey per season. In the initial versions of occupancy models, MacKenzie et al. (2002) advised at least two surveys for a single season when the occupancy rate and detection probability are sufficiently large (0.7 and 0.3, respectively), and more surveys when these probabilities are smaller. Lele et al. (2012) investigated the single-survey scenario further and found that occupancy and detection probabilities can be estimated with a single survey when the number of spatial locations is large, provided they depend on covariates with at least one numeric covariate in detection and the two sets of covariates differ by at least one variable. Hepler et al. (2018) extended earlier work on occupancy models and proposed a spatio-temporal occupancy model for the multiple-season scenario and found the model could identify covariate effects in the single survey per season case, provided the number of seasons was large enough. For the multivariate spatio-temporal model implemented in `multiocc`, Hepler and Erhardt (2021) found the model could identify covariate effects in the single-survey case with at least 10-20 seasons. Careful investigation of the posterior distributions for dependence across the two parameters would be one way to investigate confounding and identifiability. Note that interactions between two or more variables may also be used and defined in the usual way.

The definition of the threshold governs the structure of the adjacency matrix A . If two sites are within this distance, then they are neighbors, and the corresponding element of the adjacency matrix is 1. When the data are collected on a regular grid, it is common to use either a rook (shared edge) or queen (shared edge or vertex) relationship to determine which grid points are neighbors. These neighborhood specifications can be implemented in this package by determining the distance between grid centroids and setting the threshold to be larger than that distance. In general, lower values of 'threshold' result in fewer neighbors and a sparser matrix A , while larger values result in more neighbors but a less sparse matrix A . There are computational advantages to a sparse adjacency matrix, and a low threshold can mimic the first-order Markov dependence commonly used when modeling aerial data (Gelfand et al., 2010).

The output of `multioccbuild()` is a list called 'model.input'. This list contains the following elements:

- 'DataNames', the precise list with three elements, always titled 'species', 'detection', and 'occupancy' that the user specified for the `multioccbuild()` function. These are used downstream in the package.
- 'X', a matrix which serves as the design matrix for the occupancy portion of the model. The first column is a column of all 1s, and latter columns are from 'names\$occupancy'. The matrix has $\sum_{t=1}^T n_t$ rows - one for each site/season combination.
- 'W', a matrix which serves as the design matrix for the detection portion of the model. The first column is a column of all 1s, and latter columns are from 'names\$detection'. The matrix has $\sum_{t=1}^T \sum_{i \in I_t} V_{it}$ rows - one for each site/season/survey combination.
- 'y', a $\sum_{t=1}^T \sum_{i \in I_t} V_{it} \times S$ matrix of the binary values of detection for the S species. The columns are in the same order that the species are listed in in the vector 'DataNames\$species'.
- 'A', the symmetric adjacency matrix of 1s and 0s indicating which rows of 'occupancy' are neighbors. 'A' is a block diagonal matrix, where the t th block is the $n_t \times n_t$ dimensional matrix A_t .
- 'detection.info' is a data frame with the detection information containing the five columns 'siteID' (a numeric index of the site), 'site' (the factor identifier of the site), 'season', 'survey', and an indicator variable named 'observations' that indicates whether the survey result was missing (indicated by an NA in the 'y' matrix) or not.
- 'occupancy.info' is a data frame with 'siteID', 'site', and 'season' columns as well as columns for the x/y coordinates merged from 'coords'.

The function `GibbsSampler(M.iter, M.burn, M.thin, model.input, q, sv, every, WAIC, param2keep)` performs the Gibbs sampling MCMC algorithm to sample from the posterior distribution. This function is written entirely in R.

- 'M.iter' is a required input and is the number of iterations in the MCMC.
- 'M.burn' is the desired length of the burn in. Default is half of 'M.iter'.

- 'M. thin' is the desired thinning of the MCMC output. M. thin=1 is no thinning, M. thin=10 saves every tenth, etc. Default is 1.
- 'model.input' is the output from the `multioccbuild()` function described above.
- 'q' is the number of Moran's I spatial basis functions used in the restricted spatial regression model. Default is 10 percent of the minimum number of locations in any season (rounded).
- 'sv' is a logical indicator controlling whether or not the chain should be saved at regular iterations or not. The number of iterations is set by 'every'. Default for 'sv' is FALSE.
- 'every' a numeric value controlling the frequency that the MCMC chain results are saved. Default is 1000.
- 'WAIC' is a logical value indicating whether or not the MCMC should compute and save WAIC. Defaults to false. Note that computing WAIC increases the computational storage required to fit this model.
- 'param2keep' is a character vector that governs which outputs are saved. Permissible entries in this vector are 'alpha', 'beta', 'gamma', 'rho', 'sigma', 'psi', 'z', 'p', and 'loglik'. The default is to save 'alpha', 'beta', 'gamma', 'rho', 'sigma', and 'psi'.

A number of comments are needed to discuss issues with computational cost and computer memory. First, including more parameters in 'param2keep' increases the amount of storage required. In particular, saving the samples of 'z', 'psi', 'p', and 'loglik' can create storage issues depending on the amount of available RAM. This can be avoided by not saving those quantities as they can easily be computed based on their definitions provided 'alpha', 'beta', and 'gamma' are saved. For example, the detection probabilities 'p' can be computed for each stored value of β as $\text{probit}(\mathbf{W}\beta)$. However, it is also possible to store these outputs if memory is not a concern. If the user specifies `WAIC=TRUE`, 'loglik' will be created and computed but not stored by default; the user would additionally need to add 'loglik' to 'param2keep' to store these values. This is a higher cost to memory, but allows the user to perform other types of model assessment, e.g. leave-one-out cross-validation as in [Vehtari et al. \(2017\)](#).

The output of `GibbsSampler()` is a list. Specifically, it contains the following elements.

- 'samples' is a list with one element for each of the items specified in 'param2keep'. Each individual element of this list is named according to 'param2keep' and is a valid MCMC object defined from the `coda` package, and contains the saved samples of that object from the MCMC.
- 'run.time' saves the run time for the `GibbsSampler()` command.
- 'WAIC' is a vector of length S which contains WAIC values by species, if this argument was true.
- 'basis.K' is an $\sum_{t=1}^T n_t \times q \cdot T$ matrix of the basis functions created using the `make.basis()` function.

Additionally, the list has elements 'occupancy.info', 'detection.info', 'X', 'W', and 'y' which are the same quantities that were outputted by `multioccbuild()` and are saved as output here as well for convenience. The ordering of the output in 'psi' corresponds to the site/season ordering in 'occupancy.info'. Similarly, if 'p' and/or 'loglik' are included in 'param2keep' then the ordering of the values in 'samples' corresponds to the site/season/survey ordering in 'detection.info'.

There are several limitations of the model implemented in the current version of the **multiocc** package. First, the code requires the same set of covariates are used for all species being modeled. Second, the package currently does not allow the user to specify their own prior distributions or initial values for model parameters. Third, the package currently requires the use of Moran's I basis functions and assumes an intrinsic, conditional autoregressive structure where the neighborhood matrix is specific through a distance threshold. Fourth, while the use of basis functions makes the model scalable in the number of spatial locations, the computational expense could still be prohibitive if the number of species, S , or the number of seasons, T are very large. It also remains unclear how large S can be before computational and identifiability issues arise with estimating Σ . [Taylor-Rodriguez et al. \(2017\)](#) proposed a Dirichlet process approach to dimension reduction to overcome these issues in settings where the number of species is large ($\sim 10^3$). In a simulation study, they found their dimension reduced model outperformed a model using the full Σ when $S = 100$, although they noted that both models still performed well. However, their model did not include multiple seasons, which will improve estimation. It is important the user perform MCMC diagnostic checks to assess inferential stability for the elements of Σ .

4 Example

The data are taken from the Swiss MHB ("Monitoring Häufige Brutvögel") Breeding Bird Survey from the Swiss Ornithological Institute. This survey covers common bird species in Switzerland

and first began in 1999. Specifically, we consider the presence or absence of six types of Swiss Tits obtained at 267 locations over a ten year period (2004 - 2013). Each year is one season in our model. These data are freely available as ‘SwissTits’ in the R package [AHMbook](#) (Kéry et al., 2017), and are stored in our package [multiocc](#) as ‘detection’, ‘occupancy’, and ‘coords’, which can be called with `data(detection)`, `data(occupancy)`, `data(coords)`. The locations ‘coords’ are 1km by 1km quadrats roughly evenly spread across the entire country of Switzerland. Locations are shown in Figure 1.

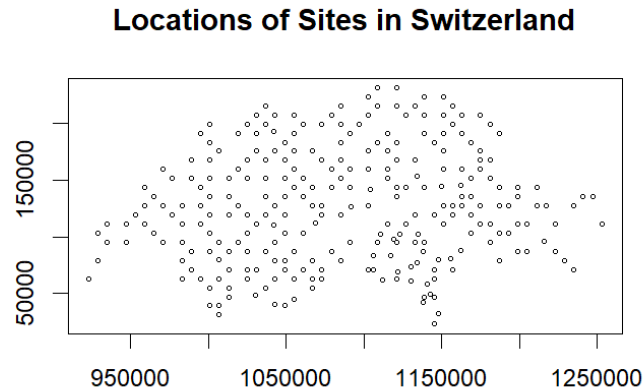


Figure 1: Locations of the 267 1km by 1km square quadrats laid out across the entirety of Switzerland in the MHB survey. In general, the spacing of locations supports the use of an areal data model with neighbors defined through an adjacency matrix A .

The response is the observed detection ($Y_{itv}^{(s)} = 1$) or not ($Y_{itv}^{(s)} = 0$) of species s at location i , time period t and survey v . There were 266 to 267 sites observed in each season, with the number of surveys per season varying from 2 to 3. Covariates on occupancy in X include the standardized percentage of forest cover at the location as well as the standardized elevation. The latent variable presence $Z_{it}^{(s)} = 1$ indicates that species s was present at location i and time period t , with detection probability $p_{it}^{(s)}$ modeled with standardized covariate duration, which measures the length of the particular survey. Histograms of the values for the covariates are shown in Figure 2. These same covariates are plotted across space in Figure 3. The left column shows raw data, and the right column shows interpolated covariates using a thin plate spline for ease of visualization only.

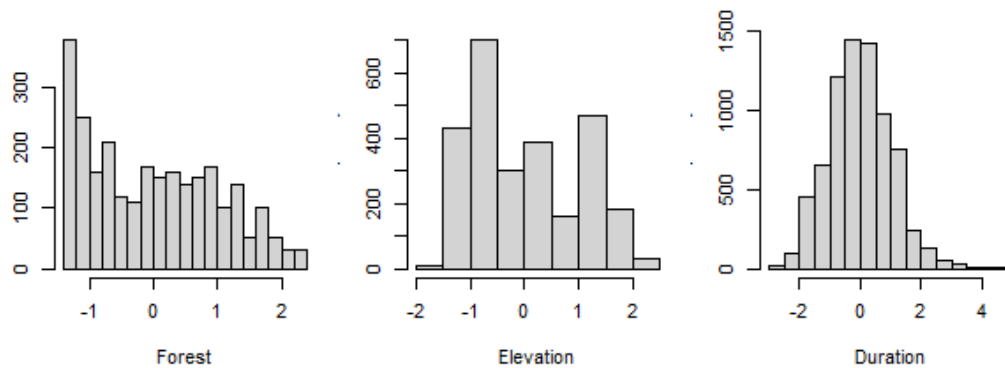


Figure 2: Histograms of the three covariates. Left: Percent forest cover across the 267 locations. Center: Elevation of the 267 sites. Right: Durations of the unique surveys. Each covariate is standardized and therefore unitless on this scale.

To define the neighborhood structure used for the spatial random effect, we selected a threshold of 15000 (meters, as these are UTM coordinates) and defined any two locations whose distance is less than 15000 to be neighbors. This threshold gave each location somewhere between 1 and 8 neighbors, with an average of 4.23 neighbors. We assume $q = 10$ spatial basis functions in each season.

The `multioccbuild(detection, occupancy, coords, DataNames, threshold)` function takes five arguments as input, and outputs ‘model.input’ which is then an argument for `GibbsSampler()` which implements the Gibbs Sampling MCMC algorithm to simulate from the posterior distribution. Examples of the inputs ‘occupancy’, ‘detection’, ‘coords’, ‘DataNames’ and ‘threshold’ for our application are shown below:

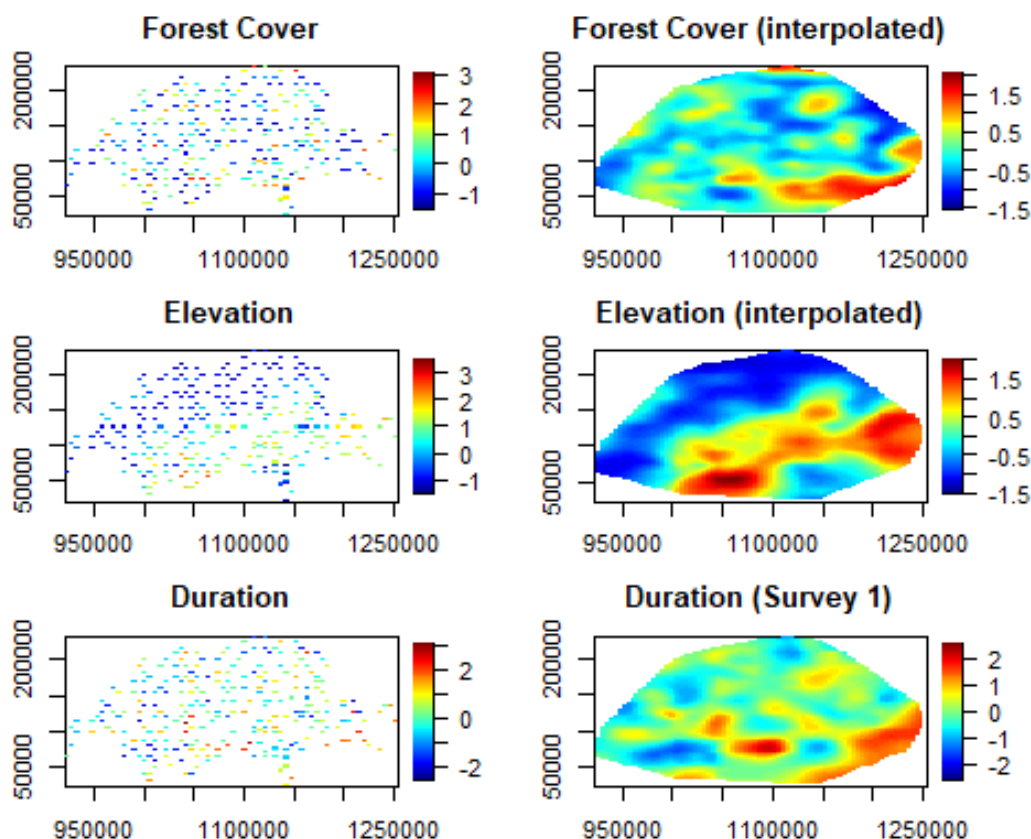


Figure 3: Spatial plots of the three covariates forest cover, elevation, and duration. The left column shows raw data, and the right column shows interpolated data using a thin plate spline for ease of visualization only. Forest and Duration are shown for season one, as these two covariates vary in time.

```
> head(occupancy, n=4L)
  site season    forest    elev
1 Q001      1 -1.148171612 -1.14938671
2 Q002      1 -0.496936229 -1.14938671
3 Q003      1 -0.098959051 -0.21468405
4 Q004      1 -0.931093151 -0.37046783

> head(detection, n=4L)
  site season survey Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit duration
1 Q001      1      1      NA      NA      NA      NA      NA      NA      0.1563255
2 Q002      1      1      1      1      0      0      1      0     -0.9274962
3 Q003      1      1      1      0      1      0      1      0     -0.2605290
4 Q004      1      1      NA      NA      NA      NA      NA      NA     -0.8441253
```

Observe that some surveys have NAs for all detections. Our model and package can easily accommodate missing values for the detections, but our model cannot handle missing values for occupancy covariates nor missing values for detection covariates when the corresponding 'y' are not missing. Missing values in the covariates will result in a warning when the `multioccbuild()` function is run, and rows corresponding to those observations in 'detection' and possibly in 'occupancy' will be removed. We suggest that all quantitative variables to be used as detection or occupancy covariates be standardized to aid with mixing in the MCMC algorithm. The remainder of the input items are:

```
> head(coords, n=4L)
  site    x    y
1 Q001 922942 63276
2 Q002 928942 79276
3 Q003 928942 103276
4 Q004 934942 95276

> DataNames = list("species"=colnames(detection)[4:9], "detection"=c("duration"),
```



```

"occupancy"=c("forest", "elev"))
> DataNames
$species
[1] "Great.tit" "Blue.tit" "Coal.tit" "Crested.tit" "Marsh.tit" "Willow.tit"

$detection
[1] "duration"

$occupancy
[1] "forest" "elev"

```

With each of these inputs properly defined, we run the main data function `multioccbuild()`:

```

> model.input = multioccbuild(detection, occupancy, coords, DataNames, threshold = 15000)
Warning: Rows in detection with missing covariates have been removed for purposes of
fitting the model, but the site/season combination is retained in occupancy and therefore
predictions will be outputted.

```

The above message arises from cases in which we have NAs for covariates in detection **W**, but we have actual observations of y for these same site/season/survey combinations. Our model cannot be fit using rows with missing detection covariates but non-missing detections in y . There are 544 NAs in the covariate, but only 4 of these cases have non-missing detections:

```

> sum(is.na(detection$duration))
[1] 544
> sum(!is.na(detection$Great.tit) & is.na(detection$duration))
[1] 4

```

One issue is site Q091, season 8:

```

> detection[detection$site == "Q091" & detection$season==8,]
  site season survey Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit duration
5698 Q091      8      1        0        0        1          1          0          1 -1.177609
5965 Q091      8      2        1        0        1          1          0          1 -1.094238
6232 Q091      8      3        0        0        1          1          0          1      NA

```

The second issue is site Q125, season 2:

```

> detection[detection$site=="Q125" & detection$season==2,]
  site season survey Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit duration
926  Q125      2      1        1        1        1          1          1          0      NA
1193 Q125      2      2        1        1        1          0          1          0      NA
1460 Q125      2      3        1        1        1          1          1          0      NA

```

The `multioccbuild()` function replaces the observed occupancy data y with NAs for these 4 cases. The output is 'model.input':

```

model.input = list("DataNames"=DataNames, "X"=X, "W"=W, "y"=y, "A"=A,
"detection.info"=detection[,c("siteID", "site", "season", "survey", "observations")],
"occupancy.info"=occupancy[,c("siteID", "site", "season", "x", "y")])

```

Objects saved here reflect handling the missing data for the 4 rows described above:

```

>model.input$y[(model.input$detection.info$site == "Q091" &
model.input$detection.info$season == 8),]
  Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit
5698      0      0        1          1          0          1
5965      1      0        1          1          0          1
6232     NA     NA      NA          NA      NA          NA

> model.input$y[(model.input$detection.info$site == "Q125" &
model.input$detection.info$season == 2),]
  Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit
926      NA     NA      NA          NA      NA          NA
1193      NA     NA      NA          NA      NA          NA
1460      NA     NA      NA          NA      NA          NA

```

The package reorders the data as site, then season, then survey, and ensures all matrices are in this order. The resulting design matrices for occupancy \mathbf{X} and detection \mathbf{W} are

```
> head(model.input$X, n=4L)
      forest      elev
1  1 -1.14817161 -1.1493867
11 1 -0.49693623 -1.1493867
21 1 -0.09895905 -0.2146840
32 1 -0.93109315 -0.3704678

> nrow(model.input$X)
[1] 2670

> head(model.input$W, n=4L)
      [,1]      [,2]
[1,] 1 0.1563255
[2,] 1 1.1567763
[3,] 1 1.1567763
[4,] 1 -0.9274962

> nrow(model.input$W)
[1] 8010
```

The threshold is the Euclidean distance which defines whether or not two locations are neighbors. In our, setting we set threshold = 15000 which results in the 2670×2670 block diagonal adjacency matrix

```
> model.input$A[1:5,1:5]
      1 11 21 32 41
1  0  0  0  0  0
11 0  0  0  0  0
21 0  0  0  1  1
32 0  0  1  0  0
41 0  0  1  0  0
```

Observe the detections y are reordered because all components of the data were reordered according to site, season, and survey:

```
> head(model.input$y, n=4L)
      Great.tit Blue.tit Coal.tit Crested.tit Marsh.tit Willow.tit
1           NA      NA      NA           NA      NA      NA
268         NA      NA      NA           NA      NA      NA
535         NA      NA      NA           NA      NA      NA
2           1       1       0           0       1       0

> head(model.input$detection.info, n=4L)
      siteID site season survey observations
1         1  Q001      1      1            0
268        1  Q001      1      2            0
535        1  Q001      1      3            0
2         2  Q002      1      1            1

> nrow(model.input$detection.info)
[1] 8010

> head(model.input$occupancy.info, n=4L)
      siteID site season      x      y
1         1  Q001      1 922942 63276
11        2  Q002      1 928942 79276
21        3  Q003      1 928942 103276
32        4  Q004      1 934942 95276

> nrow(model.input$occupancy.info)
[1] 2670
```

We ran our full analysis using the GibbsSampler ($M.iter = 50000, M.burn = 20000, M.thin = 10, model.input, q=10, sv=TRUE$). This took 4.03 hours to run on a single node on the Wake Forest University DEAC Cluster ([Information Systems and Wake Forest University, 2021](#)).

Results from the analysis are shown in Tables 1 and 2. It is important to note that we standardized all covariates to aid with MCMC mixing, and so we focus on the relative ordering of parameter estimates when comparing to other studies rather than absolute estimates. Table 1 shows the ordering of elevation's impact on occupancy as Willow Tit, Crested and Coal Tit, Great Tit, Blue Tit and Marsh Tit. This ordering largely matches the ordering published in [Tobler et al. \(2019\)](#) (Table S2), which considered the same data and six species in a somewhat similar Joint Species Distribution Model. [Chamberlain et al. \(2016\)](#) published that Willow Tits occupy higher elevations than Coal Tits in a nearby region of Italy, consistent with the ordering presented here. Turning to percent forest cover, our results show higher occupancy with higher forest cover for the Willow, Coal and Crested Tits, middling occupancy for Marsh Tit, and lower occupancy for the Great and Blue Tits, again matching the ordering of [Tobler et al. \(2019\)](#).

| Species | Intercept | Forest | Elevation |
|-------------|----------------------|-------------------------|-------------------------|
| Great Tit | 0.847 (0.769, 0.930) | -0.143 (-0.215, -0.072) | -0.224 (-0.300, -0.149) |
| Blue Tit | 0.639 (0.547, 0.729) | -0.139 (-0.221, -0.058) | -0.261 (-0.349, -0.177) |
| Coal Tit | 0.996 (0.919, 1.076) | -0.007 (-0.080, 0.066) | -0.130 (-0.204, -0.055) |
| Crested Tit | 0.764 (0.667, 0.867) | -0.014 (-0.097, 0.071) | -0.138 (-0.223, -0.052) |
| Marsh Tit | 0.611 (0.509, 0.724) | -0.124 (-0.219, -0.031) | -0.283 (-0.386, -0.182) |
| Willow Tit | 0.383 (0.267, 0.506) | 0.108 (0.013, 0.199) | 0.068 (-0.028, 0.163) |

Table 1: Posterior means and 95% credible intervals for the occupancy regression coefficients α .

| Species | Intercept | Duration |
|-------------|-------------------------|-------------------------|
| Great Tit | 0.593 (0.553, 0.633) | -0.187 (-0.228, -0.147) |
| Blue Tit | 0.109 (0.069, 0.149) | -0.257 (-0.299, -0.214) |
| Coal Tit | 0.738 (0.696, 0.779) | 0.303 (0.263, 0.346) |
| Crested Tit | -0.030 (-0.072, 0.011) | 0.352 (0.309, 0.395) |
| Marsh Tit | -0.307 (-0.351, -0.264) | -0.090 (-0.133, -0.048) |
| Willow Tit | -0.482 (-0.532, -0.434) | 0.430 (0.377, 0.482) |

Table 2: Posterior means and 95% credible intervals for the detection regression coefficients β

Turning next to the residual dependence structure, Figure 4 shows the posterior mean of the correlation matrix computed from Σ for the six species, with notable differences across pairs of species. We see a mixture of positive and negative correlations across different pairs of species, with, for instance, a strong positive signal for the Coal and Crested Tits but a strong negative correlation for the Willow and Marsh Tits. The positive correlation estimated here for Great tit and Blue tit is consistent with [Stenseth et al. \(2015\)](#), who noted these species occupy the same geographical areas and compete for resources, which can result in negative correlation for abundance, but positive correlation for occupancy. Our estimated correlations for the six species somewhat differ from those estimated by [Tobler et al. \(2019\)](#), who showed results on residual correlation of occupancy probability for the six tit species for a single season, but computed under a latent factor model. Their results showed all positive correlations among the six species, a result they noted with surprise. However, this is reflective of the factor model structure in their assumed model which is less flexible than the multivariate random effect we used here. They also assumed different covariates and only modeled a single season, and as both model's dependence structures are residual to covariate effects, the estimates are not readily comparable.

Figures 5 and 6 show posterior means for occupancy probabilities ψ for seasons 1 and 7 for all six species. Hollow circles show the actual detections of each species in each corresponding season. For all $T = 10$ years, we observe a very strong relationship between predicted occupancy probabilities ψ and species detections for all species and all seasons.

5 Conclusion

In this article, we introduced the R package **multiocc** for implementing the multivariate spatio-temporal occupancy model proposed by [Hepler and Erhardt \(2021\)](#). This model overcomes many of the limitations of occupancy models that can be implemented with already existing R packages. More

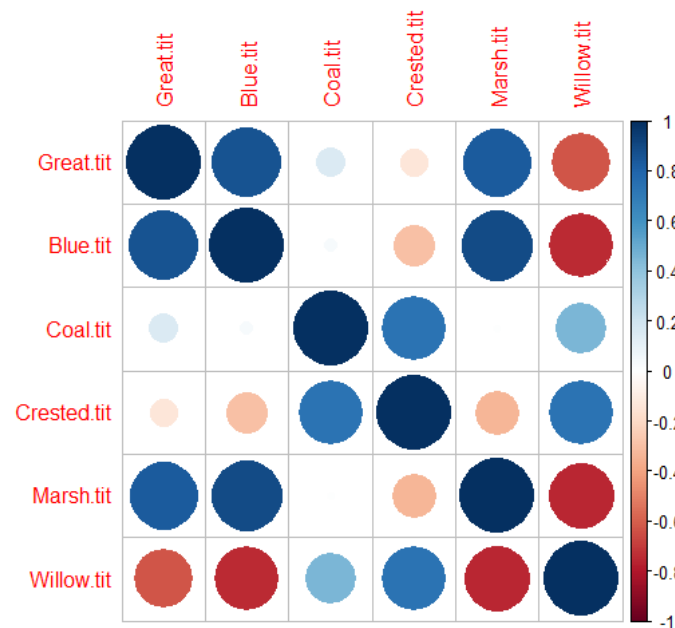


Figure 4: Posterior mean of the correlation matrix computed from Σ . Plot shows examples of pairs of strongly correlated species (e.g. Great Tit and Blue Tit), strongly negative correlated species (e.g. Marsh Tit and Willow Tit), and largely uncorrelated species (e.g. Marsh Tit and Coal Tit).

specifically, this model can jointly analyze $S \geq 1$ species over multiple seasons with a varying number of surveys at each site per season. The use of spatial basis functions makes this method feasible even with a large number of spatial locations.

Using this package to analyze imperfectly detected presence/absence data requires two main steps. First, the `multioccbuild()` function takes the raw data and outputs all matrices required to run the MCMC algorithm. Then, the `GibbsSampler()` function performs the Gibbs sampling algorithm to generate samples from the posterior distribution which can then be used to perform inference. We illustrated use of this package by analyzing occupancy data on six bird species in Switzerland from 2004 - 2013. The development of **multiocc** makes this multivariate spatio-temporal occupancy model accessible to the large community of researchers who use R for their data analysis needs.

References

- J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993. [p39]
- J. R. Bradley, S. H. Holan, and C. K. Wikle. Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics*, 9(4):1761–1791, 2015. [p39, 41]
- D. Chamberlain, M. Brambilla, E. Caprio, P. Pedrini, and A. Rolando. Alpine bird distributions along elevation gradients: the consistency of climate and habitat effects across geographic regions. *Oecologia*, 181(4):1139–1150, 2016. [p47]
- J. S. Clark, A. E. Gelfand, C. W. Woodall, and K. Zhu. More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999, 2014. [p37]
- J. S. Clark, D. Nemergut, B. Seyednasrollah, P. J. Turner, and S. Zhang. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56, 2017. [p37]
- J. W. Doser, A. O. Finley, M. Kéry, and E. F. Zipkin. spoccupancy: An r package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, 2022. [p37]

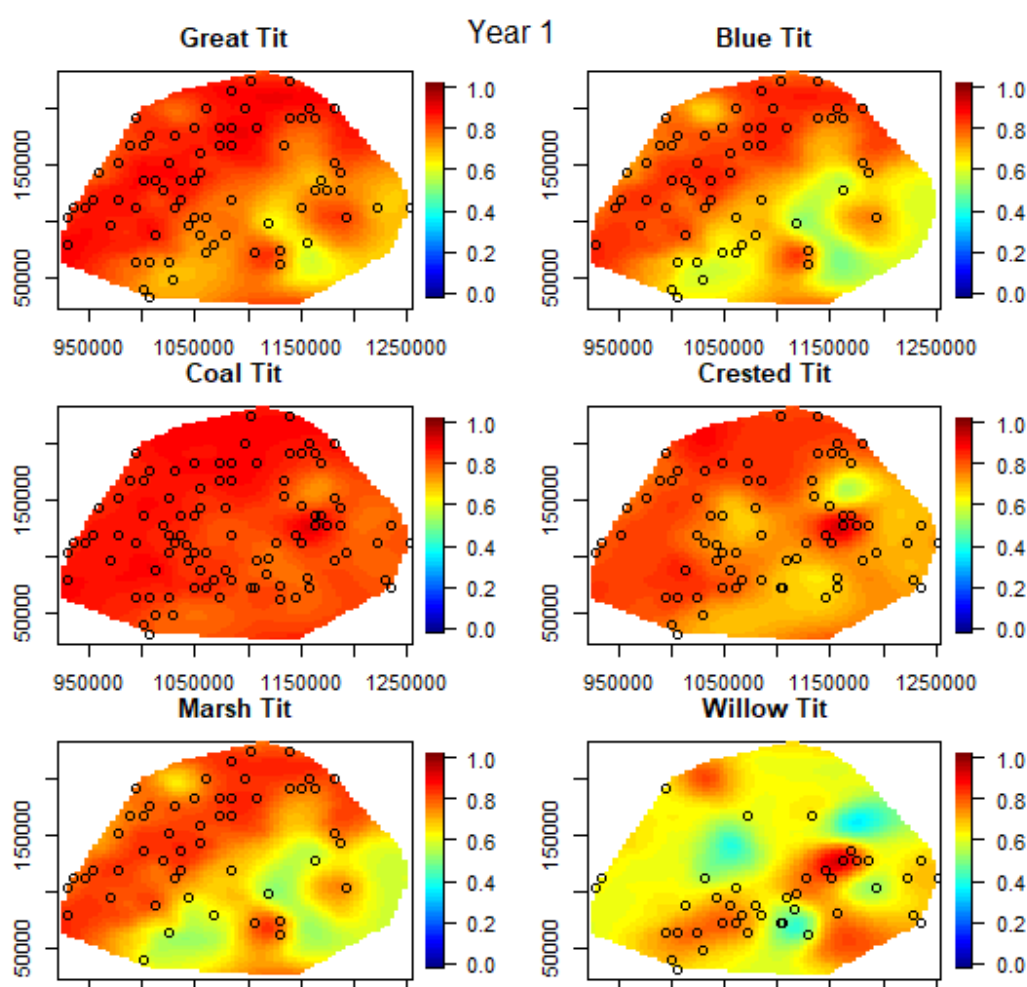


Figure 5: An example of occupancy predictions for season 1. Colors show posterior means of ψ for each location and species, interpolated across space using a thin plate spline for ease of visualization only. Hollow circles show points with detections of the species in that particular season. As expected, regions of higher posterior probability match regions with greater density of detections.

- A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC press, 2010. [p41]
- G. Guillerá-Arroita, M. Kéry, and J. Jhaz-Monfort. Inferring species richness using multispecies occupancy modeling: Estimation performance and interpretation. *Ecology and Evolution*, 9(2): 780–792, 2019. [p37]
- A. Guisan and C. Rahbek. Sesam—a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages, 2011. [p37]
- S. Hepler, R. Erhardt, and T. Anderson. Identifying drivers of spatial variation in occupancy with limited replication camera trap data. *Ecology*, 99(10):2152–2158, 2018. [p41]
- S. Hepler, K. Kaufeld, K. Benedict, M. Toda, B. Jackson, X. Liu, and D. Kline. Integrating public health surveillance and environmental data to model presence of *Histoplasma* in the United States. *Epidemiology*, 2022. doi: 10.1097/EDE.0000000000001499. [p37]
- S. A. Hepler and R. J. Erhardt. A spatiotemporal model for multivariate occupancy data. *Environmetrics*, 32(2):e2657, 2021. [p37, 38, 39, 41, 47]
- J. Hughes and M. Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1): 139–159, 2013. [p39]

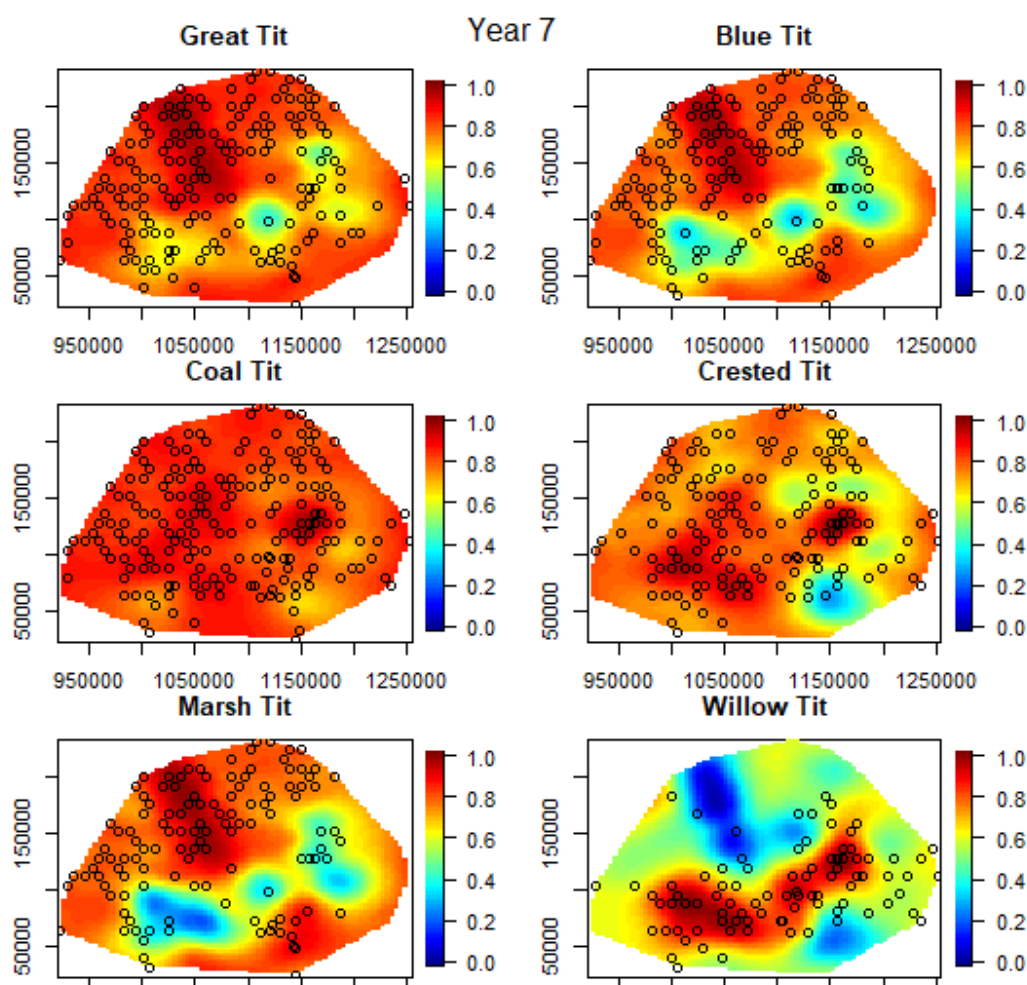


Figure 6: An example of occupancy predictions for season 7. Colors show posterior means of ψ for each location and species, interpolated across space using a thin plate spline for ease of visualization only. Hollow circles show points with detections of the species in that particular season. As expected, regions of higher posterior probability match regions with greater density of detections.

Information Systems and Wake Forest University. WFU High Performance Computing Facility, 2021. URL <https://hpc.wfu.edu>. [p47]

D. Johnson, P. Conn, M. Hooten, J. Ray, and B. Pond. Spatial occupancy models for large data sets. *Ecology*, 94(4):801–808, 2013. [p37]

M. Kéry, A. Royle, and M. Meredith. Ahmbook: functions and data for the book ‘applied hierarchical modeling in ecology’. *R package version 0.1, 3*, 2017. [p43]

S. Lele, M. Moreno, and E. Bayne. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5(1):22–31, 2012. [p41]

D. MacKenzie, J. Nichols, G. Lachman, S. Droege, J. Royle, and C. Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002. [p38, 41]

K. Mardia. Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284, 1988. [p39]

O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology letters*, 20(5):561–576, 2017. [p37]

- L. J. Pollock, R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014. [p37]
- H. Rahman, K. McCarthy, J. McCarthy, and M. Faisal. Application of multi-species occupancy modeling to assess mammal diversity in northeast Bangladesh. *Global Ecology and Conservation*, 25, 2021. [p37]
- C. Rota, C. Winkle, R. Kays, T. Forrester, W. McShea, A. Parsons, and J. Millsaugh. A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. *Ecology*, 97(1):48–53, 2016. [p37]
- C. T. Rota, R. J. Fletcher Jr, R. M. Dorazio, and M. G. Betts. Occupancy estimation and the closure assumption. *Journal of Applied Ecology*, 46(6):1173–1181, 2009. [p38]
- J. A. Royle and R. M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, 2008. [p38]
- J. Sanderlin, J. Golding, T. Wilcox, D. Mason, K. McKelvey, D. Pearson, and M. Schwartz. Occupancy modeling and resampling overcomes low test sensitivity to produce accurate SARS-CoV-2 prevalence estimates. *BMC Public Health*, 21(577), 2021. [p37]
- N. C. Stenseth, J. M. Durant, M. S. Fowler, E. Matthysen, F. Adriaensen, N. Jonzén, K.-S. Chan, H. Liu, J. De Laet, B. C. Sheldon, et al. Testing for effects of climate change on competitive relationships and coexistence between two bird species. *Proceedings of the Royal Society B: Biological Sciences*, 282(1807):20141958, 2015. [p47]
- D. Taylor-Rodriguez, K. Kaufeld, E. M. Schliep, J. S. Clark, and A. E. Gelfand. Joint species distribution modeling: dimension reduction using dirichlet processes. *Bayesian Analysis*, 12(4):939–967, 2017. [p37, 42]
- G. Tikhonov, Ø. H. Opedal, N. Abrego, A. Lehtikainen, M. M. de Jonge, J. Oksanen, and O. Ovaskainen. Joint species distribution modelling with the r-package hmsc. *Methods in ecology and evolution*, 11(3): 442–447, 2020. [p37]
- M. W. Tobler, M. Kéry, F. K. Hui, G. Guillera-Arroita, P. Knaus, and T. Sattler. Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8):e02754, 2019. [p47]
- A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017. [p42]

Staci Hepler

Department of Statistical Sciences, Wake Forest University
Winston-Salem NC, 27103 USA
heplersa@wfu.edu

Robert Erhardt

Department of Statistical Sciences, Wake Forest University
Winston-Salem NC, 27103 USA
erhardrj@wfu.edu