Population Ecology BIOS 4770/5770, Fall 2019 – Viorel Popescu

OCCUPANCY MODELING

Occupancy models have gained prominence since the seminal paper by Mackenzie et al. (2002) in Ecology, and the least known but equally important paper by Tyre et al. (2003) in Ecological Applications (see below) that articulated a category of models that explicilty incorporated and predicted detection probabilities for a single spcies during a single season. Since then, the methodological developments of occupancy modeling have been exponential, and extended to include multiple species, multiple seasons, multiple states (e.g., reproducing), multiple spatial scales, count data, and the applications have been increasingly complex (see new book by Mackenzie et al 2018).

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. Andrew Royle, and C. A. Langtimm. 2002. *Estimating site occupancy rates when detection probabilities are less than one.* Ecology 83:2248–2255. Tyre, A. J., B. Tenhumberg, S. A. Field, D. Niejalke, K. Parris, and H. P. Possingham. 2003. *Improving precision and reducing bias in biological surveys: Estimating false-negative error rates.* Ecological Applications 13:1790–1801. MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. Hines. 2018. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence.* 2nd edition. Academic Press, Boston.

There is huge hype around this framework, but what do these models do? The theory is strikingly simple despite the complex math. Basically, by surveying a single location multiple times within a *season* (defined based on species ecology and life history), and building a history of species detection/non-detection at a site, one can explicitly evaluate the probability of detecting the species (this is what we call the **observation model**, and it is essentially a logistic regression). This detection probability is then incorporated into the **state model**, which evaluates the probability of a species occurring at a site (this is just another logistic regression). Ideally, if the variabes used to model **detection probability** (e.g., time of day, time of year, weather during the survey, etc.) and **occurrence probability** (e.g., site-scpeifc variables such as land cover type, elevation, habitat structure, etc.), then the predicted **occupancy** (the proportion of sites occupied by the target species) should be higher than the *naive* (or observed) occupancy. This is because the model can differentiate between absence of a speices and non-detection as inferred from the detection history and its associated variables. For example, a site may happen to be surveyed (by chance) during times when detection of the species is low; as such, the species is never detected, but the species occupies the site. If relevant data on survey timing and site characteristics is collected and integrated into the model, then that site will be predicted to be occupied, as the site characteristics would indicate that the species should be there, while the survey variables would indicate that surveys were done at the wrong time.

Next, we will implement 2 scripts tha will introduce you to occupancy modeling using package *unmarked* (Fiske and Chandler 2011) for program R.

Fiske, I., and R. Chandler. 2011. *unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance.* Journal of Statistical Software 43:1–23. https://cran.r-project.org/web/packages/unmarked/vignettes/unmarked.pdf

LET'S START CODING!!!

The objectives of this exercise are: 1) to introduce you to simulating occupancy data 2) to set up occupancy data for package 'unmarked' 3) to run simple (single-season) occupancy models and evaluate outputs

First, install the R packages that we will use to implement occupancy models

```
install.packages(c("unmarked", "MuMIn"))

library(unmarked)
library(MuMIn)
```

DATA SIMULATION AND ANALYSIS

The best way to understand how modeling works in general is to simulate your own data, then run predictive models and evaluate the outputs. Because we know what to expect from simulated datasets (in this case, the TRUE occupancy of sites, the probbaility of detection), this exercise is a great way to both understand occupancy data structure, and how to evaluate the outputs of complex models.

First, we will simulate occupancy data, with the aim to understand the structure of single-season occupancy data, and evaluate patterns in simulated data

We will set up our sampling regime (number of visits, number of sites) and an empty vector of 'true' occupancy (**STATE PROCESS**) name *tocc* (stands for true occupancy).

```
n_sites=100
n_vis = 4

tocc <- rep(0, n_sites)
tocc <- as.data.frame(tocc)
names(tocc) <- c("T1")
tocc
```

Next, we set up an empty matrix for our detection/non-detection field observations (**OBSERVATION PROCESS**)

```
obsocc <- matrix(rep(0, n_sites*n_vis), ncol=(n_vis),
                 dimnames=list(paste("site", 1:n_sites, sep=""),
                 paste("v", 1:n_vis, sep="")))
obsocc <- as.data.frame(obsocc)
obsocc
```

Now let's set values for psi (occupancy) and p (detection). We will tweak these variables for future simulations. Let's consider that psi = 0.7 (in other words, around 70% of the 100 sites are occupied), and p = 0.6 (detection probability is decently high, meaning that 4 survyes may be enough to conclude whether a species is absent with a 95% confidence; *see Table on slide 30 of Powerpoint lecture*).

```
psi = 0.7
p = 0.6
```

We are now ready to SIMULATE OCCUPANCY DATA, and fill the observed occupancy matrix. We will use *for loop* that iterates through the sites and assigns a 0 (absent) or 1 (present) for true occupancy based on random draws from a binomial distribution with a probability of psi=0.7. For sites that are occupied, the observations are also assigned a 0 (non-detection) or a 1 (detection) based on random draws from a binomial distribution with a probability p=0.6.

```
for(i in 1:n_sites) {
  tocc[i,1] = rbinom(1, 1, psi)                    # Simulate true site occupancy
  obsocc[i,1:n_vis] = rbinom(n_vis, 1, tocc[i,1]*p)  # Simulate observed data
}
```

Let's examine the simulated probability of occupancy, differences (if any) between true and observed occupancy. Are the results intuitive?

```
sim_psi = sum(tocc)/n_sites
sim_psi

# examine the true and observed occupancy
tocc
sum(tocc)

obsocc
sum(obsocc)
```

Let's now assume that the 100 sites are distributed between 2 regions, G and B. WE need to declare a new variable (Region) that we can then use in our model for predicting occupancy. In this case, because we simulated the data randomly across the 100 sites, declaring this is moot, as very likely we will not find differences between the 2 Regions. However, it is a good exercise to see how the data needs to be formatted for *unmarked* (similar to the datasets for marked animals, *unmarked* will take the data and create a new dataframe, called *dataUMF* in our case that contains all the pieces of the data that can be used in the models). The new object has 3 input requirements: (1) observed data (*obsocc*), (2) site-specific covariates used to model occupancy, and (3) survey-specific covariates used to model detection (we will not use this one for now since we did not simulate variables for modeling detection).

```
Region <- c(rep("G",50), rep("B",50))

dataUMF <-  unmarkedFrameOccu(
  y = obsocc,
  siteCovs = data.frame(Region)
  #obsCovs = list()
)

plot(dataUMF)
str(dataUMF)
```

We are now ready to run an occupancy model with 2 components, detection and occupancy in this exact order. We will run a Null model (intercept only for both detection and occupancy), and a more complex model, with Region as a covariate for occupancy (i.e., to evaluate if there is a difference in occupancy between the two Regions).

```
model.Null = occu(~1 ~1, data=dataUMF)
summary(model.Null)

model.1 = occu(~1 ~Group, data=dataUMF)
summary(model.1)
```

Of course, we find that there is no difference between the two regions in terms of occupancy, since the data was simulated at random across the 100 sites. Next, let's simulate differences between the 2 regions in terms of occupancy, and see if the models are able to pick up the differences. We repeat the process of creating observed data, and then we merge the 2 datasets (each corresponding to a Region) and run models with and without Region.

```
# We will simulate another 100 sites with psi = 0.3 and p=0.3

tocc1 <- rep(0, n_sites)
```

```r
tocc1 <- as.data.frame(tocc1)
names(tocc1) <- c("T1")
tocc1

### set up an empty observed occupancy matrix (OBSERVATION PROCESS)
obsocc1 <- matrix(rep(0, n_sites*n_vis), ncol=(n_vis),
                  dimnames=list(paste("site", 1:n_sites, sep=""),
                  paste("v", 1:n_vis, sep="")))
obsocc1 <- as.data.frame(obsocc1)
obsocc1

##declare new values for psi (occupancy) and p (detection)
psi1 = 0.3
p1 = 0.3

# fill the observed occupancy matrix
for(i in 1:n_sites) {
  tocc1[i,1] = rbinom(1, 1, psi1)                      # Simulate site occupancy
  obsocc1[i,1:n_vis] = rbinom(n_vis, 1, tocc1[i,1]*p1) # Simulate observed data
}

tocc1
sum(tocc1)

obsocc1
sum(obsocc1)

# put together the 2 datasets: tocc and tocc1; obsocc and obsocc1

tocc2 = rbind(tocc, tocc1)
tocc2

obsocc2 = rbind(obsocc, obsocc1)
obsocc2

# create new object for unmarked and run models
# we declare the obsocc data as Region G and obsocc1 data as Region B
Group2 <- c(rep("G",100), rep("B",100))

dataUMF2 <-  unmarkedFrameOccu(
  y = obsocc2,
  siteCovs = data.frame(Group2),
  #obsCovs = list()
)

plot(dataUMF2)
str(dataUMF2)

# run occupancy models

# Null model
model.Null = occu(~1 ~1, data=dataUMF2)
summary(model.Null)
```

```r
# psi as a function of Group
model.1 = occu(~1 ~Group2-1, data=dataUMF2)
summary(model.1)

# psi and p as a function of Group
model.2 = occu(~Group2-1 ~Group2-1, data=dataUMF2)
summary(model.2)
```

To evaluate which model had the best fit to the data, we create a model selection table (using package *MuMIn*)

```r
# create model selection table
modlist <- list(Null = model.Null,
                psiGroup = model.1,
                psiGroup.pGroup = model.2)

# create model selection table
selection = model.sel(modlist)
selection
```

Which model had the best fit to the data? Does it make sense?

Now, let's extract the model coefficients from the best model and contrast them to the initial **psi** and **p** values that we declared for each Region of study. We use function *coef* to extract the coefficients and function *plogis* to convert them back to probabilities

```r
plogis(coef(model.2))

sum(tocc) # sites occupied in Region G
sum(tocc1) # sites occupied in Region B
```

END OF OCCUPANCY SIMULATIONS

ANALYSIS OF AN OCCUPANCY DATASET

After simulating and analyzing simulated data, let's analyze an occupancy dataset of mink frog (Lithobates septentrionalis) occurrence from the Adirondack Mountains, New York State

The specific objectives of this document are: To document the data management and occupancy analysis process; To conduct a simple occupancy analysis in R using the unmarked package; and To explore occupancy predictions and prepare the fondation for multi-year (dynamic) occupancy models

```r
setwd('C:/Users/popescu/Dropbox/OhioU/teaching/PopEco_Fall2019/occupancy')

mink <- read.csv("minkfrogs_occ.csv", header=T)

names(mink)
str(mink)
```

```r
# we need to 'attach' the file for unmarked to work; usually it is not
# recommended to use this function
attach(mink)

library(unmarked)
library(MuMIn)
```

First, calculate the naive occupancy (not adjusted for detection probability) by dividing the number of sites where species was detected to the total number of sites

```r
minkfrog_naive <- sum(apply(mink[,3:6], 1, sum) > 0) / nrow(mink)
minkfrog_naive
```

Create an R object that can be read by the 'unmarked' package. The object contains 3 types of data: 1) detection/nondetection data (the observetions grouped into a matrix "y") 2) site-level covariates; these can be any variables that do not vary with survey occasion, such as landscape characteristics, habitat types, site characteristics etc. - these variables are used to model OCCUPANCY; they can also be used to model DETECTION (not shown here) 3) observation-level coveriates; these can be any variables recorded at the time of survey that may influence detection - these variables are used to model DETECTION

```r
minkUMF <- with(mink, {
  unmarkedFrameOccu(

    y = cbind(LISE1, LISE2, LISE3, LISE4),

    siteCovs = data.frame(scale(JulyTemp), scale(DO), scale(DistMFbreed), Beaver),

    obsCovs = list(Date = scale(cbind(Julian1, Julian2, Julian3, Julian4)),
                   Wind = cbind(Wind1, Wind2, Wind3, Wind4),
                   Sky = cbind(Sky1, Sky2, Sky3, Sky4)))
})
```

The analysis has 2 stages: First we need to evaluate the best predictors for detection. Therefore, we run several models that contains the same suite of site-related variables (for modeling occupancy), but various combinations of survey-specific variables for modeling detection. The goal is to identify the best variables that explain detection probbaility, when then use them in the next step of the analysis.

```r
mink.d1 = occu(~Date + Wind + Sky ~JulyTemp + DO + DistMFbreed, data=minkUMF)
#summary(mink.d1)

mink.d2 = occu(~Date + Sky ~JulyTemp + DO + DistMFbreed, data=minkUMF)
summary(mink.d2)

mink.d3 = occu(~Date  ~JulyTemp + DO + DistMFbreed, data=minkUMF)
#summary(mink.d3)

mink.d4 = occu(~Sky  ~JulyTemp + DO + DistMFbreed, data=minkUMF)
#summary(mink.d4)

mink.d5 = occu(~Date + Wind  ~JulyTemp + DO + DistMFbreed, data=minkUMF)
#summary(mink.d5)
```

Next we implement model selection via AICc; we need to create a list of models, give them names that make sense, create a selection table and identify the best model that explains detection.

```
modlist_d <- list (DateWindSky = mink.d1, DateSky = mink.d2,
                   Date = mink.d3, Sky = mink.d4, DateWind = mink.d5)

selection_d = model.sel(modlist_d)
selection_d
```

The best model is mink.d2: variables Date and Sky are best for explaining variation in detection probability. We now use these 2 variables to build several models that explore predictors for occupancy.

```
# climate predictors
mink.o1 = occu(~Date + Sky ~JulyTemp, data=minkUMF)

# landscape predictors
mink.o2 = occu(~Date + Sky ~DistMFbreed + factor(Beaver), data=minkUMF)

# pond predictors
mink.o3 = occu(~Date + Sky ~DO, data=minkUMF)

# climate and pond predictors
mink.o4 = occu(~Date + Sky ~JulyTemp + DO, data=minkUMF)

# landscape and pond predictors
mink.o5 = occu(~Date + Sky ~DistMFbreed + factor(Beaver) + DO, data=minkUMF)

# landscape and climate predictors
mink.o6 = occu(~Date + Sky ~DistMFbreed + factor(Beaver) + JulyTemp, data=minkUMF)
```

We are ready to implement model selection via AICc; we need to create a list of models, give them names that make sense, create a selection table and identify the best model that explains detection.

```
modlist_o <- list(JTemp = mink.o1, MFBreedBeaver = mink.o2,
                  DO = mink.o3, JTempDO = mink.o4,
                  MFBreedBeaverDO = mink.o5, FBreedBeaverJTemp = mink.o6)

selection_o = model.sel(modlist_o)
selection_o
```

Model Averaging: instead of choosing the BEST model, we should draw inferences on occupancy and detection based on all models, weighed by their AICc weight, and use all models whose cumulative AICc weight <=0.95. The function below returns the averaged model coefficients (RECENT WORK CLAIMS THIS IS WRONG, SO WE WILL NOT DWELL ON THEM)

```
mod.avg = model.avg(selection_o, cumsum(weight) <= .95)
mod.avg
```

We can estimate detection probability and occupancy probability using model averaging (THIS IS CORRECT)

```
detpred = predict(mod.avg, type="det")
detpred
hist(detpred$fit)

psipred = predict(mod.avg, type="state")
psipred
hist(psipred$fit)
```

Using the predicted occupancy estimates, we can now get the probability of occupancy CORRECTED for imperfect detection, also known as Proportion of Areas Occupied (PAO) = mean of 'psipred', and evaluate the difference between naive and occupancy predictions

```
PAO = mean(psipred$fit)
PAO

PAO-minkfrog_naive
```

There is a difference between the estimated Proportion of Area Occupied and the naive occupancy estimates, suggesting that one or more sites that were surveyed are likely occupied by mink frogs, but were never detected during the study. As such, not accounting for imperfetc detection would have underestimated the true occupancy of mink frogs in the study area.

THE END