Multispecies Occupancy Modeling

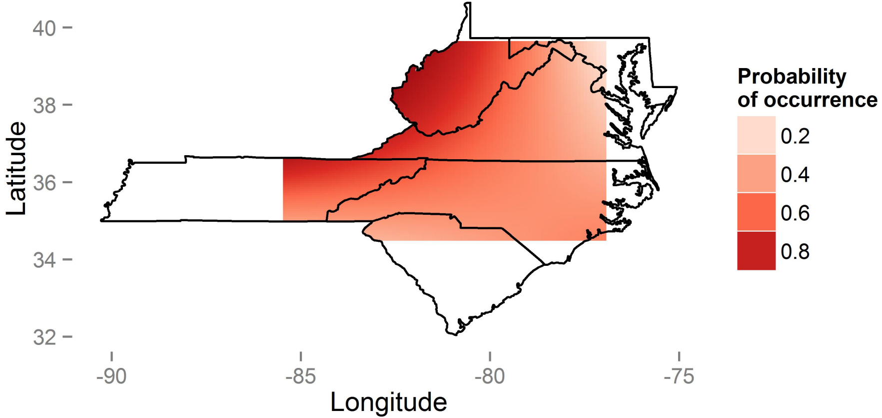
Christopher Rota
Arielle Parsons

West Virginia University
Morgantown, West Virginia
&
North Carolina State University
Raleigh, North Carolina

christopher.rota@mail.wvu.edu
http://bit.ly/RotaLab

October 7, 2018

Introduction



Understanding how species are distributed in space is a fundamental question in wildlife ecology.

Introduction

Species distributions are often driven by environmental variables.

Introduction

Species distributions can also be a function of interspecific interactions:

- Competitive exclusion
- Facilitation
- Predator-prey interactions

Introduction

- Multi-species occupancy models simultaneously model environmental variables and interspecific dependence (all while accounting for imperfect detection)
- This workshop focuses exclusively on Rota et al. 2016, *Methods in Ecology and Evolution*

Published Methods

2-species models Richmond et al. 2010, *Ecological Applications*
Waddle et al. 2010, *Ecological Applications*
Rota et al. 2016, *Ecology*

$>$ 2-species models Ovaskainen et al. 2010, *Ecology*
Pollock et al. 2014, *MEE*
Clark et al. 2014, *Ecological Applications*

Random-effects models Zipkin et al. 2010, *Biological Conservation*

Plus many more!

Sampling Scheme

Equivalent to single-species occupancy models:

- survey for S species at N sites selected from a population of interest
- at each site i, conduct J_i replicate surveys
- for each species s, during replicate survey j at site i, record detection ($y_{sij} = 1$) or non-detection ($y_{sij} = 0$)
- record detection level covariates during each replicate detection / non-detection survey
- record site-level covariates at each site i

Assumptions

Assumptions the same as single-season occupancy models. For example:

- Closure between replicate surveys
- No unmodeled heterogeneity in detection probability
- etc.

Model Description

As with single-season occupancy models, we assume detection / non-detection is a Bernoulli random variable, conditional on the presence of species s:

$$y_{sij} \sim \text{Bernoulli}(p_{sij} z_{si})$$

p_{sij}: probability of detection species s during replicate survey j at site i

$$z_{si}: \begin{cases} = 1 \text{ if species } s \text{ present at site } i \\ = 0 \text{ if species } s \text{ absent from site } i \end{cases}$$

Detection probability can modeled as a function of covariates:

$$p_{sij} = \text{logit}^{-1}(\boldsymbol{w}_{sij}\boldsymbol{\alpha})$$

Model Description

We model occupancy probability as a multivariate Bernoulli random variable:

$$\boldsymbol{Z_i} \sim \text{Multivariate Bernoulli}(\boldsymbol{\Psi}_i)$$

- $\boldsymbol{Z_i}$: a vector of length S with each element indicating presence ($z_{si} = 1$) or absence ($z_{si} = 0$) of species s at site i
- $\boldsymbol{\Psi}_i$: a 2^S dimensional vector indicating the probability of all possible realizations of Z_i, such that $\sum \boldsymbol{\Psi}_i = 1$

For example, when $S = 2$:

$$\begin{aligned} Pr(Z_i) = \{11\} &\Rightarrow \psi_{i11} \\ Pr(Z_i) = \{10\} &\Rightarrow \psi_{i10} \\ Pr(Z_i) = \{01\} &\Rightarrow \psi_{i01} \\ Pr(Z_i) = \{00\} &\Rightarrow \psi_{i00} \end{aligned}$$

Multi-species occupancy model with 1 species

A single-species occupancy model is really a multi-species occupancy model with $S = 1$. To demonstrate:

$$\boldsymbol{Z_i} \sim \text{Multivariate Bernoulli}(\boldsymbol{\Psi}_i)$$

$$\begin{aligned} Pr(Z_i) = \{1\} \quad &\Rightarrow \quad \psi_{i1} \\ Pr(Z_i) = \{0\} \quad &\Rightarrow \quad \psi_{i0} = 1 - \psi_{i1} \end{aligned}$$

The probability mass function (PMF) can be written as:

$$\begin{aligned} f(z_i) =& \psi_{i1}^{z_i}(1 - \psi_{i1})^{(1-z_i)} \\ =& \psi_{i1}^{z_i}\psi_{i0}^{1-z_i} \\ =& \exp\Big(z_i \log\big(\frac{\psi_{i1}}{\psi_{i0}}\big) + \log(\psi_{i0})\Big) \end{aligned}$$

Multi-species occupancy model with 1 species

From the previous slide, we can write the PMF as:

$$f(z_i) = \exp\Big(z_i \log\big(\frac{\psi_{i1}}{\psi_{i0}}\big) + \log(\psi_{i0})\Big)$$

Hopefully, the quantity "$\log\big(\frac{\psi_{i1}}{\psi_{i0}}\big)$" is recognized as the log odds of success. In binomial regression (including occupancy models, logistic regression, etc.), we model the log odds as a function of covariates:

$$\log\Big(\frac{\psi_{i1}}{\psi_{i0}}\Big) = \boldsymbol{x}_i \boldsymbol{\beta} = f_{i1},$$

and we model the *probability* of success with the logit link as:

$$\psi_{i1} = \frac{\exp(f_{i1})}{1 + \exp(f_{i1})}$$

Multi-species occupancy model with 2 species

Now, assume $S = 2$. We can write the PMF as:

$$
\begin{aligned}
f(\boldsymbol{Z}_i) =& \psi_{11}^{z_1 z_2} \psi_{10}^{z_1(1-z_2)} \psi_{01}^{(1-z_1)z_2} \psi_{00}^{(1-z_1)(1-z_2)} \\
=& \exp\left(z_1 \log\left(\frac{\psi_{10}}{\psi_{00}}\right) + z_2 \log\left(\frac{\psi_{01}}{\psi_{00}}\right) + z_1 z_2 \log\left(\frac{\psi_{11}\psi_{00}}{\psi_{10}\psi_{01}}\right) + \log(\psi_{00}) \right)
\end{aligned}
$$

yuck ... what do we learn from this?

Natural Parameters

Hopefully, you see that we now have 3 natural parameters, all of which we can model as a function of covariates:

$$
\begin{aligned}
f_1 =& \log \frac{\psi_{10}}{\psi_{00}} = \boldsymbol{c_i}\gamma \\
f_2 =& \log \frac{\psi_{01}}{\psi_{00}} = \boldsymbol{d_i}\delta \\
f_{12} =& \log \frac{\psi_{11}\psi_{00}}{\psi_{10}\psi_{01}} = \boldsymbol{e_i}\epsilon
\end{aligned}
$$

Deriving probability of occurrence from natural parameters

And from these natural parameters, we can obtain the probability of all combinations of 1s and 0s via the multinomial logit link:

$$
\begin{aligned}
\psi_{11} =& \frac{\exp(f_1 + f_2 + f_{12})}{1 + \exp(f_1) + \exp(f_2) + \exp(f_1 + f_2 + f_{12})} \\
\psi_{10} =& \frac{\exp(f_1)}{1 + \exp(f_1) + \exp(f_2) + \exp(f_1 + f_2 + f_{12})} \\
\psi_{01} =& \frac{\exp(f_2)}{1 + \exp(f_1) + \exp(f_2) + \exp(f_1 + f_2 + f_{12})} \\
\psi_{00} =& \frac{1}{1 + \exp(f_1) + \exp(f_2) + \exp(f_1 + f_2 + f_{12})}
\end{aligned}
$$

Interpretation of natural parameters

In general, there will be $2^S - 1$ natural parameters, all of which can be modeled as a function of covariates. These can be divided into $1^{st}, 2^{nd}, ..., S^{th}$ order natural parameters.

Order	Interpretation
1^{st}	log odds species s occurs, conditional on absence of all other species
2^{nd}	difference in log odds species s occurs when another species is present and absent
S^{th}	difference in log odds all S species occur together from log odds that not all species occur together

Fixing 2^{nd} order and higher natural parameters at 0 assures independence between species at that order.

Independence between species

We can fix pairwise (or higher) independence between species by fixing 2^{nd} order (or higher) natural parameters at 0.

For example, a 3-species model has $2^3 - 1 = 7$ natural parameters:

$$1^{st} \text{ order: } \begin{cases} f_1 \\ f_2 \\ f_3 \end{cases}$$

$$2^{nd} \text{ order: } \begin{cases} f_{12} \\ f_{13} \\ f_{23} \end{cases}$$

$$3^{rd} \text{ order: } \{ \; f_{123}$$

Fixing f_{12}, f_{13}, f_{23}, and $f_{123} = 0$ assures independence among all 3 species and is equivalent to fitting 3 single-species occupancy models.

Independence between species

Alternatively, we can estimate dependence between species by estimating 2^{nd} order (or higher) natural parameters.

$$1^{st}\text{ order: } \begin{cases} f_1 \\ f_2 \\ f_3 \end{cases}$$

$$2^{nd}\text{ order: } \begin{cases} f_{12} \\ f_{13} \\ f_{23} \end{cases}$$

$$3^{rd}\text{ order: } \{ \quad f_{123}$$

For our 3-species example, we can estimate pairwise dependence by estimating $f_{12}, f_{13},$ and f_{23}.

Relation to other multi-species occupancy models

- Ability to model interactions in more detail ...
- ... which can come at the expense of many more parameters
- In principle, can handle arbitrary number of species, but in practice may face computational limits
- This method may be better suited to smaller numbers of species, for which more detailed information on interactions is desired
- Other covariance matrix based methods may be better suited to larger species assemblages

Challenges

- MARK good for relatively simple models
- May need to use restricted maximum likelihood (I've not used) or Bayesian techniques to fit more complicated models (to shrink coefficients toward 0)
- Model selection – how to choose a parsimonious model when you can fit 10+ linear models?
- Will be implemented soon in Unmarked.