Occupancy simulation and analysis

Understanding the simulations and the basic occupancy model

Diego J. Lizcano, Ph.D.

2022-06-26

Índice general

Capítulo 1

Prerequisites

This is a tutorial book written with **Markdown**

Using R y R studio, using the 'bookdown', 'knitr' and 'rmarkdown' packages.

This *book-tutorial* is part of the mini course on occupation methods with R.

Before starting please install the JAGS program on your computer, then from R studio install the unmarked, raster, spatstat, jagsUI, mcmcplots and ggmcmc packages.

```
install.packages("unmarked", dependencies = TRUE)
install.packages("raster", "spatstat", "jagsUI", "mcmcplots", "ggmcmc", dependencies =
```

1.1 Please cite this work as:

Lizcano D.J. (2019). Simulación y análisis de ocupación. Entendiendo las simulaciones y el modelo básico de ocupación (Version 1).

Capítulo 2

Why to do simulations?

2.1 why simulations are useful:

1. When doing simulations, the true parameters are known, so we can ensure that the code we execute (R or BUGS) estimates what we want, and that the estimates are equal to or close to the true parameters, allowing us to debug errors in the code.
2. We can calibrate a derived and/or more complex model more easily. Simulations can be viewed as a controlled experiment, or as simplified versions of a real system, in which we can test how certain parameters vary and affect estimates of other parameters. Conducting controlled experiments in the real world is often impractical or impossible in ecology, so simulation is the most consistent way to study the ecological system.
3. Sampling error is experienced firsthand and becomes a fantastic learning process.
4. We can check the quality (frequentist) of the estimates, as well as the precision and the effect of sample size, by computing the difference between the mean of the estimate and the true value (bias) and the variance of the estimates (the precision).
5. It is the most flexible and direct way to carry out power analysis, solving the great problem of determining the sample size necessary to detect an effect of a certain magnitude, with a given probability.
6. We can visualize how identifiable the parameters are in more complex models.
7. We can check how robust the model is to violations of the assumptions.
8. Being able to simulate data under a particular model ensures that one understands the model, its constraints, and limitations.

Capítulo 3

The Occupancy

Obtaining data for studies of animal populations is costly and wasteful, and it is not always possible to measure population density or demographic parameters such as birth or mortality rates. That is why the estimation of habitat occupation (ψ) is a good study tool since it is a reflection of other important population parameters such as abundance and density, which require a high number of records, with the economic and logistical costs involved. Additionally, and because detectability (p) in wild animals is not complete, the use of raw data generates underestimates of habitat occupation. Using repeated sampling, it is possible to generate estimates of detectability and, with this estimate, obtain unbiased values of habitat occupancy. Occupancy analysis methods were initially developed by (**?**) and later expanded by other authors (**??????**). These types of models allow inferences to be made about the effects of continuous and categorical variables on habitat occupancy. Furthermore, if sampling is done over long periods, it is also possible to estimate extinction and recolonization rates, which are useful in metapopulation studies (**?**). This is a field of great development in biostatistics that has produced a great explosion of studies that use occupation taking detectability into account (**??????**).

Capítulo 4

Our Example:

The data set that we are going to simulate mimics the spatial and temporal way in which we imagine the repeated measures of presence-absence in ecology originate. Which are a combination of an ecological process and an observation process. The first process contains the mechanisms under which Spatio-temporal distribution patterns originate, while the second process contains the different facets in which sources of error originate when taking the data.

To be more concrete we are going to call our imaginary species by a real name. We will call it the Mountain Tapir (*Tapirus pinchaque*), a large and conspicuous mammal, distributed from Colombia to Ecuador and Northern Peru, and listed as endangered in terms of conservation.

Mountain tapir in "Los Nevados" National Park. Colombia

Though it is the smallest (and furriest!) of tapir species, the Mountain tapir is the largest mammal in the tropical Andes mountain range. Their long hair is brownish to black, and their lips are lined in a white color.

The data set contains J replicated data of detection or non-detection of the species in M sites, taking into account that we assume that it is a closed population ('closure' assumption). This means that during the sampling there were no changes due to births, deaths, immigration or emigration. In other words, the sampling was short in time and the occurrence of species z did not change due to demographic effects.

Clearly, we must distinguish two processes, the first is the ecological process, which generates (partially) a latent state of the occurrence z. The second is the observation process, which produces the observed tapir detection or non-detection data. Here we assume that the observation process is governed by an imperfect detection mechanism. In other words, some tapir could have escaped my observation, which generates false negatives. We also assume that false positives are absent, meaning that anything I identify as a tapir is indeed a tapir and not a deer or a bear. To make the example more realistic, we include the effects of elevation (altitude) and forest cover on occurrence, as factors that affect occurrence linearly, decreasing it in the case of elevation, and increasing occupation linearly in the case of forest cover. In the end the two variables interact negatively with each other. These effects are introduced in the logarithmic scale occurrence as a generalized linear model (GLM) is traditionally done.

In our simulation we are going to make it explicit that it is not possible to detect all the tapirs from a sample site, so we are facing a type of error that makes us underestimate the abundance of the population. There are many reasons why we fail to spot an individual in the wild, it can be because we got distracted while the tapir passed by, because the binoculars did not have enough magnification, or simply because the tapir hid behind a tree upon smelling us, or because some other reason. In this way, we are going to register the presence (z=1) with a probability of detection p which we are also going to make dependent (on the logarithmic scale) of the elevation and of a co-variable that affects the detection, the temperature. In general terms, animals are more difficult to observe when the temperature is higher, and generally the higher the elevation, the lower the temperature. In this way, we assume that detection is negatively related to elevation and temperature. But it should also be noted that the negative effect on p can also be mediated by a decrease in abundance with elevation, which also causes the probability of occupancy to decrease with elevation. Note that a co-variable, elevation affects both the ecological process (the occurrence) and the observational process (the probability of detection). This has a purpose and is likely to happen in nature many times. Occupancy models have a "mechanistic" basis producing a spatial variation in abundance. That is, we will have sites with greater abundance and others with less abundance. But hierarchical models, like the one we are about to build, are capable of unraveling these complex relationships between occurrence and probability of detection (**???**). Finally, for this first example, we are going to leave out the effect of the interaction between elevation and temperature, setting it to zero. Then we can vary this parameter to consider that effect. In summary, we are going to generate data under the following model, where the sites are indexed as i and the repeated

counts on the site are going to be referred to as j.

4.0.1 Ecological Model:

$$z_i = Bernoulli(\psi_i) \tag{4.1}$$

$$logit(\psi_i) = \beta_0 + \beta_1 * Elevation_i + \beta_2 * CovForest_i + \beta_3 * Elevation_i * CovForest_i \tag{4.2}$$

4.0.2 Observation Model:

$$y_{ij} = Bernoulli(z_i * p_{ij}) \tag{4.3}$$

$$logit(p_{ij}) = \alpha_0 + \alpha_1 * Elevation_i + \alpha_2 * Temperature_{ij} + \alpha_3 * Elevation_i * Temperature_{ij} \tag{4.4}$$

Where ψ is the occupancy and p the probability of detection. With β as the regression coefficient for the occupancy covariates and α the regression coefficient for the detection covariates.

We are going to generate data from the "inside out" and from the top down. For this, we first choose the sample size and create the values for the covariates. Second, we select the values of the ecological model parameters (the occupancy) and assemble the expected occurrence (the parameter ψ, occupancy) and then obtain the random variable z which has a Bernoulli distribution. Third, we select the values of the parameters of the observation model (the detection), to assemble the probability of detection p and obtain the second set of a random variable y (observed or unobserved detection of a tapir) which also has distribution Bernoulli.

To simulate the data we will use the statistical programming language R (**?**), which provides a wide variety of graphical and statistical modeling techniques and a large ecosystem of packages for statistical and ecological analysis. If you haven't already, download and install R on your computer, then do the same with RStudio.

4.1 Initial steps: sample size and covariate values

Start RStudio, copy, paste and execute the commands in the gray window.

We first choose the sample size, the number of sites, and the number of repeated measures (number of visits) of presence/absence at each site.

```
M <- 60 # Number of spatial replicas (sites)
J <- 30 # Number of temporal replicas (repeat counts)
```

We then create the values for the covariates. We have elevation and forest cover as co-variables for each site. They differ from site to site but for each sampling, they are the same. While the temperature is a co-variable of the observation, so it does vary in each sampling and also in each site. Remember that the sub-index i refers to the site and the j to each sampling. To keep things simple our covariates are going to have a normal distribution with a mean-centered at zero and not going to extend very far on either side of zero. In real data analysis, we will have to standardize the co-variables to avoid numerical problems of difference in the scales of the co-variables and to be able to calculate the value of maximum likelihood (ML), as well as to obtain convergence in the Markov chains of the Bayesian model. Here we are going to ignore a fact of real life, and that is that the co-variables are not totally independent of each other, that is, in nature, forest cover can be related to elevation, but this is not going to be relevant, for now.

To initialize the random number generator and always get the same results we can add the following line:

```
set.seed(24) # Can choose seed of your choice
```

In this way, we can always obtain the same estimates. But then when we want to get the sampling error we will have to remove that line. For this example, we will generate values already standardized for the covariates, which are centered at zero and ranging from -1 to 1.

```
elev <- runif(n = M, -1, 1)              # Scaled elevation of a site
forest <- runif(n = M, -1, 1)            # Scaled forest cover at each site
temp <- array(runif(n = M*J, -1, 1), dim = c(M, J)) # Scaled temperature
```

4.2 Simulating the ecological process and its result: the occurrence of The Tapir

To simulate the occurrence of tapirs at each site, we choose the values for the parameters that govern the spatial variation in occurrence β_0 to β_3. The first parameter is the expected average occurrence of tapir (occupation probability) when all covariates have a value of zero, in other words the intercept of the occurrence model. We prefer to think of tapir in terms of their occurrence rather than logit (occurrence). Here we choose the occupancy intercept first and then transform it from the logarithmic scale with the logit link function.

```
mean.occupancy <- 0.60          # Mean expected occurrence of tapir
beta0 <- plogis(mean.occupancy) # Same on logit scale (= logit-scale intercept)
beta1 <- -2                     # Effect (slope) of elevation
beta2 <- 2                      # Effect (slope) of forest cover
beta3 <- 1                      # Interaction effect (slope) of elev and forest
```

Here we apply the linear model (to the logarithmic scale) and obtain the logit transformation of the occupancy probability, which we invert with the logit transformation to obtain the tapir occupancy and plot everything.

```
logit.psi <- beta0 + beta1 * elev + beta2 * forest + beta3 * elev * forest
psi <- plogis(logit.psi)      # Inverse link transformation

# par()               # view current settings
opar <- par()       # make a copy of current settings
par(mfrow = c(2, 2), mar = c(5,4,2,2), cex.main = 1)
curve(plogis(beta0 + beta1*x), -1, 1, col = "red", frame.plot = FALSE, ylim = c(0, 1),
      xlab = "Altitud", ylab = "psi", lwd = 2)
text(0.9, 0.95, "A", cex = 1.5)
plot(elev, psi, frame.plot = FALSE, ylim = c(0, 1), xlab = "Altitud", ylab = "")
text(0.9, 0.95, "B", cex = 1.5)
curve(plogis(beta0 + beta2*x), -1, 1, col = "red", frame.plot = FALSE, ylim = c(0, 1),
      xlab = "Forest cover", ylab = "psi", lwd = 2)
text(-0.9, 0.95, "C", cex = 1.5)
plot(forest, psi, frame.plot = FALSE, ylim = c(0, 1), xlab = "Forest cover", ylab = "")
text(-0.9, 0.95, "D", cex = 1.5)
```

```
# dev.off()
par(opar)           # restore original par settings
```

To better show the joint relationship between the two covariates and psi, we need to make a surface plot. Here we have not changed anything about the simulation, we have only added more data to it to better visualize it.

```
# Compute expected occurrence for a grid of elevation and forest cover
cov1 <- seq(-1, 1, , 100)                       # Values for elevation
cov2 <- seq(-1, 1, , 100)                       # Values for forest cover
psi.matrix <- array(NA, dim = c(100, 100))      # Prediction matrix, for every
# combination of values of elevation and forest cover

for(i in 1:100){
   for(j in 1:100){
      psi.matrix[i, j] <- plogis(beta0 +
                                 beta1 * cov1[i] +
                                 beta2 * cov2[j] +
                                 beta3 * cov1[i] * cov2[j] )
   }
```

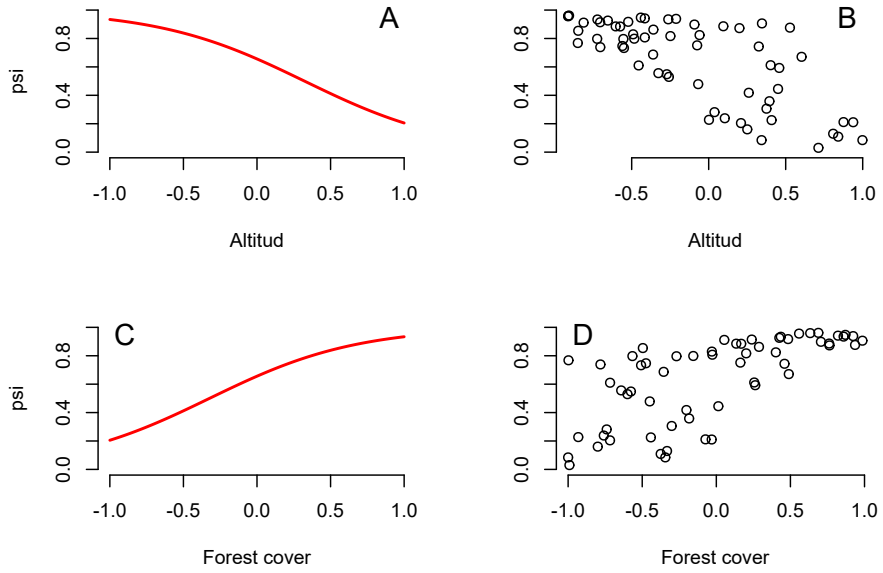

Figura 4.1: Two ways to show the relationship between the probability of occurrence of tapir and the covariates. (A) Relationship between psi and elevation for a constant value (mean equal to zero) of forest cover. (B) Relationship between psi and elevation in an observed value of forest cover. (C) Forest cover psi ratio for a constant elevation (at mean zero). (D) Relationship psi forest cover for the observed value of elevation.

```
}

mapPalette <- colorRampPalette(c("grey", "yellow", "orange", "red"))
image(x = cov1, y = cov2, z = psi.matrix, col = mapPalette(100), xlab = "Altitud",
      ylab = "Forest cover", cex.lab = 1.2)
contour(x = cov1, y = cov2, z = psi.matrix, add = TRUE, lwd = 1)
matpoints(elev, forest, pch="+", cex=0.8)
```

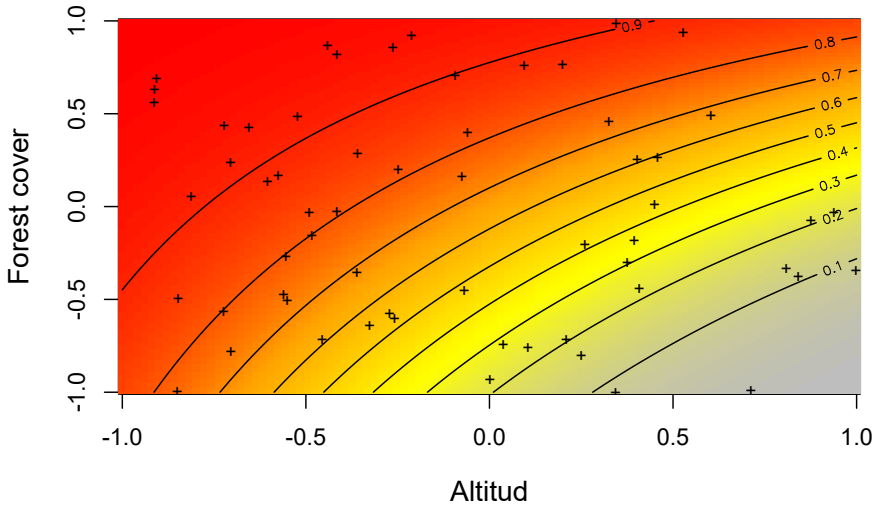

Figura 4.2: Relationship constructed between the simulated data of the expected occurrence (occupation) of tapir (psi) represented with the color scale from gray to red, against elevation and forest cover simultaneously. In this case the interaction between the two covariates is given by the value of beta3 = 1 that we have established previously.

So far we have not introduced any stochastic variation in the relationship between tapir occurrence and covariates. To do this we must make use of some statistical models, or statistical distributions, to describe the random variability around the expected value of psi. The typical way to introduce this random variation is to obtain the occurrence of tapir at each site i, z_i , from a Bernoulli distribution with the expected values (ψ_i).

4.2.1 Why Bernoulli?

In the ecological process z_i tapir occurrence is represented by a Bernoulli-type distribution where tapir is present at a site represented as the occupancy ψ at a site where it is present, or not present 1-ψ. The Bernoulli distribution is a special case of the binomial distribution, and its best example is a single toss of a coin. If you require a more extensive, basic, detailed explanation and with more examples, I recommend you visit khanacademy.

```
z <- rbinom(n = M, size = 1, prob = psi)  # Realised occurrence. A Bernoulli
sum(z)                                    # Total number of occupied sites
```

```
## [1] 40
```

```
table(z)                                  # Frequency distribution of tapir occurrence
```

```
## z
##  0  1
## 20 40
```

Here we have created the result of the ecological process: site-specific occurrence z_1. We see that 20 sites are not occupied and the remaining 40 sites are occupied

4.3 Simulating the observation process and its result: the detection

Occurrence z is not what we normally see, as there is a chance that we will fail to observe an individual. Hence there is a binary measure of error when we measure occurrence (we observe it or we don't observe it). We assume that we can make only one of the two possible sightings (yes, no), but we may have missed a tapir sighting somewhere, so the probability of detection is less than one, and the error measure is affected by coverage. of forest and temperature. Keep in mind that we will never register the presence of a tapir when in fact there are no tapir. In other words, we are assuming that we have no false positives. To make it explicit that we have an interaction effect between two covariates in our data, we are going to allow an interaction effect in the code, but set it to zero and thus have no effect in the model that generates the data. We first select the values for α_0 to α_3, where the first is the probability of detection for the tapir, on the logit scale, when all detection covariates have a value of zero. We have chosen the intercept of the detection model and then transformed it with the plogis link function. This is not the same as the average detection probability, which is higher in our simulation model, as we will see later.

```
mean.detection <- 0.3             # Mean expected detection
alpha0 <- qlogis(mean.detection) # same on logit scale (intercept)
alpha1 <- -1                      # Effect (slope) of elevation
alpha2 <- -3                      # Effect (slope) of temperature
```