Spatially-explicit occupancy modeling with the spOccupancy R package

Jeff Doser and Elise Zipkin
Michigan State University
TWS 2023
November 9, 2023

Course Website

- https://doserjef.github.io/TWS23-spOccupancy/
- Single-species non-spatial/spatial occupancy models
- Multi-species non-spatial/spatial occupancy models
- Multi-season non-spatial/spatial occupancy models

Occupancy Models

Ecological Motivation

- Conservation management often requires information on species distributions across space and time

Ecological Motivation

- Conservation management often requires information on species distributions across space and time
- Examples
  - Species occurrence in protected vs. unprotected areas
  - Management effects on a community of species
  - Trends in species occurrence over time as a measure of population change

Ecological Motivation

- Conservation management often requires information on species distributions across space and time
- Examples
  - Species occurrence in protected vs. unprotected areas
  - Management effects on a community of species
  - Trends in species occurrence over time as a measure of population change
- Two important complexities when modeling species distributions:
  - Imperfect detection

Ecological Motivation

- Conservation management often requires information on species distributions across space and time
- Examples
  - Species occurrence in protected vs. unprotected areas
  - Management effects on a community of species
  - Trends in species occurrence over time as a measure of population change
- Two important complexities when modeling species distributions:
  - Imperfect detection
  - Spatial autocorrelation

What data do we use to quantify species distributions?

- Presence-only data: locations where species occur

What data do we use to quantify species distributions?

- Presence-only data: locations where species occur
- Presence-absence data: data takes value 1 if a species is present at some location and/or time point and 0 if the species is absent

What data do we use to quantify species distributions?

- Presence-only data: locations where species occur
- Presence-absence data: data takes value 1 if a species is present at some location and/or time point and 0 if the species is absent
- Count data: a non-negative integer value indicating the number of individuals of a species at some location and/or time point

What data do we use to quantify species distributions?

- Presence-only data: locations where species occur
- **Presence-absence data: data takes value 1 if a species is present at some location and/or time point and 0 if the species is absent**
- Count data: a non-negative integer value indicating the number of individuals of a species at some location and/or time point

Presence-absence data

Site	Survey
1	1
2	0
3	1
4	1
5	0
6	0

Presence-absence data

Site	Survey
1	1
2	0
3	1
4	1
5	0
6	0

Assuming no false positives, if we detect the species, we know it exists at the site

Presence-absence data

Site	Survey
1	1
2	0
3	1
4	1
5	0
6	0

Assuming no false positives, if we detect the species, we know it exists at the site

A 0 (or nondetection) could mean:

Presence-absence data

Site	Survey
1	1
2	0
3	1
4	1
5	0
6	0

Assuming no false positives, if we detect the species, we know it exists at the site

A 0 (or nondetection) could mean:

1. The species does not exist at the site

Presence-absence data

Site	Survey
1	1
2	0
3	1
4	1
5	0
6	0

Assuming no false positives, if we detect the species, we know it exists at the site

A 0 (or nondetection) could mean:

1. The species does not exist at the site
2. The species exists at the site, but we failed to detect it.

Occupancy modeling

- Developed to more accurately estimate species distributions when *imperfect detection* (false negatives) exists (MacKenzie et al. 2002; Tyre et al. 2003)

Occupancy modeling

- Developed to more accurately estimate species distributions when *imperfect detection* (false negatives) exists (MacKenzie et al. 2002; Tyre et al. 2003)
- **Fundamental concept**: obtain "repeated surveys" at a given site during some period of closure

Occupancy modeling

- Developed to more accurately estimate species distributions when *imperfect detection* (false negatives) exists (MacKenzie et al. 2002; Tyre et al. 2003)
- **Fundamental concept**: obtain "repeated surveys" at a given site during some period of closure
    - Key assumption: the species does not move in or out of the site during this time period

Occupancy modeling

- Developed to more accurately estimate species distributions when *imperfect detection* (false negatives) exists (MacKenzie et al. 2002; Tyre et al. 2003)
- **Fundamental concept**: obtain "repeated surveys" at a given site during some period of closure
  - Key assumption: the species does not move in or out of the site during this time period
- "Repeated surveys" usually come in the form of multiple visits to a site during some time period, but can also take different forms (e.g., multiple observers, spatial replicates)

Data for occupancy modeling

Detection-nondetection matrix (y)

k

j

Site	Survey 1	Survey 2	Survey 3	Survey 4
1	1	0	0	1
2	0	0	0	0
3	1	1	0	NA
4	1	NA	0	NA
5	0	1	1	1
6	0	0	0	1

$y_{j,\,k}$

- *J* sites with K_j replicate surveys at each site *j*
- Assume no false positives
- Any variation in the observed data values across surveys is assumed to arise from imperfect detection.

Occupancy model structure

- Two distinct sub-models
    1. Model occupancy probability as a function of site-level covariates

Occupancy model structure

- Two distinct sub-models
  1. Model occupancy probability as a function of site-level covariates
  2. Model detection probability as a function of site and/or survey-level covariates
     - Can only detect a species if it truly occupies a site
     - Detection probability is modeled "conditional" on true occupancy

Single-species occupancy model

$j = 1, \ldots, J$ (site)
$k = 1, \ldots, K_j$ (replicate)

Occupancy (ecological) sub-model

$$z_j \sim \text{Bernoulli}(\psi_j)$$
$$\text{logit}(\psi_j) = \beta_1 + \beta_2 \cdot X_{2,j} + \cdots + \beta_r \cdot X_{r,j}$$

z_j True occurrence of the species at site *j*

ψ_j Occurrence probability at site *j*

$X_{r,j}$ The *r*th covariate at site *j* (e.g., habitat variable)

Single-species occupancy model

$j = 1, \ldots, J$ (site)
$k = 1, \ldots, K_j$ (replicate)

Detection (observation) sub-model

$$y_{j,k} \sim \text{Bernoulli}(p_{j,k} \cdot z_j)$$
$$\text{logit}(p_{j,k}) = \alpha_1 + \alpha_2 \cdot V_{2,j,k} + \cdots + \alpha_r \cdot V_{r,j,k}$$

$y_{j,k}$ Detection-nondetection data at site *j* during replicate *k*

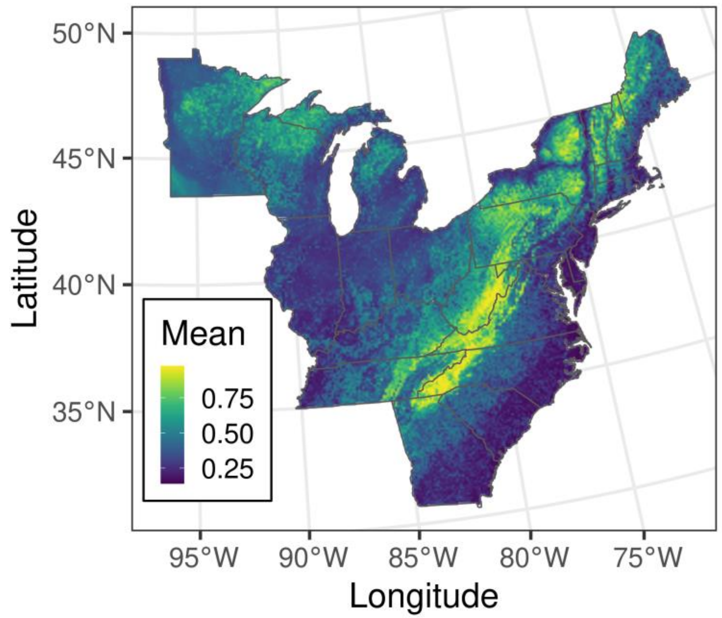
$p_{j,k}$ Detection probability at site *j* during replicate *k*

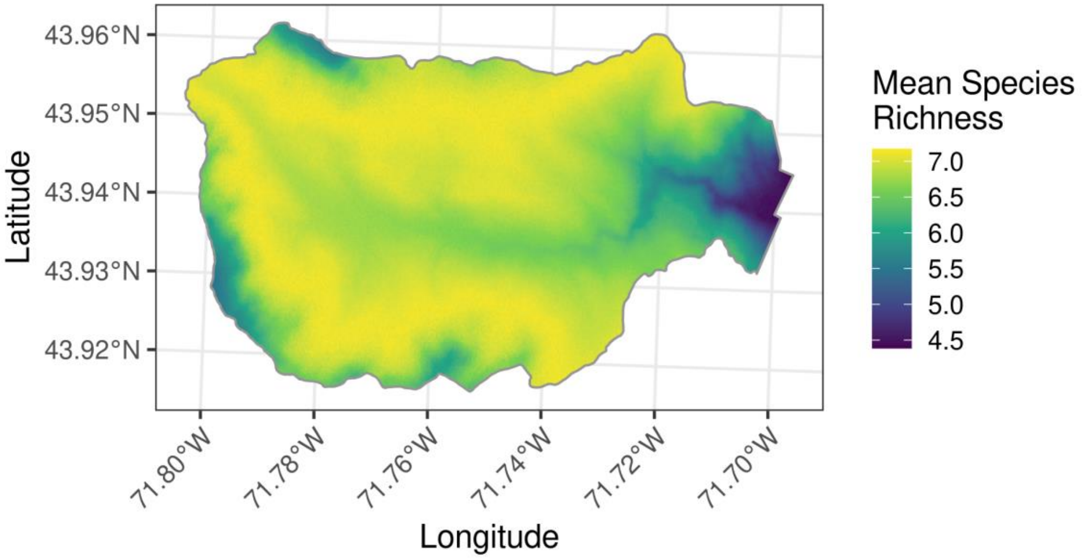
$V_{r,j,k}$ Covariate affecting detection at site j during replicate k

Spatial Occupancy Models

Spatial autocorrelation

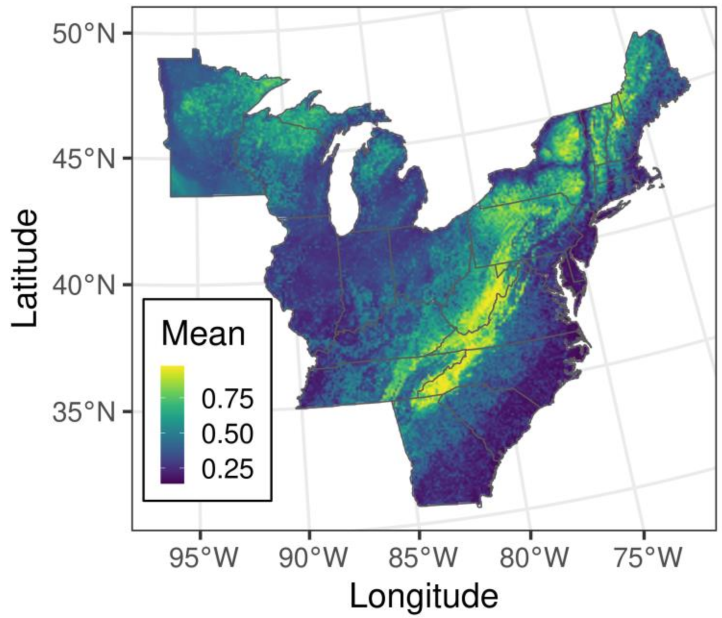
- First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." - Waldo Tobler

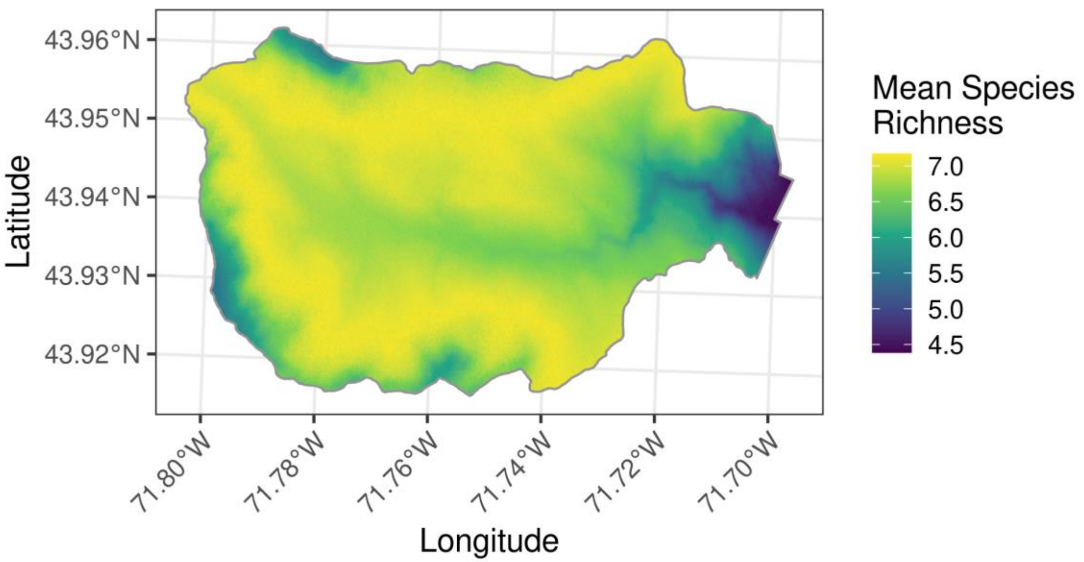




Spatial autocorrelation

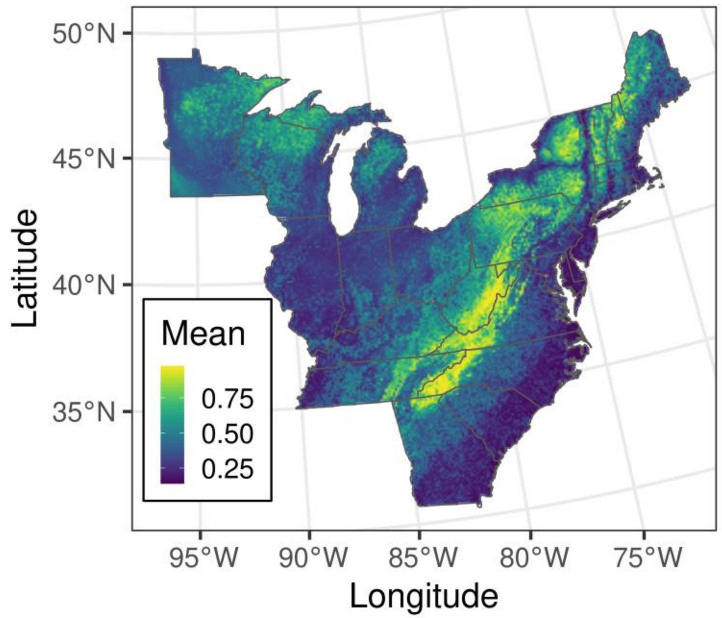
- First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." - Waldo Tobler
- What leads to spatial autocorrelation in species distributions?
  - Environmental drivers, habitat requirements
  - Biotic factors (dispersal, conspecific attraction)

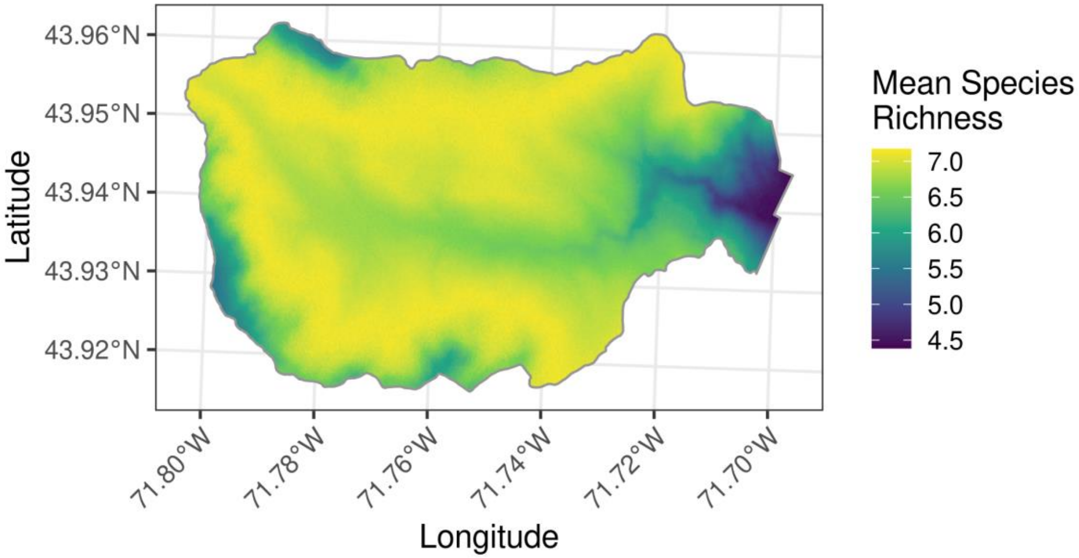




Spatial autocorrelation

- First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." - Waldo Tobler
- What leads to spatial autocorrelation in species distributions?
  - Environmental drivers, habitat requirements
  - Biotic factors (dispersal, conspecific attraction)
- Initial approach: attempt to explain spatial variation in species distributions with covariates (e.g., forest cover, temperature, elevation)





Residual spatial autocorrelation

Residual spatial autocorrelation

- Spatial correlation in occupancy probability *after* including spatial covariates

Residual spatial autocorrelation

- Spatial correlation in occupancy probability *after* including spatial covariates
- Often arises from missing/unavailable covariates

Residual spatial autocorrelation

- Spatial correlation in occupancy probability *after* including spatial covariates
- Often arises from missing/unavailable covariates
- Can lead to bias if unaddressed

Residual spatial autocorrelation

- Spatial correlation in occupancy probability *after* including spatial covariates
- Often arises from missing/unavailable covariates
- Can lead to bias if unaddressed
- Account for using spatial random effects
  - Each site has a local adjustment in occupancy probability
  - The local adjustments are given a spatial structure
  - Estimated parameters: spatial variance and spatial decay

Single-species spatial occupancy model

$j = 1, \ldots, J$ (site)
$k = 1, \ldots, K_j$ (replicate)

Occupancy (ecological) sub-model

$$z_j \sim \text{Bernoulli}(\psi_j)$$
$$\text{logit}(\psi_j) = \beta_1 + \beta_2 \cdot X_{2,j} + \cdots + \beta_r \cdot X_{r,j} + \text{w}_j$$
$$\text{w}_j \sim \text{Normal}(0, \boldsymbol{\Sigma})$$

Detection (observation) sub-model

$$y_{j,k} \sim \text{Bernoulli}(p_{j,k} \cdot z_j)$$
$$\text{logit}(p_{j,k}) = \alpha_1 + \alpha_2 \cdot V_{2,j,k} + \cdots + \alpha_r \cdot V_{r,j,k}$$

Single-species spatial occupancy model

$j = 1, \ldots, J$ (site)

$k = 1, \ldots, K_j$ (replicate)

Occupancy (ecological) sub-model

$$z_j \sim \text{Bernoulli}(\psi_j)$$

$$\text{logit}(\psi_j) = \beta_1 + \beta_2 \cdot X_{2,j} + \cdots + \beta_r \cdot X_{r,j} + \text{w}_j$$

$$\boxed{\text{w}_j \sim \text{Normal}(0, \boldsymbol{\Sigma})}$$

Detection (observation) sub-model

$$y_{j,k} \sim \text{Bernoulli}(p_{j,k} \cdot z_j)$$

$$\text{logit}(p_{j,k}) = \alpha_1 + \alpha_2 \cdot V_{2,j,k} + \cdots + \alpha_r \cdot V_{r,j,k}$$

Gaussian process

- "Gold standard" for modeling spatial data

Gaussian process

- "Gold standard" for modeling spatial data
- Covariance between two sites is determined by:
  - Distance between the sites
  - A covariance function

Gaussian process

- "Gold standard" for modeling spatial data
- Covariance between two sites is determined by:
  - Distance between the sites
  - A covariance function
- spOccupancy supports four covariance functions: **exponential**, Gaussian, spherical, Matérn

Gaussian process

- "Gold standard" for modeling spatial data
- Covariance between two sites is determined by:
  - Distance between the sites
  - A covariance function
- spOccupancy supports four covariance functions: **exponential**, Gaussian, spherical, Matérn
- Covariance between site A and site B using exponential covariance function:

$$\boldsymbol{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \text{exp}(-\phi d_{A,B})$$

Intuition on spatial covariance

$$\mathbf{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \exp(-\phi d_{A,B})$$

$d_{A,B}$ Euclidean (linear) distance between site A and B

Intuition on spatial covariance

$$\boldsymbol{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \exp(-\phi d_{A,B})$$

$d_{A,B}$ Euclidean (linear) distance between site A and B

σ^2 Spatial variance. Controls magnitude of random effects

Intuition on spatial covariance

$$\mathbf{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \exp(-\phi d_{A,B})$$

$d_{A,B}$ Euclidean (linear) distance between site A and B

σ^2 Spatial variance. Controls magnitude of random effects

ϕ Spatial decay. Controls how quickly the correlation between sites decays across space

Intuition on spatial covariance

$$\boldsymbol{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \exp(-\phi d_{A,B})$$

$d_{A,B}$ Euclidean (linear) distance between site A and B

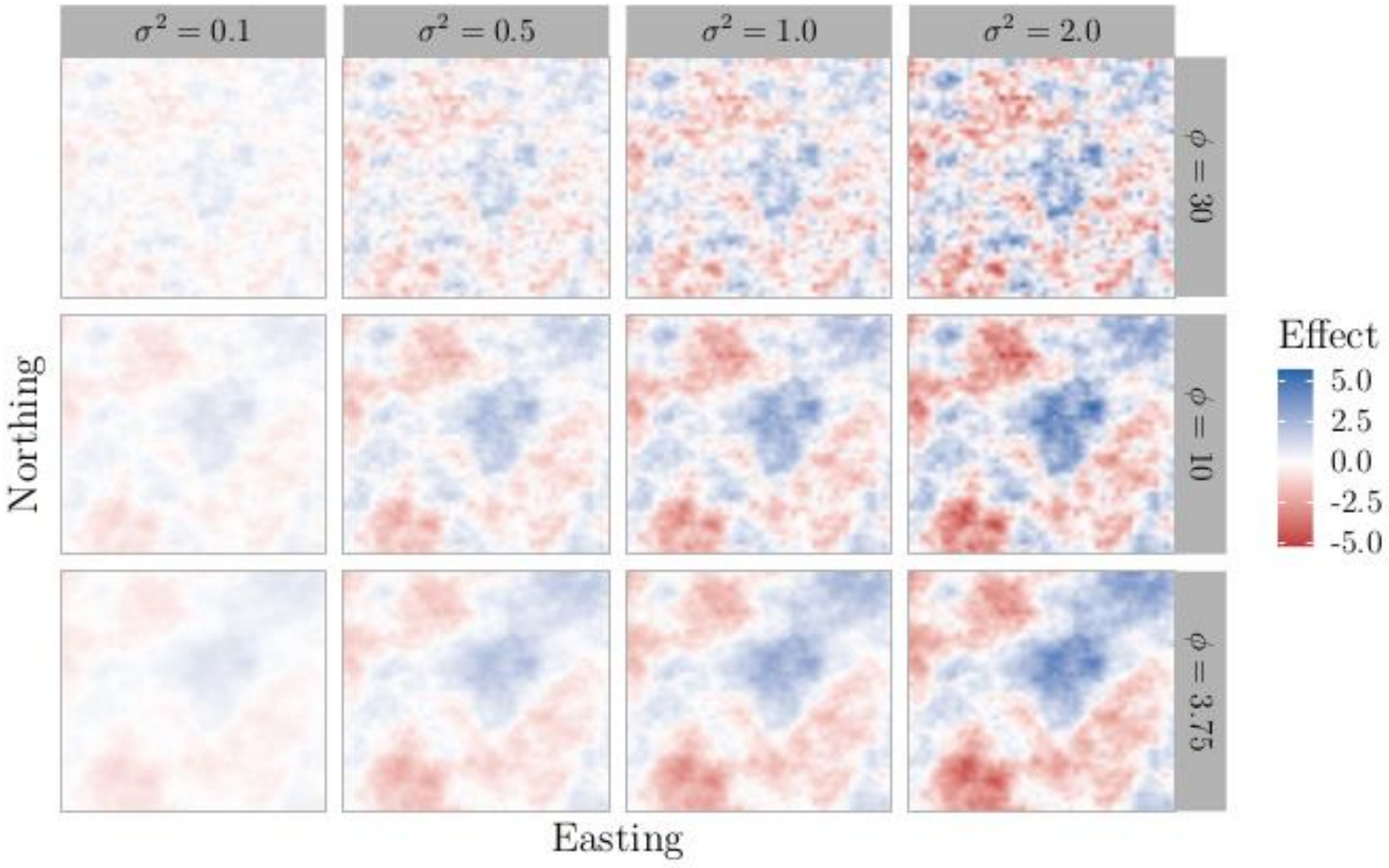
σ^2 Spatial variance. Controls magnitude of random effects

ϕ Spatial decay. Controls how quickly the correlation between sites decays across space

$\frac{3}{\phi}$ "Effective spatial range" when using an exponential covariance function. This is the distance at which the spatial correlation between two sites is essentially negligible (0.05)

Intuition on spatial covariance

$$\boldsymbol{\Sigma}(d_{A,B}, \sigma^2, \phi) = \sigma^2 \exp(-\phi d_{A,B})$$



Gaussian process

- Flexible approach to account for spatial autocorrelation

Gaussian process

- Flexible approach to account for spatial autocorrelation
- But... becomes extremely slow as the number of sites increases
- Not practical for data sets with hundreds of data points, let alone thousands.
- Computational bottleneck: dealing with a large, dense J x J matrix
- Need a more efficient approach...

Nearest Neighbor Gaussian Processes (NNGPs)

- Heaton et al. (2019): overview of approaches to model big spatial data

Nearest Neighbor Gaussian Processes (NNGPs)

- Heaton et al. (2019): overview of approaches to model big spatial data
- Our approach: NNGPs (Datta et al. 2016, Finley et al. 2019). Same parameters/interpretation as full GP

Nearest Neighbor Gaussian Processes (NNGPs)

- Heaton et al. (2019): overview of approaches to model big spatial data
- Our approach: NNGPs (Datta et al. 2016, Finley et al. 2019). Same parameters/interpretation as full GP
- Conceptually:
    1. Order the spatial locations (e.g., along the x-axis)

Nearest Neighbor Gaussian Processes (NNGPs)

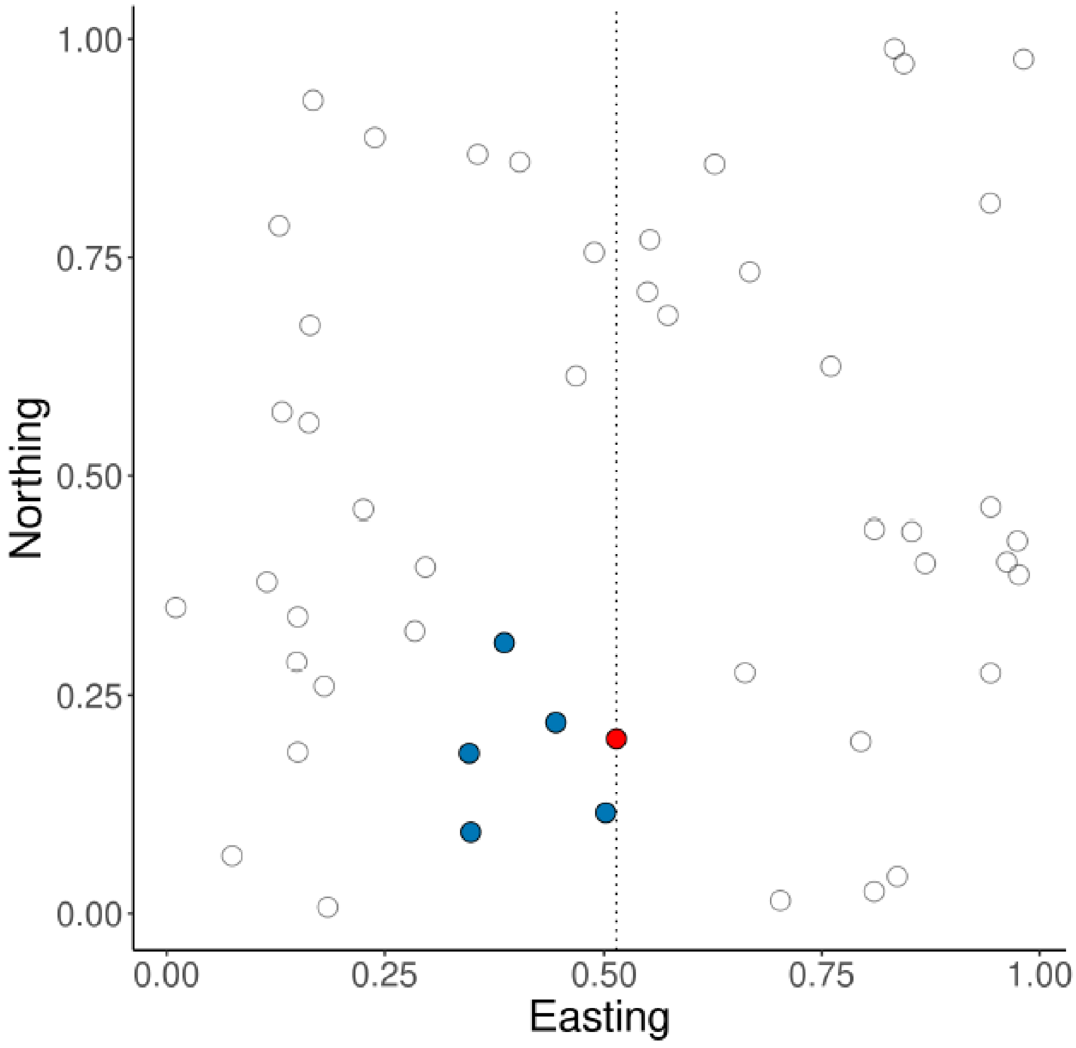
- Heaton et al. (2019): overview of approaches to model big spatial data
- Our approach: NNGPs (Datta et al. 2016, Finley et al. 2019). Same parameters/interpretation as full GP
- Conceptually:
  1. Order the spatial locations (e.g., along the x-axis)
  2. Determine the *m* nearest neighbors (subject to ordering) based on Euclidean (linear) distance

Nearest Neighbor Gaussian Processes (NNGPs)

- Heaton et al. (2019): overview of approaches to model big spatial data
- Our approach: NNGPs (Datta et al. 2016, Finley et al. 2019). Same parameters/interpretation as full GP
- Conceptually:
  1. Order the spatial locations (e.g., along the x-axis)
  2. Determine the *m* nearest neighbors (subject to ordering) based on Euclidean (linear) distance
  3. The spatial random effect at each site only depends on values of its *m* nearest neighbors and is conditionally independent of all other values

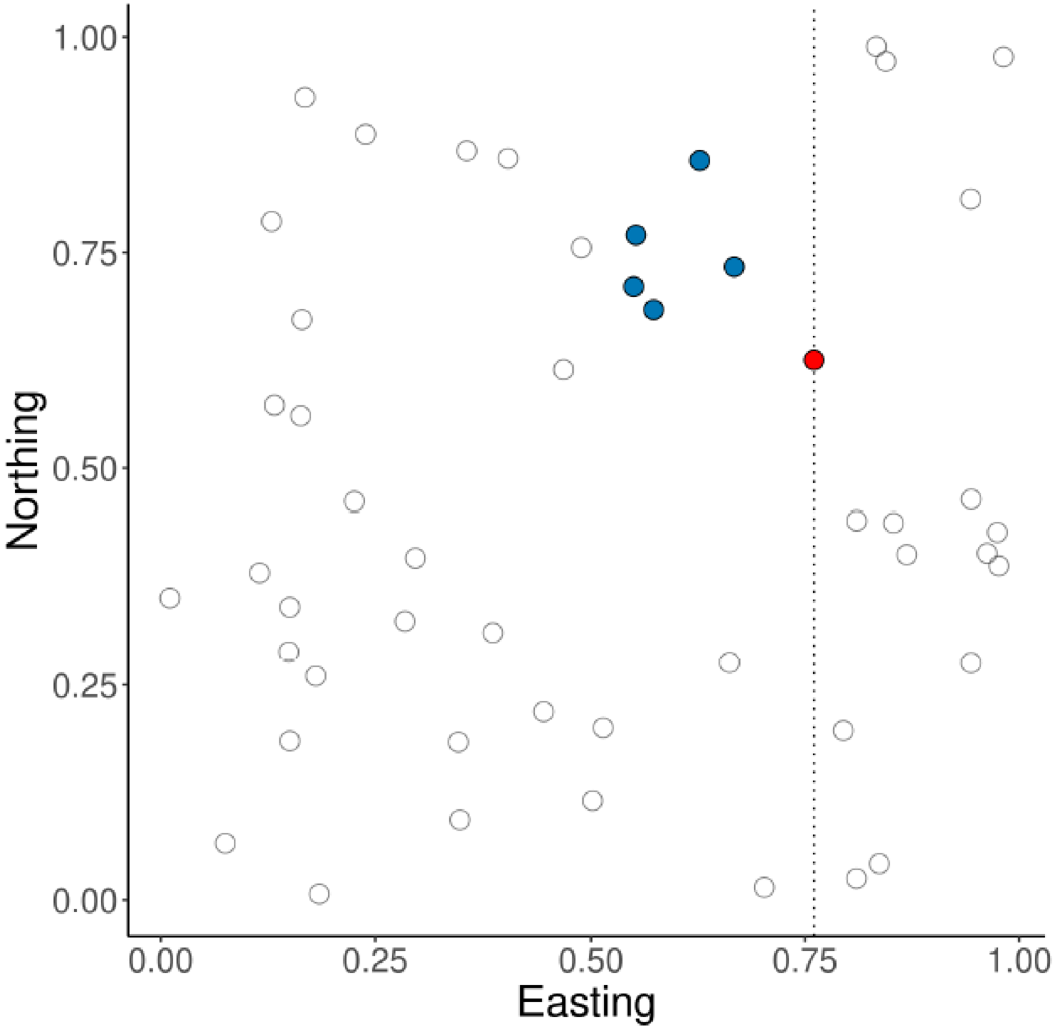
Choosing the neighbors

- spOccupancy orders sites along the horizontal axis (i.e., Easting)
- Example: NNGP with 5 neighbors
- Red point denotes the current site
- Blue points denote sites in the "neighbor set"



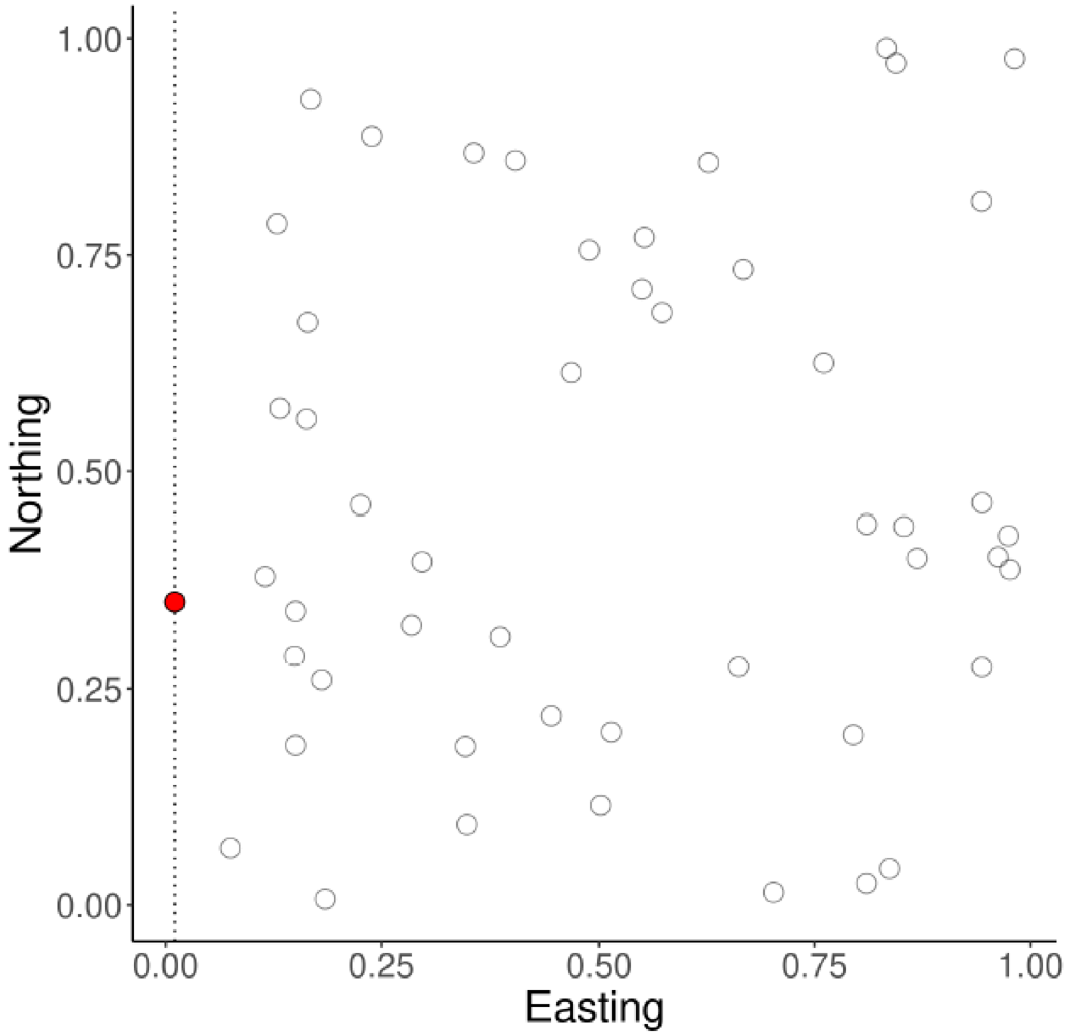
Choosing the neighbors

- spOccupancy orders sites along the horizontal axis (i.e., Easting)
- Example: NNGP with 5 neighbors
- Red point denotes the current site
- Blue points denote sites in the "neighbor set"

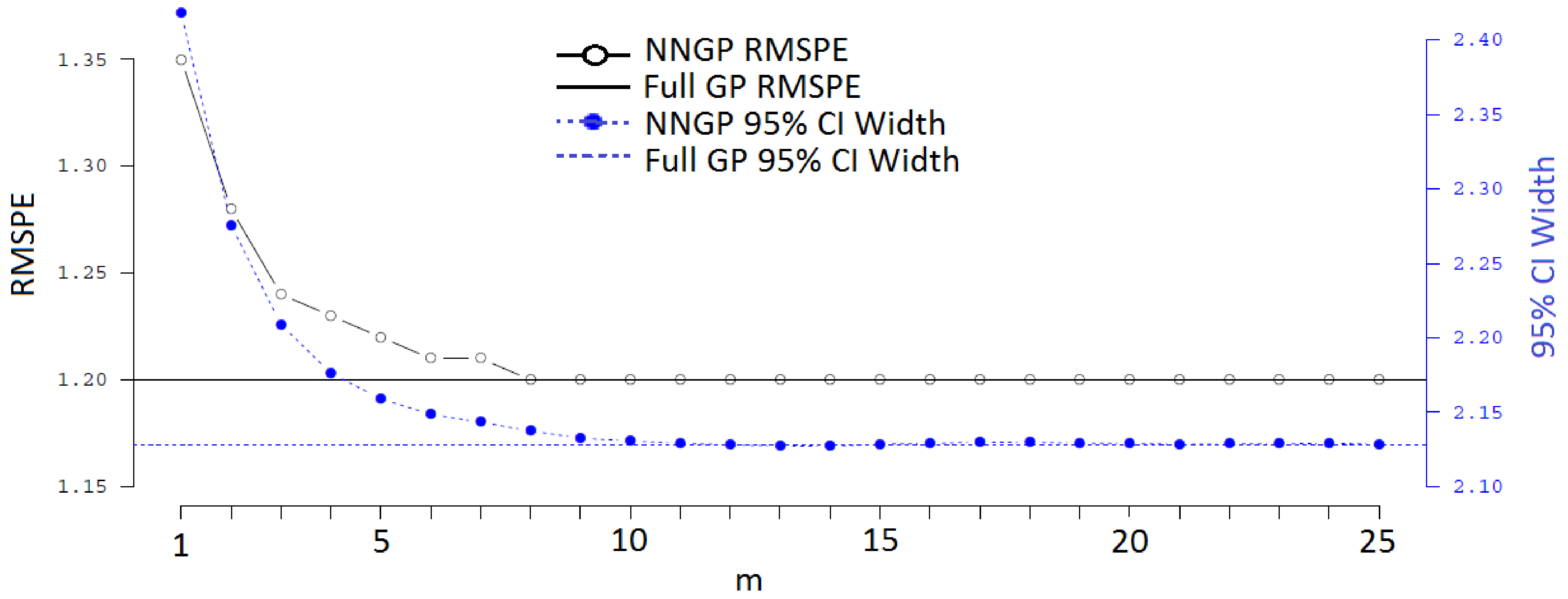


Choosing the neighbors

- spOccupancy orders sites along the horizontal axis (i.e., Easting)
- Example: NNGP with 5 neighbors
- Red point denotes the current site
- Blue points denote sites in the "neighbor set"

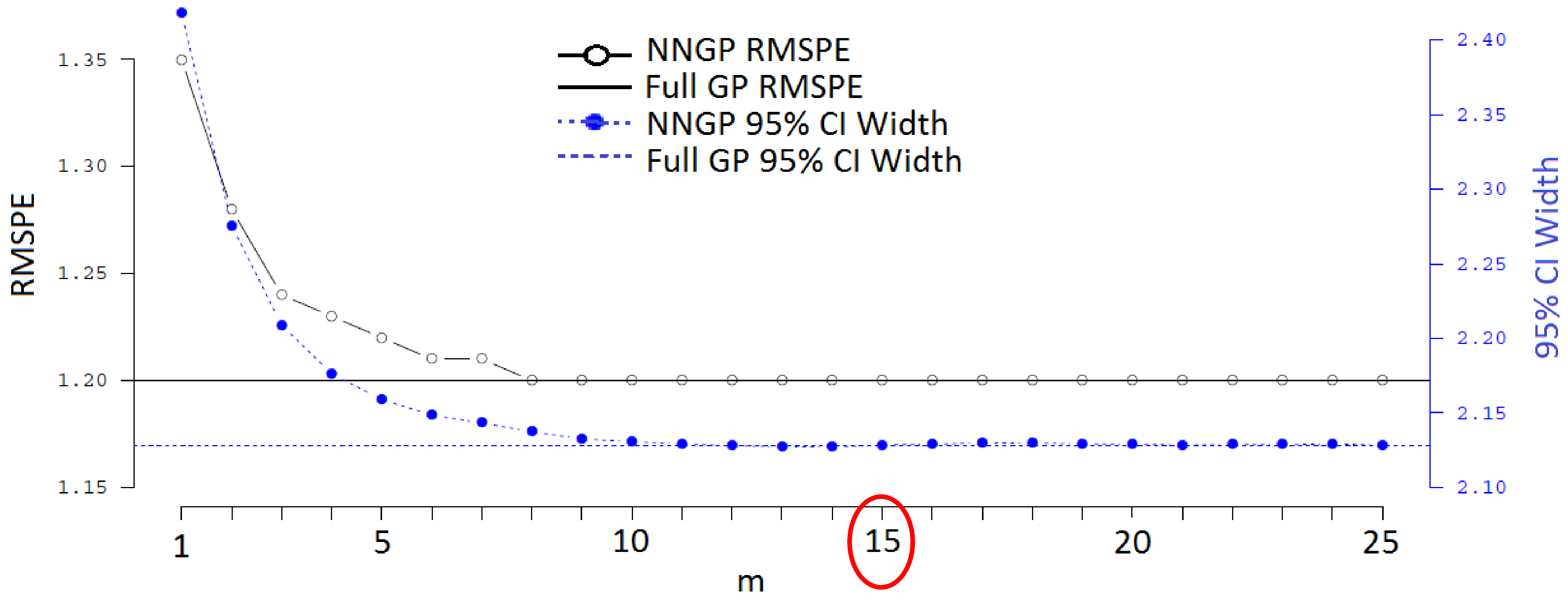


How many neighbors?



Datta et al. (2016) JASA

How many neighbors?



- m= 15 neighbors is often adequate (spOccupancy default)
- Can compare smaller m using WAIC

Datta et al. (2016) JASA

Pros/cons of spatial models

Pros

- More accurate species distribution maps (improved predictions)
- More accurate uncertainty estimates
- Provide insights on underlying drivers
- Generate new hypotheses

Pros/cons of spatial models

Pros

- More accurate species distribution maps (improved predictions)
- More accurate uncertainty estimates
- Provide insights on underlying drivers
- Generate new hypotheses

Cons

- Slower (but NNGPs help a lot!)
- Spatial confounding (Hanks et al. 2015, Mäkinen et al. 2022)
- More data hungry

Bayesian Basics

Why Bayesian for spatial occupancy models?

1. Interpretation

Why Bayesian for spatial occupancy models?

1. Interpretation
2. More flexible to accommodate spatial autocorrelation

Why Bayesian for spatial occupancy models?

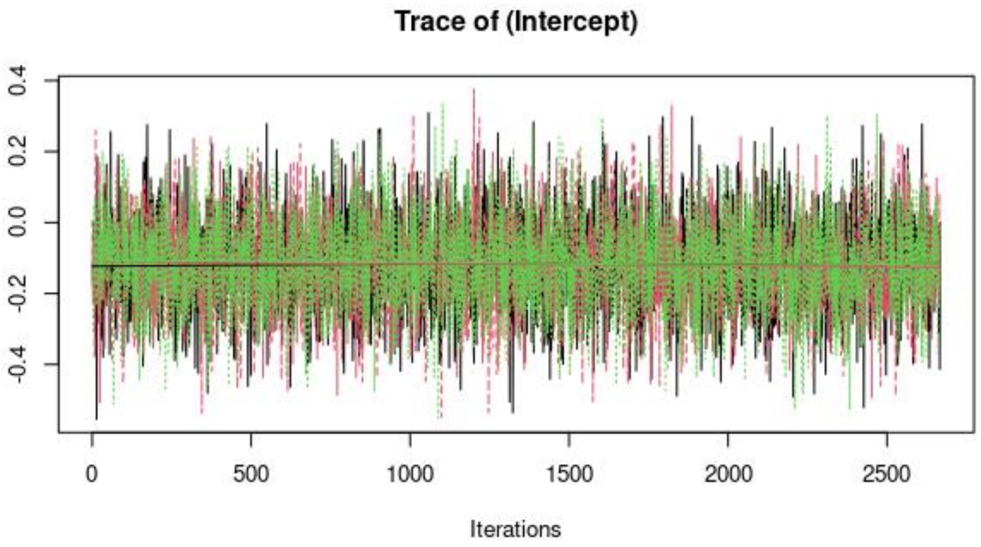
1. Interpretation
2. More flexible to accommodate spatial autocorrelation
3. Easy to extend to multispecies frameworks/integrate multiple data sources

Why Bayesian for spatial occupancy models?

1. Interpretation
2. More flexible to accommodate spatial autocorrelation
3. Easy to extend to multispecies frameworks/integrate multiple data sources
4. Fully propagate uncertainty in all estimates (and derived quantities)

Bayesian basics: what to know to get started in spOccupancy

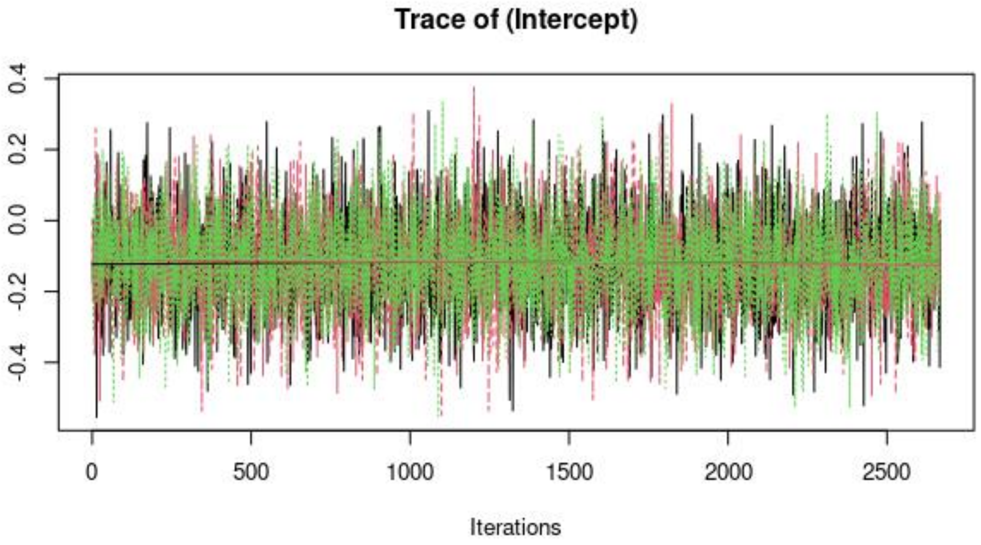
- Markov chain Monte Carlo (MCMC)
- MCMC chains eventually converge to a posterior distribution
  - Assess convergence by running multiple chains with different starting values



Bayesian basics: what to know to get started in spOccupancy

Monte

- Markov chain Monte Carlo (MCMC)
- MCMC chains eventually converge to a posterior distribution
  - Assess convergence by running multiple chains with different starting values



MCMC Step 1: Specify prior distributions

$$\begin{aligned}
\beta &\sim \mathrm{Normal}(\mu_\beta, \sigma^2_\beta) \\
\alpha &\sim \mathrm{Normal}(\mu_\alpha, \sigma^2_\alpha) \\
\sigma^2 &\sim \text{Inverse-Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
\phi &\sim \mathrm{Uniform}(a_\phi, b_\phi)
\end{aligned}$$

MCMC Step 2: Set initial values

- Set different values for each chain
- spOccupancy will set initial values by default
- Can be important for more complicated models (e.g., spatially-varying coefficient models)

MCMC Step 3: Propose new value

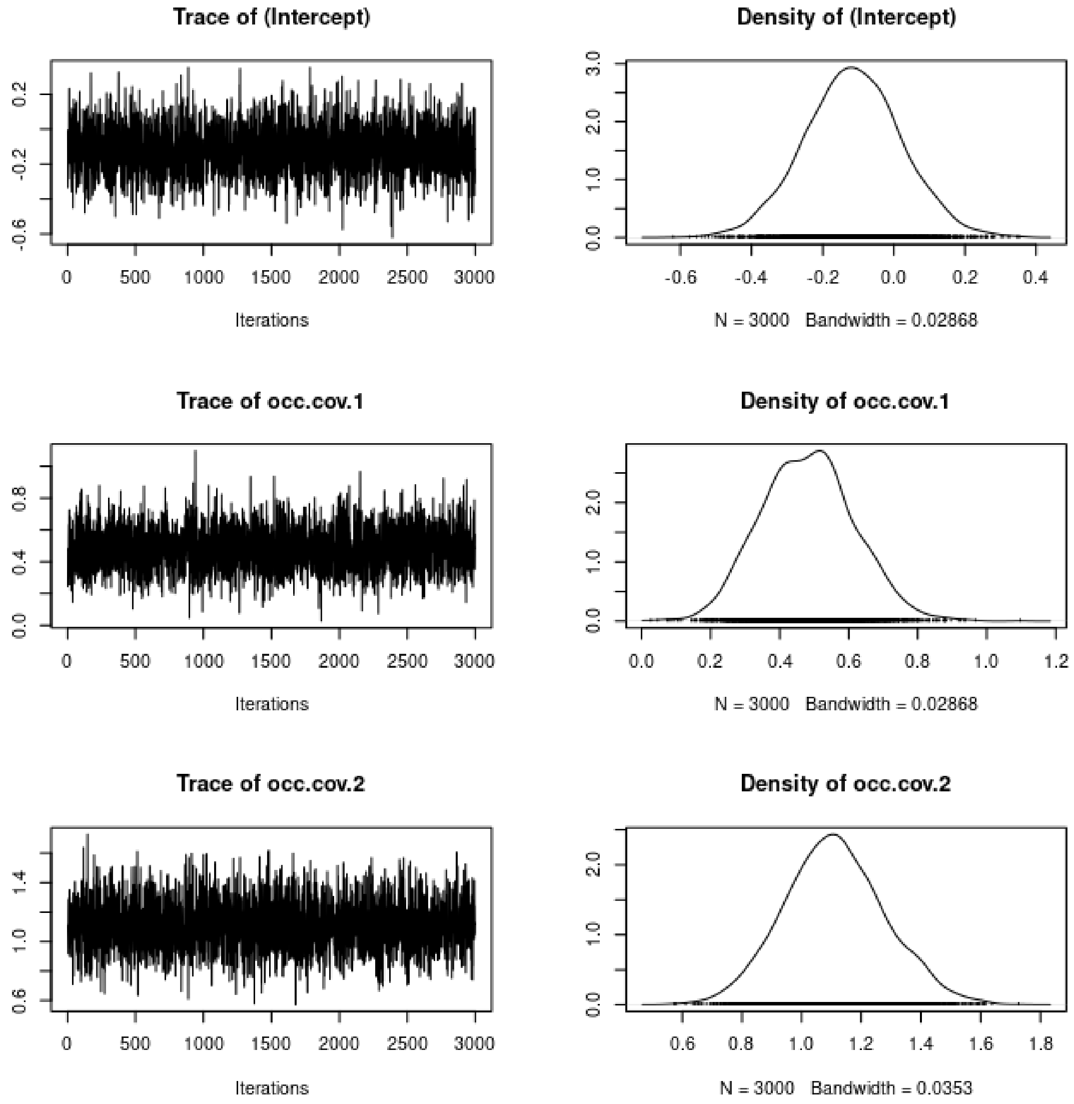
- Propose a new value for each parameter one at a time based on a statistical algorithm.
- For some parameters, we always accept the proposed value because our algorithm is efficient.
- For parameters with less efficient algorithms, we will accept the new value with some probability p.

MCMC Step 4: Repeat

- Repeat step 3 "many" times to generate a set of samples from the posterior distribution for each parameter.

MCMC Step 5: Summarize

- Point estimate: mean, median, mode
- Uncertainty: 95% credible (e.g., 2.5 and 97.5% quantiles of the samples)



What do you need to specify?

- Prior distribution (optional)
- Initial values (optional)
- Number of samples/iterations
- Burn-in: initial part of the MCMC chain that we throw away
- Thinning rate: how often do you want to save a sample?

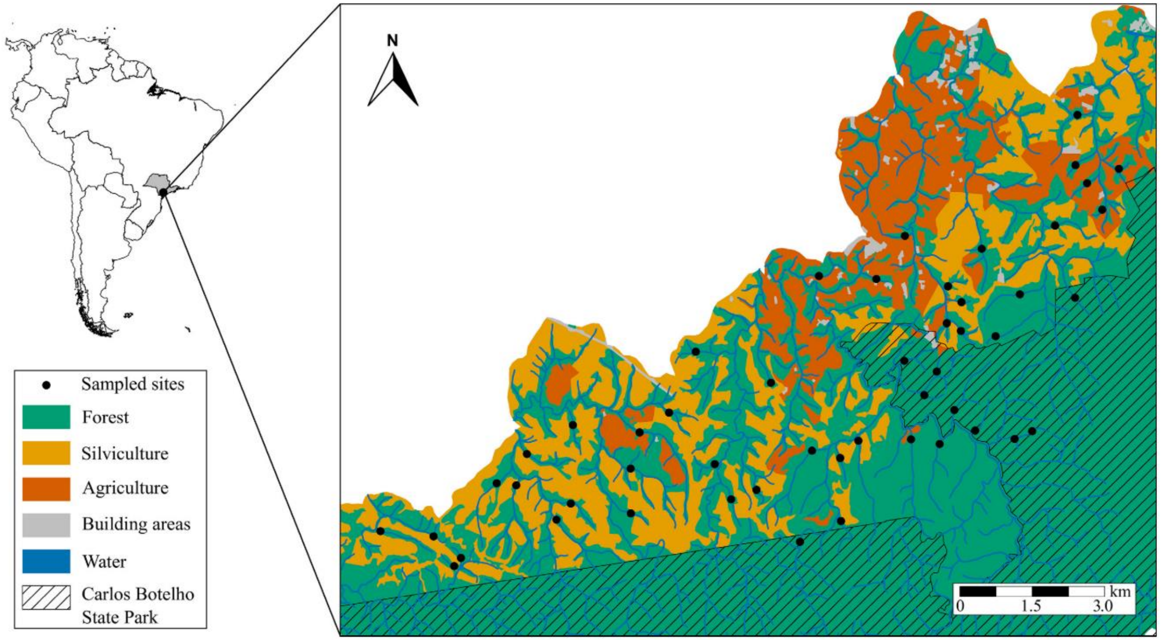
spOccupancy

spOccupancy

- Designed to fit Bayesian single-species and multi-species occupancy models
- Efficient options (NNGPs) to account for spatial autocorrelation
- Workflow completely in R (no Bayesian programming languages necessary)
- PGOcc -> single-species occupancy model
- spPGOcc -> spatial single-species occupancy model
- The "PG" stands for Pólya-Gamma (Polson et al. 2013)

Exercise 1: Amphibian occupancy in Brazil

- Data from [Ribeiro Jr. Et al (2018) *Eco Apps*]
- 50 sites along a gradient of landscape characteristics
- 3 ARU recordings at each site (repeat surveys/visits)
- 36 amphibian species analyzed
- Focus on *Crossodactylus caramaschii*



Ribeiro Jr. et al. (2018) *Eco Apps*

spOccupancy workflow

1. Data simulation/prep
2. Model fitting
3. Model validation
4. Model comparison
5. Posterior summaries
6. Prediction

Multi-species occupancy models

Multi-species detection-nondetection data

- Many types of multi-species inventories:
  - Point count surveys
  - Acoustic recording units
  - Camera traps
  - Citizen science checklists

Species	Site 1	Site 2	Site 3	Site 4
A	1	0	0	1
B	0	0	1	0
C	1	1	0	0
D	1	0	0	0
E	0	1	1	1
F	0	0	0	1

Multi-species detection-nondetection data

Visit 1

Species	Site 1	Site 2		
A	1	0		
B	0	0		
C	1	1		
D	1	0	0	0
E	0	1	1	1
F	0	0	0	1

Visit 2

Species	Site 1	Site 2		
A	0	NA		
B	0	NA		
C	1	NA		
D	0	NA		
E	0	NA	1	1
F	0	NA	0	0

Visit 3

Species	Site 1	Site 2	Site 3	Site 4
A	1	0	0	NA
B	0	1	1	NA
C	0	0	0	NA
D	0	0	0	NA
E	0	0	1	NA
F	0	0	0	NA

Ecological Motivation

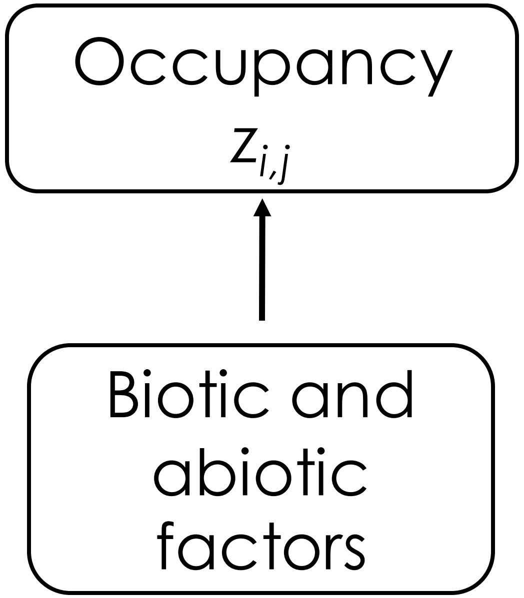
- Management has historically focused on individual species.
- Increased interest in multi-species management
- Biodiversity conservation
- Species are not independent of each other

Statistical Motivation

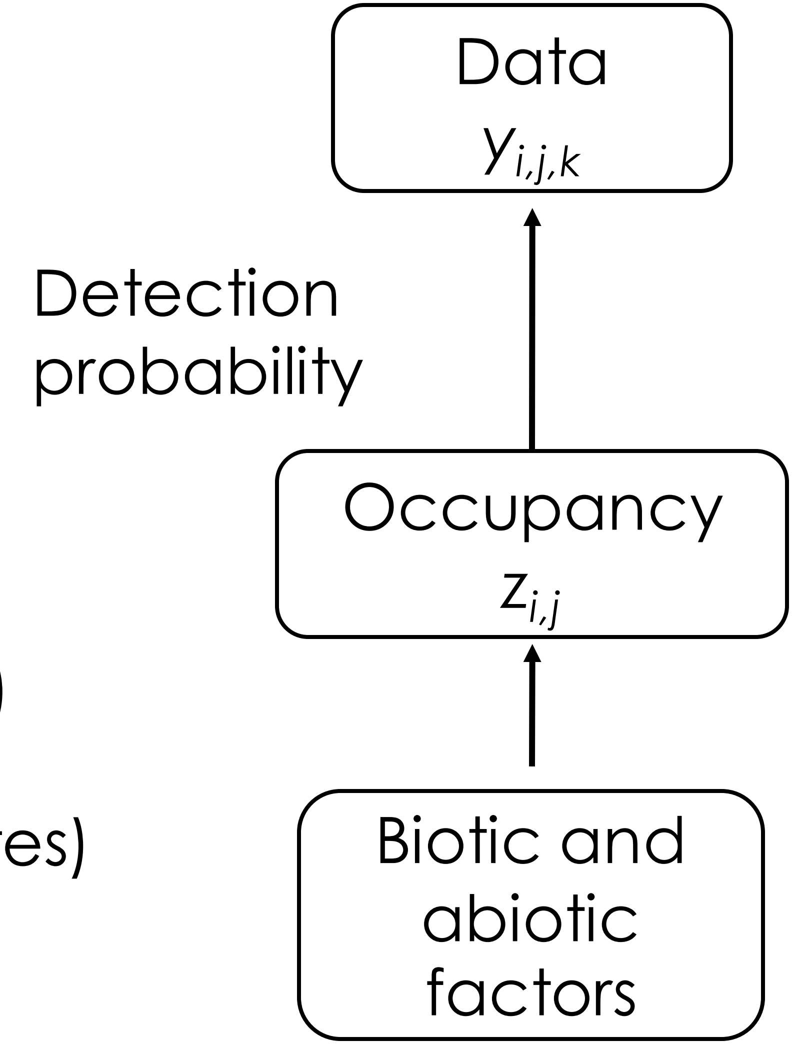
- Species of interest (e.g., SGCNs) are often the rarest species.
- Occupancy models are hard to fit when the number of detections is low
- Multi-species models can:
  - Improve ability to model rare species
  - Provide inference at both species and community-levels
  - Use information from other species to improve species-specific estimates

Multi-species occupancy model

i = 1, …, N (species)
j = 1, …, J (sites)
k = 1, …, K (replicates)

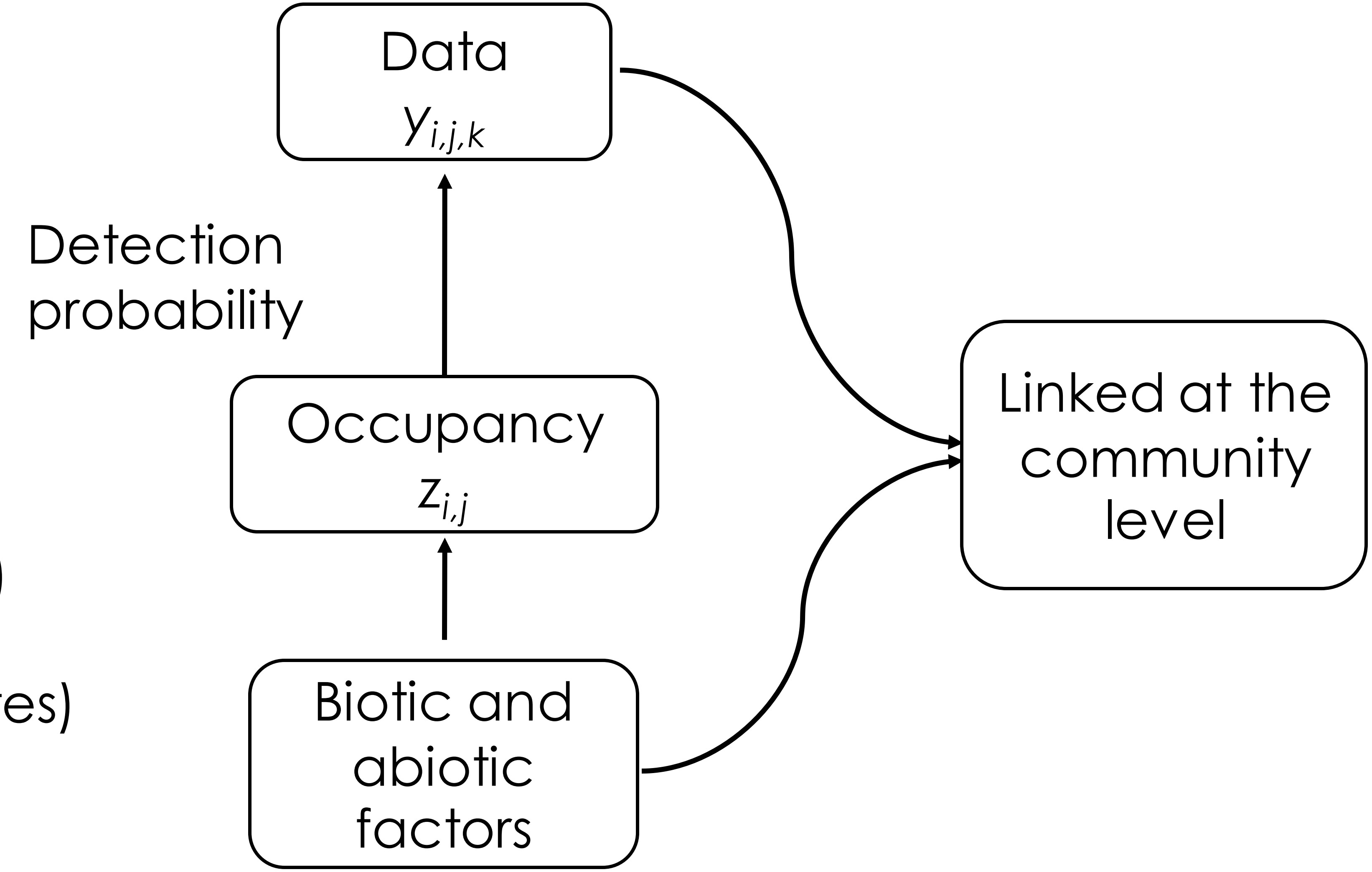


Multi-species occupancy model



i = 1, …, N (species)
j = 1, …, J (sites)
k = 1, …, K (replicates)

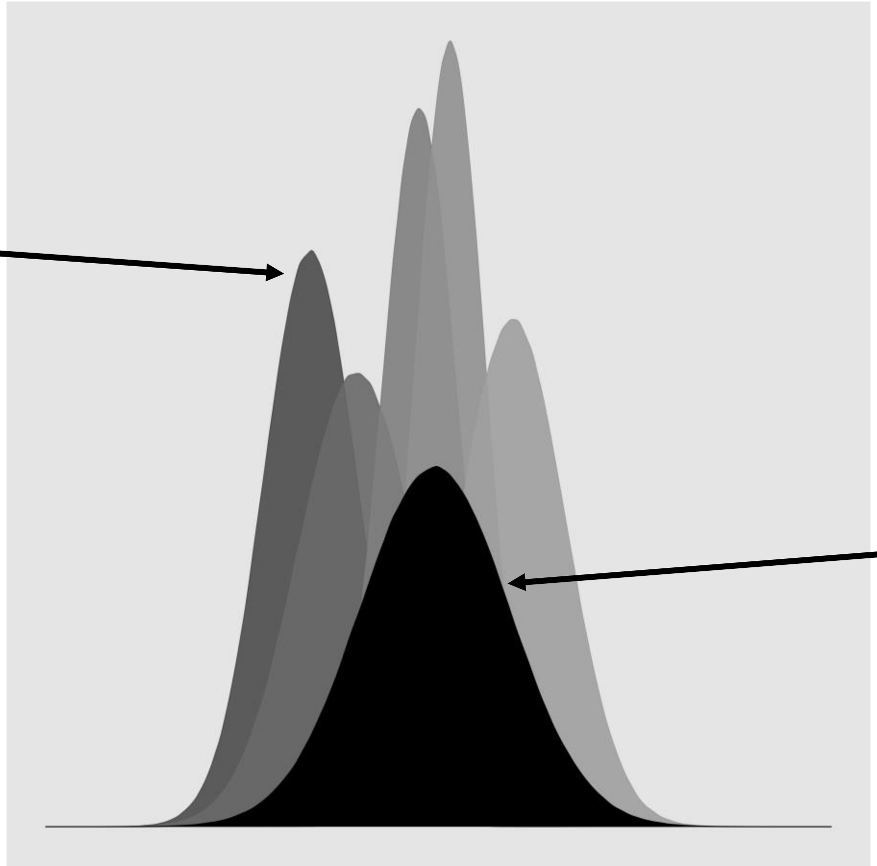
Multi-species occupancy model



i = 1, …, N (species)
j = 1, …, J (sites)
k = 1, …, K (replicates)

Species-specific and community effects

Species-specific effects



Community-level effect

Species-specific effects are drawn from a common, community-level distribution

Multi-species occupancy model

i = 1, …, N (species)
j = 1, …, J (sites)
k = 1, …, K (replicates)

Occupancy (ecological) sub-model

$$
\begin{aligned}
z_{i,j} &\sim \text{Bernoulli}(\psi_{i,j}) \\
\text{logit}(\psi_{i,j}) &= \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \cdots + \beta_{r,i} \cdot X_{r,j} \\
\beta_{r,i} &\sim \text{Normal}(\mu_{\beta_r}, \tau^2_{\beta,r})
\end{aligned}
$$

Detection (observation) sub-model

$$
\begin{aligned}
y_{i,j,k} &\sim \text{Bernoulli}(p_{i,j,k} \cdot z_{i,j}) \\
\text{logit}(p_{i,j,k}) &= \alpha_{1,i} + \alpha_{2,i} \cdot V_{2,j,k} + \cdots + \alpha_{r,i} \cdot V_{r,j,k} \\
\alpha_{r,i} &\sim \text{Normal}(\mu_{\alpha_r}, \tau^2_{\alpha,r})
\end{aligned}
$$

Multi-species occupancy model

i = 1, …, N (species)
j = 1, …, J (sites)
k = 1, …, K (replicates)

Occupancy (ecological) sub-model

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$
$$\text{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \cdots + \beta_{r,i} \cdot X_{r,j}$$
$$\boxed{\beta_{r,i} \sim \text{Normal}(\mu_{\beta_r}, \tau^2_{\beta,r})}$$

Detection (observation) sub-model

$$y_{i,j,k} \sim \text{Bernoulli}(p_{i,j,k} \cdot z_{i,j})$$
$$\text{logit}(p_{i,j,k}) = \alpha_{1,i} + \alpha_{2,i} \cdot V_{2,j,k} + \cdots + \alpha_{r,i} \cdot V_{r,j,k}$$
$$\alpha_{r,i} \sim \text{Normal}(\mu_{\alpha_r}, \tau^2_{\alpha,r})$$

Multi-species occupancy model

$$\beta_{r,i} \sim \mathrm{Normal}(\mu_{\beta_r}, \tau^2_{\beta_r})$$

μ_{β_r} Mean effect of covariate across all species

$\tau^2_{\beta_r}$ Variance of the covariate effect among all species

Multi-species occupancy model

$$\beta_{r,i} \sim \text{Normal}(\mu_{\beta_r}, \tau^2_{\beta_r})$$

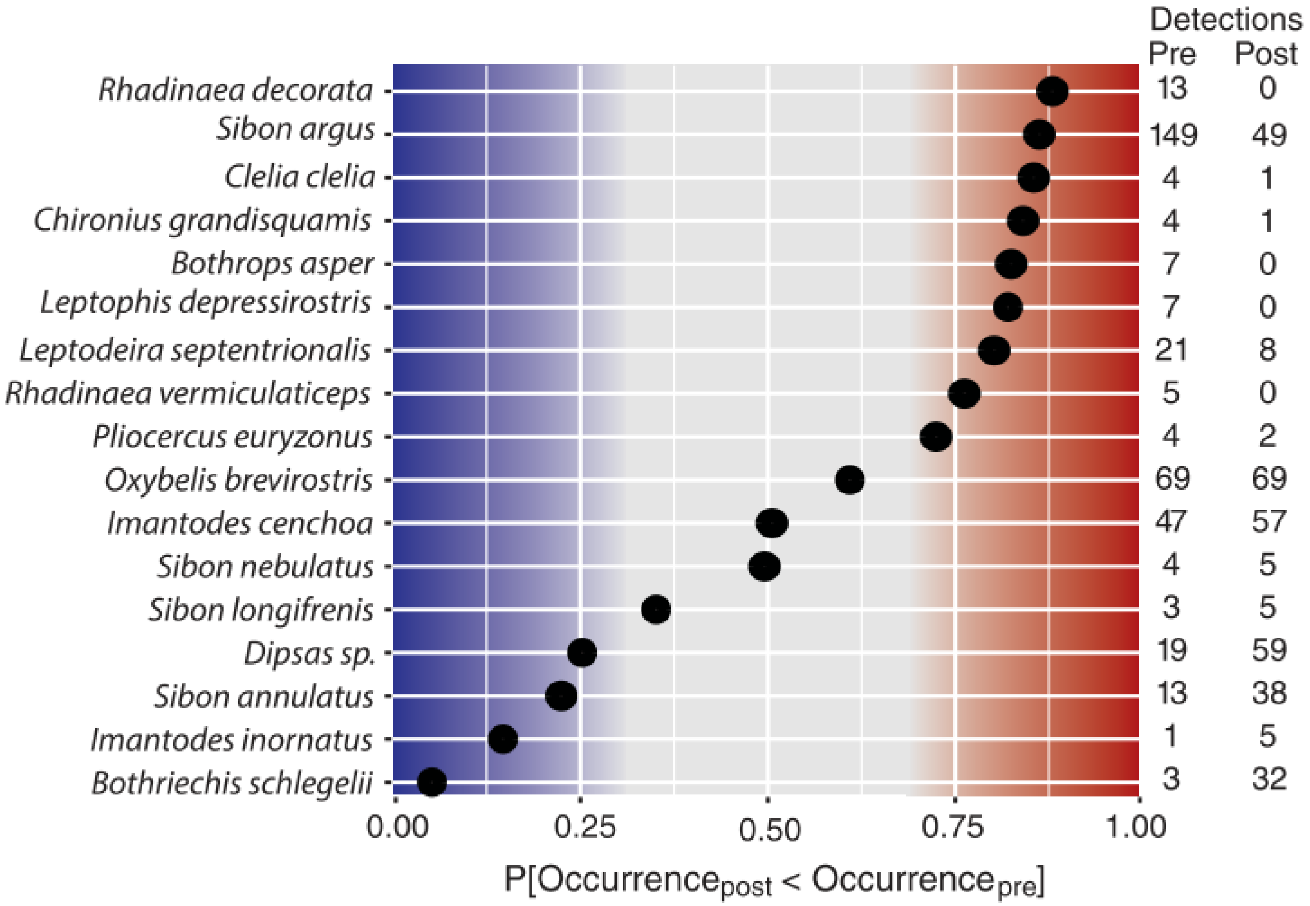
μ_{β_r} Mean effect of covariate across all species

$\tau^2_{\beta_r}$ Variance of the covariate effect among all species

Random slopes!!

Why multi-species occupancy modeling?

Improved ability to model rare species



Potential downsides

- Longer model run times
- Coding often involves working with multi-dimensional arrays (but spOccupancy simplifies this!)
- Defining a "community" is not always straightforward:
  - Pacifici et al. 2014 Ecology and Evolution
- May not be ideal for the rarest of the rare species:
  - Erickson and Smith, 2023 Ecography

Spatial multi-species occupancy models

Spatial autocorrelation in multi-species models

- Spatial autocorrelation may be more relevant in multi-species models since different species are driven by different variables.

Spatial autocorrelation in multi-species models

- Spatial autocorrelation may be more relevant in multi-species models since different species are driven by different variables.
- We could have a separate spatial random effect for each species. spMsPGOcc() function.

$$\text{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \cdots + \beta_{r,i} \cdot X_{r,j} + \boxed{\mathrm{w}_{i,j}}$$

Spatial autocorrelation in multi-species models

- Spatial autocorrelation may be more relevant in multi-species models since different species are driven by different variables.
- We could have a separate spatial random effect for each species. spMsPGOcc() function.

$$\text{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \cdots + \beta_{r,i} \cdot X_{r,j} + \boxed{\mathrm{w}_{i,j}}$$

- Each w is estimated using an NNGP as before
- Model run times become huge with even a moderate number of species (e.g., 10)

Alternative approach: spatial factor models

- Basic idea: species-specific residual spatial autocorrelation can be explained by a set of common underlying "factors"

Alternative approach: spatial factor models

- Basic idea: species-specific residual spatial autocorrelation can be explained by a set of common underlying "factors"
- View the factors as "missing covariates" with a spatial structure

Alternative approach: spatial factor models

- Basic idea: species-specific residual spatial autocorrelation can be explained by a set of common underlying "factors"
- View the factors as "missing covariates" with a spatial structure
- Each species has a unique effect of each "missing covariate"
    - Called a "factor loading"

Alternative approach: spatial factor models

- Basic idea: species-specific residual spatial autocorrelation can be explained by a set of common underlying "factors"
- View the factors as "missing covariates" with a spatial structure
- Each species has a unique effect of each "missing covariate"
  - Called a "factor loading"
- This is a form of "factor analysis" (similar to PCA)

Spatial factor models

- Example: one covariate and two factors ("missing covariates")

$$\text{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \lambda_{i,1} \cdot \mathrm{w}_{1,j} + \lambda_{i,2} \cdot \mathrm{w}_{2,j}$$

Spatial factor models

- Example: one covariate and two factors ("missing covariates")

$$\text{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \lambda_{i,1} \cdot \mathrm{w}_{1,j} + \lambda_{i,2} \cdot \mathrm{w}_{2,j}$$

"Missing covariates" that account for residual spatial autocorrelation

Spatial factor models

- Example: one covariate and two factors ("missing covariates")

$$\mathrm{logit}(\psi_{i,j}) = \beta_{1,i} + \beta_{2,i} \cdot X_{2,j} + \lambda_{i,1} \cdot \mathrm{w}_{1,j} + \lambda_{i,2} \cdot \mathrm{w}_{2,j}$$

Effects of the missing covariates

Spatial factor models

- Each factor is modeled as a spatial NNGP

Spatial factor models

- Each factor is modeled as a spatial NNGP
- Usually set the number of factors to be much smaller than the number of species
  - Guidance on the package website

Spatial factor models

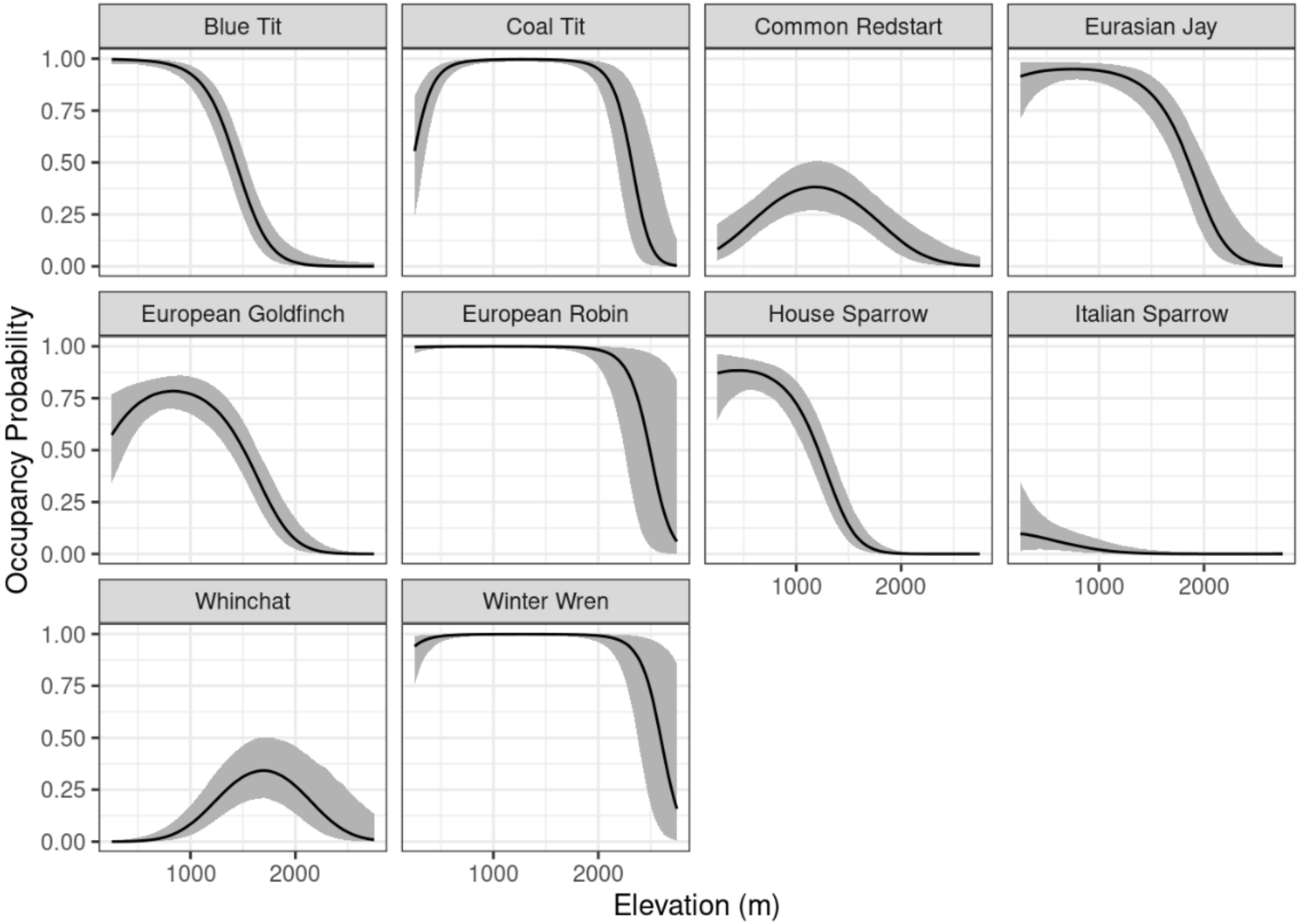
- Each factor is modeled as a spatial NNGP
- Usually set the number of factors to be much smaller than the number of species
  - Guidance on the package website
- Model run time increases with the number of factors
- See Doser et al. (2023) Ecology for details

Spatial factor models

- Each factor is modeled as a spatial NNGP
- Usually set the number of factors to be much smaller than the number of species
  - Guidance on the package website
- Model run time increases with the number of factors
- See Doser et al. (2023) Ecology for details
- Downsides
  - Convergence can be tricky (see linked vignette above)
  - Requires more data than non-spatial multi-species models

Exercise 2: Swiss songbirds

- Data from the Swiss Breeding Bird Survey
- 3 visits at 267 1km squares across Switzerland
- We will focus on 10 passerine species



Multi-season occupancy models

Ecological Motivation

- How are species distributions shifting across space and time?

Ecological Motivation

- How are species distributions shifting across space and time?
- Assessment of occupancy trends over time:
  - Detection-nondetection data are easier to collect than count data
  - Occupancy-abundance relationship
  - Exact interpretation of occupancy trends depends on how data are collected (Steenweg et al. 2018 *Ecology*)

Multi-season detection-nondetection data

- Data follow the "robust design"
- A set of J sites are sampled across a set of T seasons/years
- Within each season t, each site *j* is sampled $K_{j,t}$ times.

Multi-season detection-nondetection data

- Data follow the "robust design"
- A set of J sites are sampled across a set of T seasons/years
- Within each season t, each site *j* is sampled $K_{j,t}$ times.
- We will almost always work with an imbalanced data set
  - Each site may not be sampled each season
  - Each sampled site may not be sampled the same number of times within a season

Multi-season detection-nondetection data

- Data follow the "robust design"
- A set of J sites are sampled across a set of T seasons/years
- Within each season t, each site *j* is sampled $K_{j,t}$ times.
- We will almost always work with an imbalanced data set
  - Each site may not be sampled each season
  - Each sampled site may not be sampled the same number of times within a season
- Seasons are sometimes referred to as "primary replicates" and repeat visits within season as "secondary replicates"

Multi-species detection-nondetection data

- Example: 6 sites, 2 seasons, 3 surveys within a season

Season 1

Site	Survey 1	Survey 2	Survey 3
1	1	0	0
2	0	0	0
3	1	1	0
4	1	NA	0
5	0	1	1
6	0	0	0

Season 2

Site	Survey 1	Survey 2	Survey 3
1	0	1	NA
2	1	0	0
3	1	1	0
4	1	1	0
5	NA	NA	NA
6	0	0	1

Dynamic vs. multi-season occupancy models

Dynamic models

Multi-season models

Dynamic vs. multi-season occupancy models

Dynamic models

- Estimate colonization and survival/extinction

Multi-season models

- Estimate occupancy probability per season

Dynamic vs. multi-season occupancy models

Dynamic models

- Estimate colonization and survival/extinction
- More mechanistic

Multi-season models

- Estimate occupancy probability per season
- Less mechanistic

Dynamic vs. multi-season occupancy models

Dynamic models

- Estimate colonization and survival/extinction
- More mechanistic
- More data hungry and harder to fit

Multi-season models

- Estimate occupancy probability per season
- Less mechanistic
- Easier to fit and less data hungry

Dynamic vs. multi-season occupancy models

Dynamic models

- Estimate colonization and survival/extinction
- More mechanistic
- More data hungry and harder to fit
- Non-spatial versions in ubms package

Multi-season models

- Estimate occupancy probability per season
- Less mechanistic
- Easier to fit and less data hungry
- In spOccupancy

Dynamic vs. multi-season occupancy models

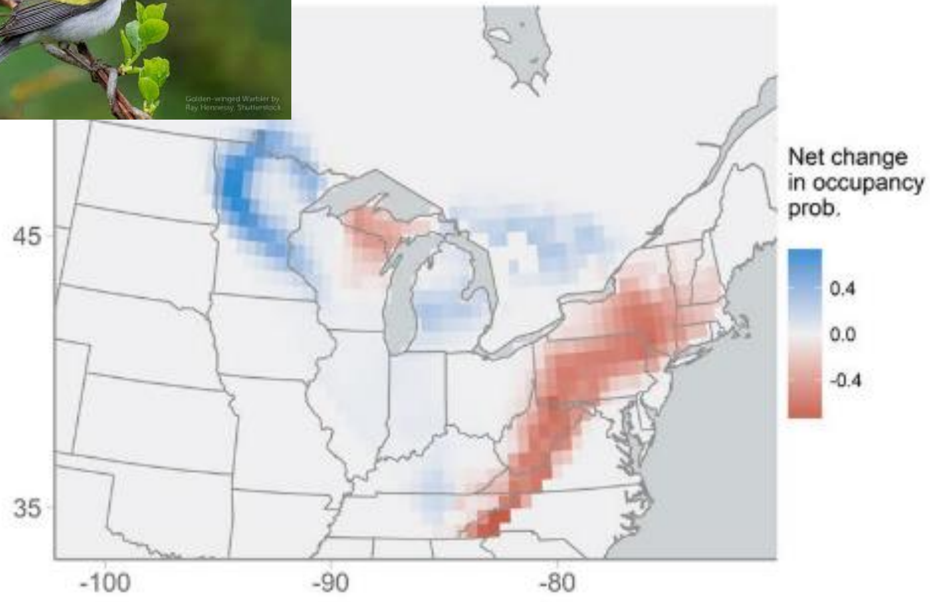
Dynamic models

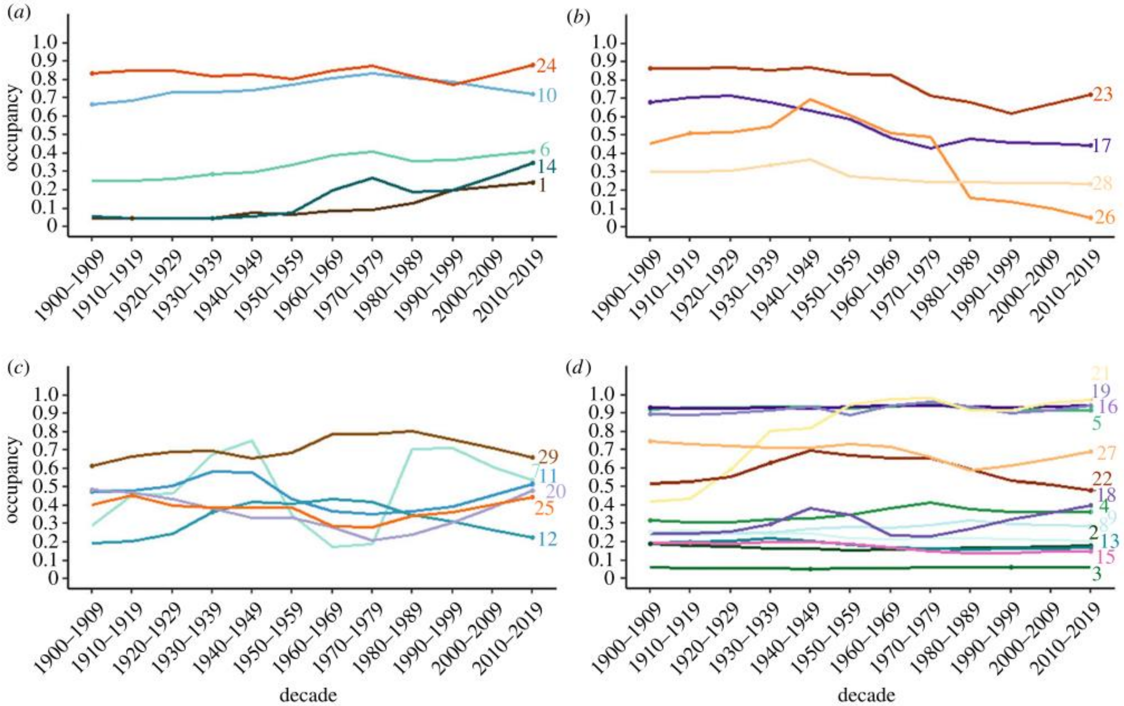
- Estimate colonization and survival/extinction
- More mechanistic
- More data hungry and harder to fit
- Non-spatial versions in ubms package
- MacKenzie et al. 2003 *Ecology*

Multi-season models

- Estimate occupancy probability per season
- Less mechanistic
- Easier to fit and less data hungry
- In spOccupancy
- Package vignette
- Sometimes called "stacked" occupancy models

Examples of multi-season occupancy models





[Rushing et al. (2019) Sci Rep]

[Rushing et al. (2020) PNAS]

[Sheard et al. (2021) Curr Bio]

Multi-season occupancy model

$j = 1, \dots, J$ (site)
$t = 1, \dots, T$ (season)
$k = 1, \dots, K_{j,t}$ (replicate)

Occupancy (ecological) sub-model

$$z_{j,t} \sim \text{Bernoulli}(\psi_{j,t})$$
$$\text{logit}(\psi_{j,t}) = \boldsymbol{x}_{j,t}\boldsymbol{\beta} + \text{w}_j + \eta_t$$

$z_{j,t}$ True occurrence of the species at site *j* in season *t*

$\psi_{j,t}$ Occurrence probability at site *j* in season *t*

$\boldsymbol{x}_{j,t}$ Site and/or season-varying covariates

w_j Site-level random effect

η_t Season-level (temporal) random effect

Multi-season occupancy model

$$j = 1, \dots, J \text{ (site)}$$
$$t = 1, \dots, T \text{ (season)}$$
$$k = 1, \dots, K_{j,t} \text{ (replicate)}$$

Occupancy (ecological) sub-model

$$z_{j,t} \sim \text{Bernoulli}(\psi_{j,t})$$
$$\text{logit}(\psi_{j,t}) = \boldsymbol{x}_{j,t}\boldsymbol{\beta} + \boxed{\text{w}_j + \eta_t}$$

$z_{j,t}$ True occurrence of the species at site *j* in season *t*

$\psi_{j,t}$ Occurrence probability at site *j* in season *t*

$\boldsymbol{x}_{j,t}$ Site and/or season-varying covariates

w_j Site-level random effect

η_t Season-level (temporal) random effect

Site-level random effects w_j

- Account for spatial autocorrelation in occupancy probability

Site-level random effects w_j

- Account for spatial autocorrelation in occupancy probability
- Alternative view: account for non-independence between occupancy probability at a site over the T seasons (i.e., pseudoreplication)

Site-level random effects w_j

- Account for spatial autocorrelation in occupancy probability
- Alternative view: account for non-independence between occupancy probability at a site over the T seasons (i.e., pseudoreplication)
- Two types:
    1. Unstructured -> a typical random intercept with the form:

$$\mathrm{w}_j \sim \mathrm{Normal}(0, \sigma^2)$$

Site-level random effects w_j

- Account for spatial autocorrelation in occupancy probability
- Alternative view: account for non-independence between occupancy probability at a site over the T seasons (i.e., pseudoreplication)
- Two types:
  1. Unstructured -> a typical random intercept with the form:

$$\mathrm{w}_j \sim \mathrm{Normal}(0, \sigma^2)$$

  2. Spatial NNGP -> same as before. This is the "spatial multi-season occupancy model"

Temporal random effects η_t

- Account for correlation in occupancy probability over time

Temporal random effects η_t

- Account for correlation in occupancy probability over time
- Alternative view: account for non-linear variation in occupancy probability when estimating a trend

Temporal random effects η_t

- Account for correlation in occupancy probability over time
- Alternative view: account for non-linear variation in occupancy probability when estimating a trend
- Two types:
    1. Unstructured -> a typical random intercept with the form:

$$\eta_t \sim \mathrm{Normal}(0, \sigma_T^2)$$

Temporal random effects η_t

- Account for correlation in occupancy probability over time
- Alternative view: account for non-linear variation in occupancy probability when estimating a trend
- Two types:
    1. Unstructured -> a typical random intercept with the form:

$$\eta_t \sim \mathrm{Normal}(0, \sigma_T^2)$$

    2. AR(1) -> random temporal effects follow an autoregressive structure. Covariance between two time points is:

$$\sigma_T^2 \rho^{|t-t'|}$$

Multi-season occupancy model

$$j = 1, \dots, J \text{ (site)}$$
$$t = 1, \dots, T \text{ (season)}$$
$$k = 1, \dots, K_{j,t} \text{ (replicate)}$$

Detection (observation) sub-model

$$y_{j,t,k} \sim \text{Bernoulli}(p_{j,t,k} \cdot z_{j,t})$$
$$\text{logit}(p_{j,t,k}) = \boldsymbol{v}_{j,t,k} \cdot \boldsymbol{\alpha}$$

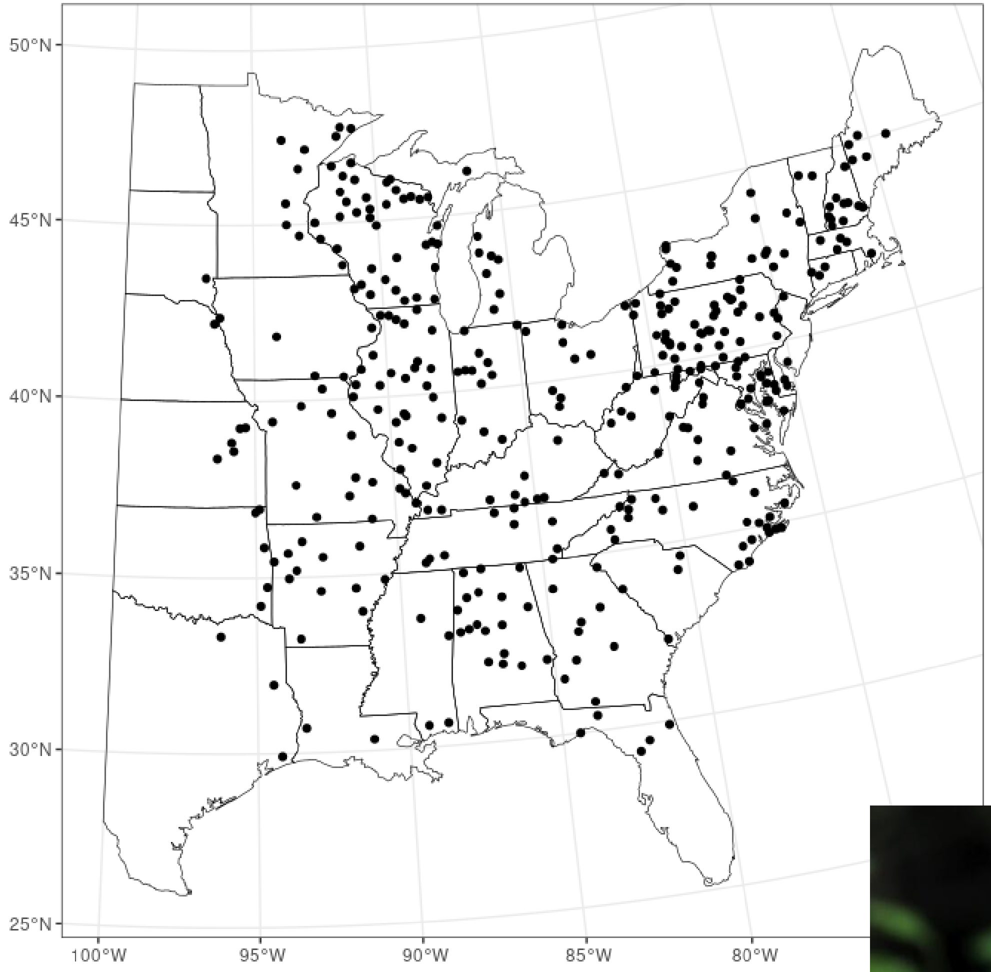
$y_{j,t,k}$ Detection-nondetection data at site *j* during replicate *k and season t*

$p_{j,t,k}$ Detection probability at site *j* during replicate *k and season t*

$\boldsymbol{v}_{j,t,k}$ Covariates affecting detection at site j during replicate k and season t

Exercise 3: Wood Thrush trend in eastern US

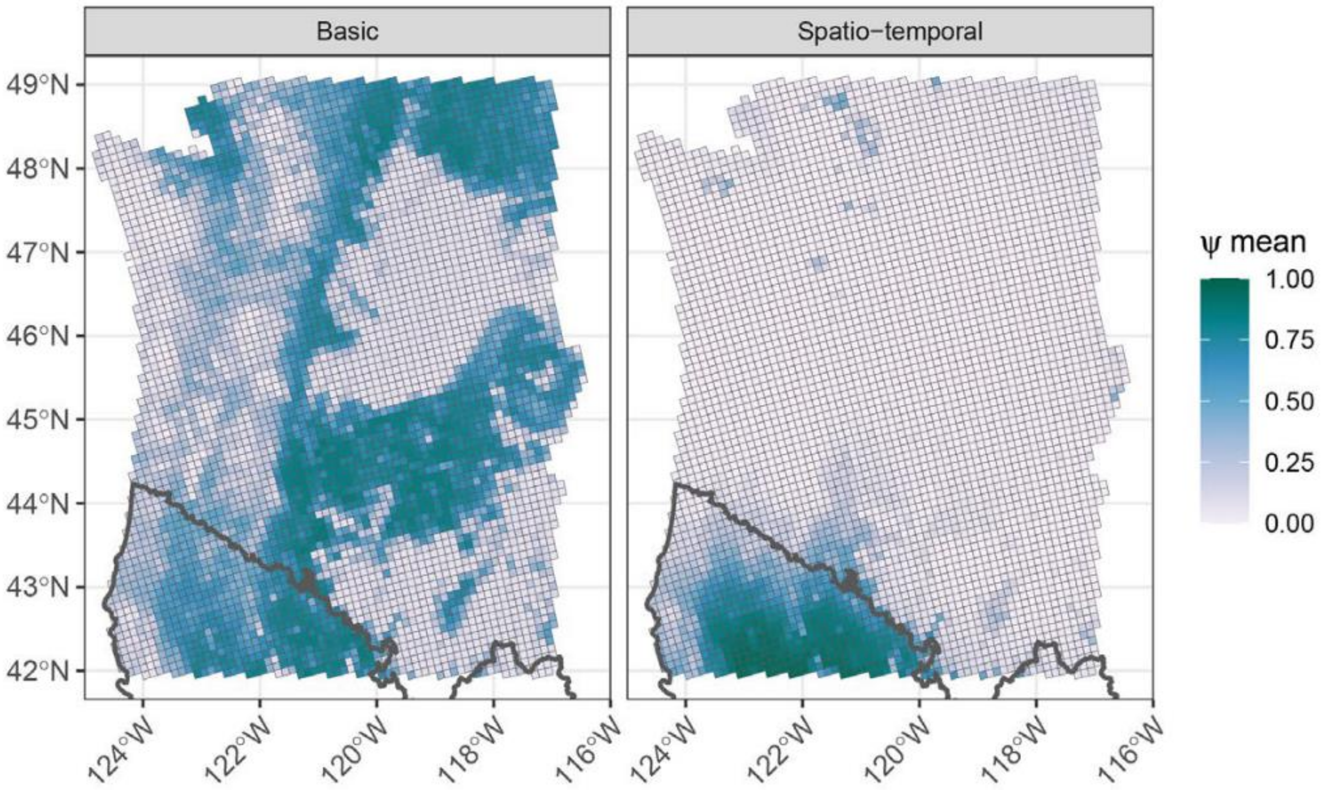
- Wood Thrush (*Hylocichla mustelina*) data from North American Breeding Bird Survey
- Replicates are spatial replicates (5 replicates per route)
- Each replicate is a group of 10 stops
- Data from 368 routes sampled in 2000-2009



Additional topics and resources

Multi-season multi-species occupancy models

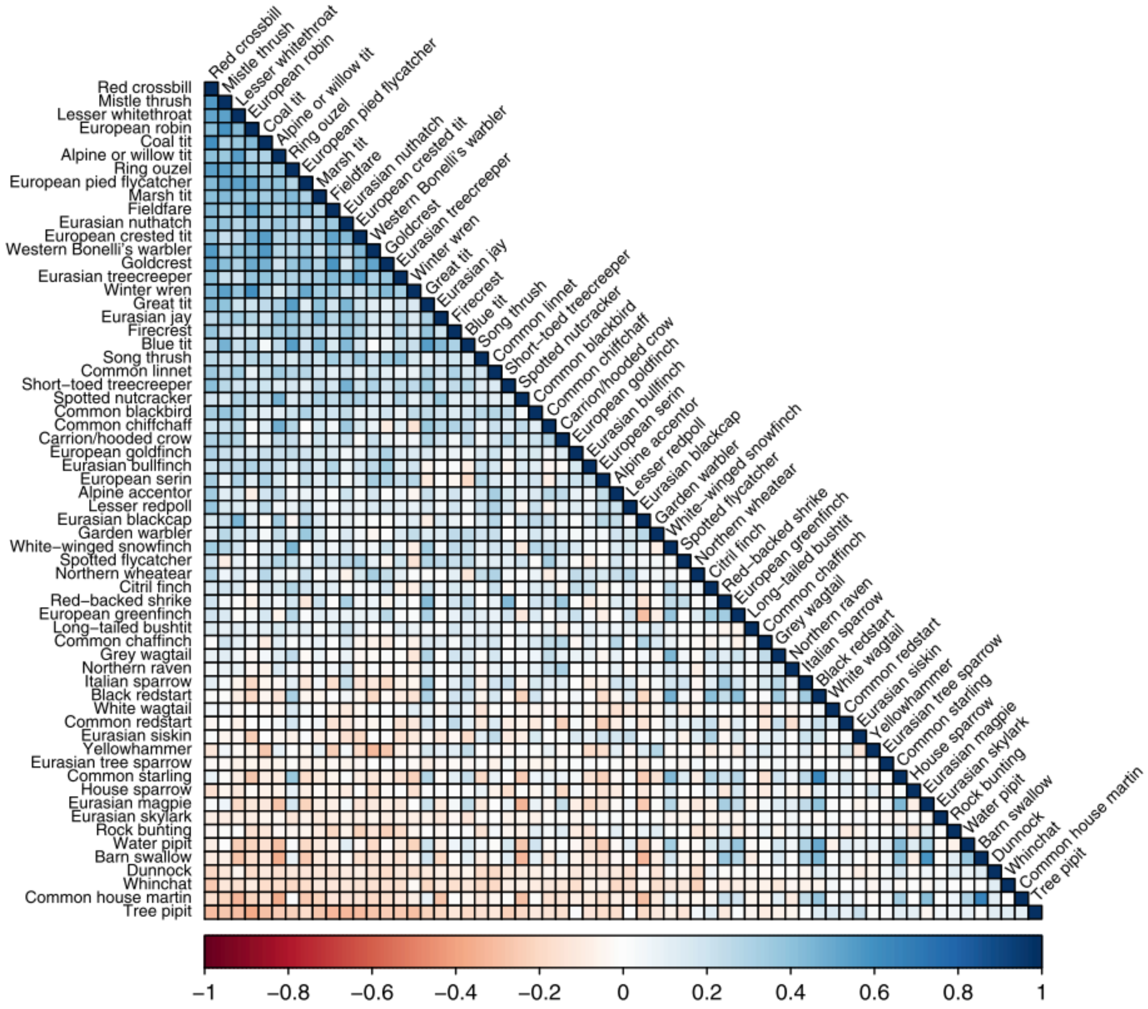
- Model spatio-temporal occupancy patterns for multiple species simultaneously
- Can help estimate trends for very rare species
- See functions tMsPGOcc() and stMsPGOcc()
- Data formatted in a four-dimensional array



[Wright et al. (2021) Eco and Evo]

Species correlations

- The factor modeling approach for multi-species models inherently accounts for residual species correlations (vignette)
- Can derive a species x species correlation matrix
- This is a spatially-explicit joint species distribution model (JSDM) with imperfect detection
- See lfMsPGOcc() function for a non-spatial JSDM



[Tobler et al. (2019) *Ecology*]

Spatially-varying coefficient occupancy models

- Allow the effects of covariates to vary spatially in addition to the intercept
- Applications: spatially-varying trends, quantify "nonstationarity" in covariate effects
- [Vignette]

Guidelines for the use of spatially-varying coefficients in species distribution models

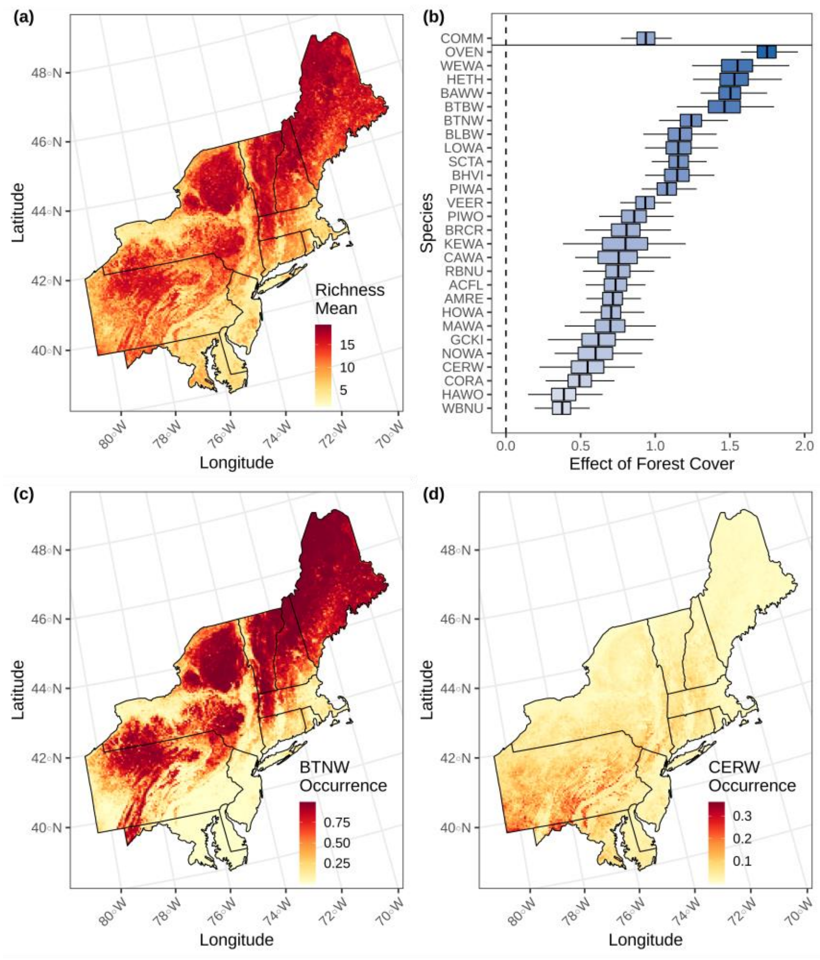
Jeffrey W. Doser[1, 2], Marc Kéry[3], Sarah P. Saunders[4], Andrew O. Finley[2,5,6], Brooke L. Bateman[4], Joanna Grand[4], Shannon Reault[4], Aaron S. Weed[7], Elise F. Zipkin[1, 2]

Modeling complex species-environment relationships through spatially-varying coefficient occupancy models

Jeffrey W. Doser[1, 2], Andrew O. Finley[2, 3, 4], Sarah P. Saunders[5], Marc Kéry[6], Aaron S. Weed[7], Elise F. Zipkin[1, 2]

Integrated occupancy models

- Fit occupancy models using multiple data sources
- Single-species: spatial and non-spatial models
- Multi-species: non-spatial models only (spatial coming soon)
- Examples:
  - Vignette for [single-species]() and [multi-species]()
  - [Code for single-species example with bottlenose dolphins]()
  - [Code for multi-species example with eBird and BBS data]()



[Zipkin et al. (2023) JAE]()

spAbundance

- Spatial and nonspatial N-mixture models, hierarchical distance sampling models, and GLMMs
- Single-species and multi-species models
- Syntax nearly identical to spOccupancy
- Website and preprint

Additional resources

Articles

Fit occupancy models

[Introduction to spOccupancy]
Learn how to get started with the core spOccupancy functionality

[Formatting data for use in spOccupancy]
Learn how to format raw data to fit occupancy models in spOccupancy

[Joint species distribution models with imperfect detection in spOccupancy]
Learn how to account for species correlations within multi-species occupancy models

[Multi-season occupancy models for assessing species trends and spatio-temporal occurrence patterns (PDF)]

[Multi-season occupancy models for assessing species trends and spatio-temporal occurrence patterns]
Learn how to fit multi-season occupancy models in spOccupancy

[Fitting occupancy models with random intercepts in spOccupancy]
Learn how to include random effects in spOccupancy

[Spatially varying coefficient models in spOccupancy]
Learn how to fit spatially varying coefficient models to quantify spatially varying trends and species-environment relationships

[Integrated multi-species occupancy models in spOccupancy]
Learn how to fit multi-species occupancy models with multiple data sources

[Convergence diagnostics and other considerations when fitting spatial occupancy models]
Ideas related to convergence, identifiability, priors, and other potential problem areas in spatial occupancy models

[Exploring model identifiability with a stress-testing framework]

- Website:
  - https://www.jeffdoser.com/files/spoccupancy-web/
- GitHub development page
  - https://github.com/doserjef/spOccupancy
- Package updates announced on Twitter/X (@jeffdoser18)
- Email: doserjef@msu.edu