Montana State University

Department of Mathematical Sciences

Writing Project

TITLE

Author:
Meaghan Winder

Supervisor:
Dr. Andrew Hoegh

Spring 2020

A writing project submitted in partial fulfillment
of the requirements for the degree

Master's of Science in Statistics

APPROVAL

of a writing project submitted by

Meaghan Winder

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Andrew Hoegh Writing Project Advisor
Date	Mark C. Greenwood Writing Projects Coordinator

Contents

Abstract

abstract text here

free write... this is very messy, basically just word vomit

This is all information extracted from WILD 502 course material or things I learned from the class

Occupancy is the presence of a particular species on a given site, this may not be the first choice of state variables to ecologists but occupancy studies are useful when there is a large spatial scale or the study is conducted over many years, when abundance or vital rates are hard to measure. Occupancy studies are also useful over capture-recapture methods when individuals cannot be marked or uniquely identified. However, sometimes patterns of species occurrence are of interest, this happens when researchers are interested in the range of a species or the spread of invasion.

The sampling units for occupancy studies are called 'sites'. We can learn about detection probabilities when multiple site visits are used. Also, when using occupancy models we need to account for imperfect detection because it is possible that the researchers could miss the species even if it is present at the site. ψ represents the occupancy probability, p_i represents the probability of detecting the species on survey i given that the species occupies the site, and $p^* = 1 - \prod_{i=1}^{t}(1 - p_i)$ is the probability of detecting the species at least one time given the species occupies the site.

The assumptions are:

- The occupancy state of sites is constant during a single season.

- The occupancy probability is constant across sites, or is modeled appropriately using site-level covariates.
- The probability of detection given occupancy status is constant across sites, or modeled appropriately using site-level covariates.
- The species is not misidentified, no false positives.

As suggested above, site-level covariates can be used to model the occupancy probabilities and the detection probabilities.

In WILD 502 when talking about multi-season occupancy models, we talked about extirpation and colonization rates, but I think that these could be modeled with a latent variable(s)? I don't think they are of particular interest here.

In this case, if we consider site to be the lake then we have replication at the site level, but if each site is the sample location within the lake then there is replication for some (very few) on different dates.

This is what I used for my project proposal... Mark said this could make it's way into our abstract

Zebra mussels are a highly invasive species that have several negative impacts, both ecologically and economically, as they suffocate native mussels and cost millions of dollars to remove from man-made structures such as power plants (USGS). Zebra mussels have been established in the United States since the 1980s, and since then "they have spread rapidly throughout the Great Lakes region and into the large rivers of the eastern Mississippi

drainage. They have also been found in Texas, Colorado, Utah, Nevada, and California" (USGS). As the threat of zebra mussels spreading into the northwestern United States grows, one concern is that the detection method used to find zebra mussels is not highly effective. The current method of sampling are plankton tows, which are essentially large nets with fine mesh, that are towed behind boats to capture zebra mussel veligers (larval stage of zebra mussels); this method only works to capture veligers because once the zebra mussels mature, they attach to substrate and are no longer able to be captured by the tows. Researchers typically go to several sites on a given day within a water body and take multiple plankton tows (usually 5). The contents of the tows is aggregated for each site, then taken to a lab, where scientists use a microscope to examine the contents for presence (sometimes counts) of zebra mussel veligers, if the veligers are not detected with the microscope, the scientists use Polymerase chain reaction (PCR) to test for presence of zebra mussels veligers. A large problem with this method, is that there is a non-zero probability that the zebra mussels are in the water body and not captured with the tows, or in the tows and not identified with the microscope or with PCR.

I have also been working on fitting models to simulated occupancy data (see code at `https://github.com/meaghanwinder/Writing-Project/blob/master/Writing%20Project%20Code.Rmd`)

- constant $\psi = 0.6$ and constant $p = 0.6$ with constant $J = 8$ samples per $M = 10$ sites – DONE

- constant $\psi = 0.6$ and constant $p = 0.6$ with j_i samples per $M = 10$ sites – DONE
- ψ and p both depend on a single site level covariate, x, with j_i samples per $M = 10$ sites

Things to think about for the simulation or questions I have:

- think about how changing the detection probabilities and occupancy probabilities impact the results:
  - High occupancy, high detection
  - High occupancy, low detection
  - Low occupancy, high detection
  - Low occupancy, low detection
- How do we/can we include sample level covariates to account for sampling effort (number of tows, if available)?
- If we are getting more data:
  - How do we account for multiple sampling seasons (years)?
  - How do the assumptions of occupancy models change when we have several sampling years versus only 1

Additional questions, not directly related to the simulation:

- How do we account for this multilevel (for lack of a better word) testing process? For example, sometimes (?) when they fail to find them in the microscope, then they test them using polymerase chain reaction (PCR), and sometimes gene sequencing (? – looks like it was only used on one observation in the sample data), or scanning electron microscopy (SEM).

- One assumption of occupancy models is that there is a non-zero probability of mis-identification. I read an article (Denise Holser) that suggests that there might be an issue with that assumption in this sitatuion because there are similar looking organisms that may be present in the waters; she then suggest that the Bureau of Reclaimation has attempted to mediate this issue with improved microscopic methods and improved PCR methods. I will look into this some more.

1 Introduction

1.1 Motivation

2 Data

This could maybe be included in the introduction section... unless I do EDA here, then probably leave this as its own section

3 Modeling Background

This whole section might be able to be merged with the methods section

3.1 Occupancy Models

$y_{i,j} = \{0, 1\}$ is the binary response for the j^{th} sample within site i

If the site is occupied: ¿ $y_{i,j} \sim Bernoulli(p)$ ¿ $y_i \sim Binomial(J, p)$ where y_i is the total number of detections from the J samples in the i^{th} site

If the site is unoccupied: ¿ $y_{i,j} = 0$ with probability 1, since we assume there are no false detections

3.2 Bayesian Modeling Background

4 Methods

Package options for Multi-season single-species occupancy models:

- nimble.dynamic.occ
- STAN
- JAGS
- Frequentist Options:
  - unmarked
  - Program MARK

- write my own package

5 Analysis

Could do EDA here

6 Conclusion

6.1 Further Investigations

7 References

8 Appendix - R Code

A script containing all code used for this analysis is available at

github link here(?)... either that or include all code here

```
# SIMULATED DATA 1
set.seed(1202020)
p <- 0.6 # constant detection probability... could change to depend on covariate
psi <- 0.6 # constant occupancy probability... could change to depend on covaria
M <- 10 # number of sites
J <- 8 # constant number of samples per site... does not need to be constant

z <- rbinom(M, 1, psi) # site occupancy

y1 <- matrix(NA, nrow = M*J, ncol = 4)
y1[, 1] <- rep(1:M, each  = J) # column indicating site
y1[, 2] <- rep(1:J, M) # column indicating sample within site
y1[, 3] <- rep(z, each = J) # column indicating true site occupancy

# creates a column of 1's and 0's indicating whether the species was detected
for(i in 1:(M*J)){
    y1[i, 4] <- rbinom(1, 1 , p*y1[i, 3]) # if the site is not occupied the speci
}

colnames(y1) <- c("Site", "Sample", "True Occupancy", "Detected")

library("car")
some(y1) # view 10 sample rows of the simulated data

# or

y2 <- rbinom(M, J, p*z) # total number of detections for the J samples within ea
# SIMULATED DATA 2
set.seed(1222020)
p <- 0.6 # constant detection probability... could change to depend on covariate
psi <- 0.6 # constant occupancy probability... could change to depend on covaria
```

```r
M <- 10 # number of sites
J <- sample(1:10, M, replace = T) # number of times each of the sites were sample

z <- rbinom(M, 1, psi) # site occupancy

y <- rep(NA, M)

for(i in 1:M){
  y[i] <- rbinom(1, J[i], p*z[i]) # total number of detections for the J samples
}
#SIMULATED DATA 3
set.seed(1222020)
M <- 10 # number of sites
x <- runif(10, 0, 10)
beta1.true <- 2
beta2.true <- 0.5
beta3.true <- 0.5
beta4.true <- 0.2
p <- exp(beta1.true - beta2.true*x)/(1 + exp(beta1.true - beta2.true*x)) # detect
psi <- exp(beta3.true + beta4.true*x)/(1 + exp(beta3.true + beta4.true*x)) # occu
J <- sample(1:10, M, replace = T) # number of times each of the sites were sample

z <- rep(NA, M)
y <- rep(NA, M)

for(i in 1:M){
  z[i] <- rbinom(1, 1, psi[i]) #site occupancy
  y[i] <- rbinom(1, J[i], p[i]*z[i]) # total number of detections for the J sampl
}
```