Occupancy models

Bayesian statistics 9 – latent variable modelling

Frédéric Barraquand (CNRS, IMB)

16/01/2023

What is a latent variable?

- A variable that is *inferred* not *observed*

What is a latent variable?

- A variable that is *inferred* not *observed*
- Random effects might fit that description

What is a latent variable?

- A variable that is *inferred* not *observed*
- Random effects might fit that description
- Can be discrete (1,2,. . .) or continuous

What is a latent variable?

- A variable that is *inferred* not *observed*
- Random effects might fit that description
- Can be discrete (1,2,. . .) or continuous
- Here we will focus on discrete

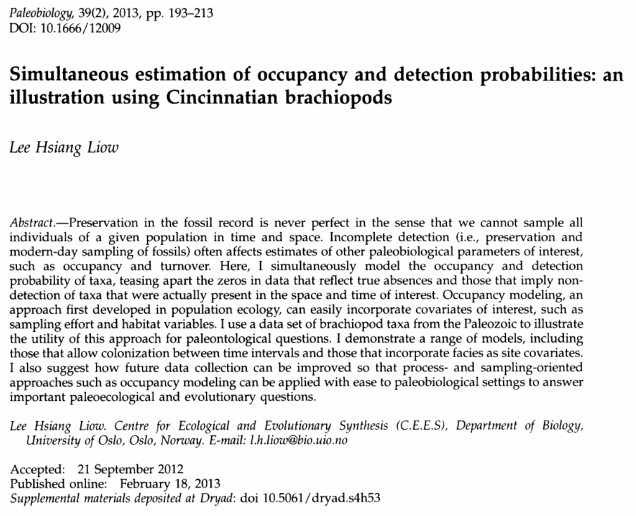
What is a latent variable?

- A variable that is *inferred* not *observed*
- Random effects might fit that description
- Can be discrete (1,2,. . .) or continuous
- Here we will focus on discrete
- Often arise in the context of modelling your observations with two submodels

What is a latent variable?

- A variable that is *inferred* not *observed*
- Random effects might fit that description
- Can be discrete (1,2,. . .) or continuous
- Here we will focus on discrete
- Often arise in the context of modelling your observations with two submodels
- Full model = Observation submodel + Process submodel (hidden)

Used for ecology, evolution and much more



Paleobiology, 39(2), 2013, pp. 193–213
DOI: 10.1666/12009

Simultaneous estimation of occupancy and detection probabilities: an illustration using Cincinnatian brachiopods

Lee Hsiang Liow

Abstract.—Preservation in the fossil record is never perfect in the sense that we cannot sample all individuals of a given population in time and space. Incomplete detection (i.e., preservation and modern-day sampling of fossils) often affects estimates of other paleobiological parameters of interest, such as occupancy and turnover. Here, I simultaneously model the occupancy and detection probability of taxa, teasing apart the zeros in data that reflect true absences and those that imply non-detection of taxa that were actually present in the space and time of interest. Occupancy modeling, an approach first developed in population ecology, can easily incorporate covariates of interest, such as sampling effort and habitat variables. I use a data set of brachiopod taxa from the Paleozoic to illustrate the utility of this approach for paleontological questions. I demonstrate a range of models, including those that allow colonization between time intervals and those that incorporate facies as site covariates. I also suggest how future data collection can be improved so that process- and sampling-oriented approaches such as occupancy modeling can be applied with ease to paleobiological settings to answer important paleoecological and evolutionary questions.

Lee Hsiang Liow. Centre for Ecological and Evolutionary Synthesis (C.E.E.S), Department of Biology, University of Oslo, Oslo, Norway. E-mail: l.h.liow@bio.uio.no

Accepted: 21 September 2012
Published online: February 18, 2013
Supplemental materials deposited at Dryad: doi 10.5061/dryad.s4h53

Does absence mean true absence or just no fossils at that time?

A Hollywood archeology example

Indy & Lara criss-cross the Amazonian jungle in search of artefacts from hidden civilizations. They have a map with 100 x 100 km quadrats. In each quadrat, there could be cultural signs but these may not be visible. Thus we consider a probability of detection p. We want to know how rich the region is, i.e., what is the probability that a quadrat is truly occupied ψ.

Upper Amazon - screenshot Google Earth

Not such a silly example, actually

What SARS-CoV-2's mild cousins reveal about COVID-19 p. 141
Specifying laws of friction pp. 150 & 200
Continued decline of sharks despite regulation p. 225
Science
$15
12 JANUARY 2024
science.org
AAAS
LOST CITY
Ancient development in the Upper Amazon p. 183

Rostain et al. Science 2024

Let's put this into equations

- We number the site $i \in \{1, .., I\}$.

Let's put this into equations

- We number the site $i \in \{1, .., I\}$.
- Variable $X_{it} = 1$ if there was an artefact observed at time t in site i, 0 otherwise. They visit the sites at various times.

Let's put this into equations

- We number the site $i \in \{1, .., I\}$.
- Variable $X_{it} = 1$ if there was an artefact observed at time t in site i, 0 otherwise. They visit the sites at various times.
- Variable Z_i is the latent state, i.e., has value 1 there truly an artefact within quadrat i.

What is the model?

[Pen & paper moment]

What is the model?

[Pen & paper moment]

Solution. For all i

$$X_{it} | Z_i \sim \text{Bernoulli}(Z_i p)$$

$$Z_i \sim \text{Bernoulli}(\psi)$$

Is that OK?

One can prove this is equivalent to

$$X_{it} \sim \text{Bernoulli}(p\psi)$$

(btw: true with binomial not just Bernoulli variables)

Problem: $p\psi$ is just one parameter.

Proof: $\mathbb{P}(X_{it}=1) = \mathbb{P}(X_{it}=1|Z_i=1)\mathbb{P}(Z_i=1) + \mathbb{P}(X_{it}=1|Z_i=0)\mathbb{P}(Z_i=0) = p \times \psi + 0 \times (1-\psi)$

Better occupancy model

i site index in $\{1, ..., I\}$

t visit index in $\{1, ..., T\}$

$$X_{it}|Z_i \sim \text{Bernoulli}(Z_i p)$$

$$Z_i \sim \text{Bernoulli}(\psi)$$

'Robust design': T repeats within each site i. Parameters identifiable now.

(McKenzie et al. 2002)

Simulating the occupancy model

```r
#set.seed(42)
I <- 250;
T <- 10;
p <- 0.4;
psi <- 0.3;

z <- rbinom(I,1,psi); # latent occupancy state
y <- matrix(NA,I,T);  # observed state
for (i in 1:I){  y[i,] <- rbinom(T,1,z[i] * p);}
```

JAGS/BUGS modelling

```
occupancy.data <- list(y=rowSums(y), T=T,nsite=I)

cat(file="occupancy.txt","
model {

  # Priors
    p~dunif(0,1)
    psi~dunif(0,1)

  # Likelihood
    for(i in 1:nsite){
      mu[i] <- p*z[i]
      z[i] ~ dbern(psi)
      y[i] ~ dbin(mu[i],T)
      }
    n<-sum(z[])
    }
")
```

Running the model I

```
# Inits function
inits <- function(){list(p = runif(1, 0, 1),
                         psi = runif(1,0,1), z = rep(1, I))}
# we need to initialize z
# see https://bcss.org.my/tut/bayes-with-jags-a-tutorial-for-wildlife-researchers/oc

# Parameters to estimate
params <- c("p","psi")

# MCMC settings
nc <- 3  ;  ni <- 2000  ;  nb <- 1000  ;  nt <- 2

# Call JAGS, check convergence and summarize posteriors
out <- jags(occupancy.data, inits, params, "occupancy.txt", n.thin = nt,
            n.chains = nc, n.burnin = nb, n.iter = ni)
```

Running the model II

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 250
##    Unobserved stochastic nodes: 252
##    Total graph size: 758
##
## Initializing model
```

```
print(out, dig = 3)     # Bayesian analysis
```

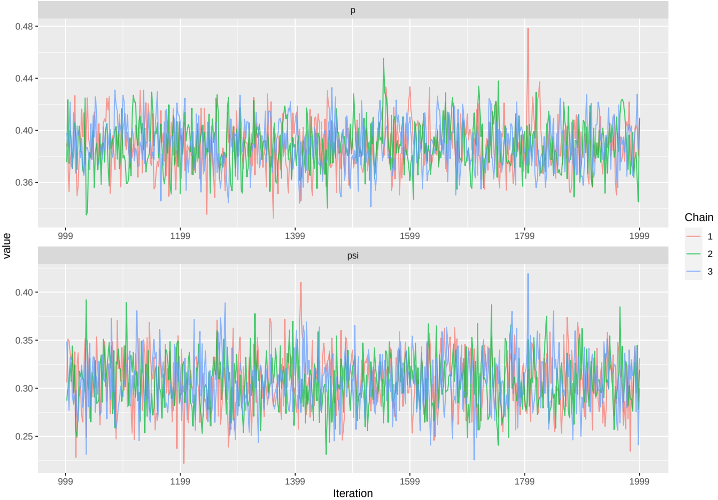
```
## Inference for Bugs model at "occupancy.txt", fit using jags,
##  3 chains, each with 2000 iterations (first 1000 discarded), n.thin = 2
##  n.sims = 1500 iterations saved
##          mu.vect sd.vect    2.5%     25%     50%     75%   97.5% Rhat n.eff
## p          0.389   0.018   0.353   0.376   0.388   0.401   0.425    1  1500
## psi        0.307   0.028   0.254   0.287   0.306   0.326   0.362    1  1500
## deviance 283.991   7.584 277.313 277.693 280.004 287.816 300.244    1  1500
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
```

Running the model III

```
## pD = 28.8 and DIC = 312.8
## DIC is an estimate of expected predictive error (lower deviance is better).
```

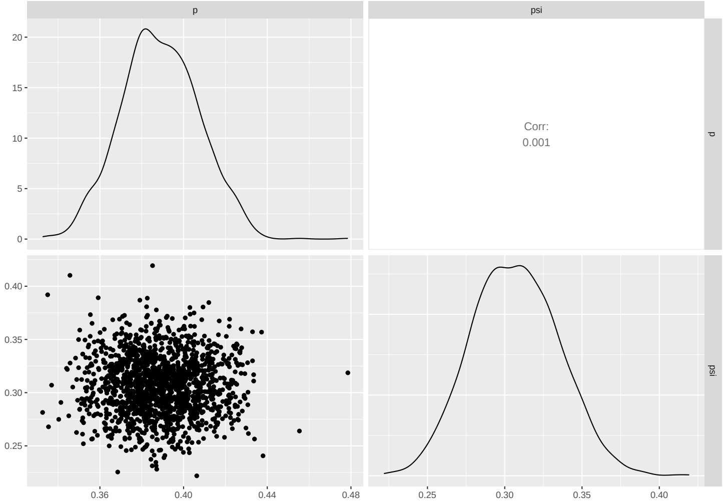
Showing traceplots

```
S<-ggs(as.mcmc(out)) #R2jags
S<-filter(S,Parameter != "deviance")
ggs_traceplot(S)
```



Showing correlations (p,ψ)

```
ggs_pairs(S)
```



Adding covariates

Possible to add covariates on detection probability

$p_{it} = \text{logistic}(\alpha_{k[i]} + \beta \times \text{survey duration}_{it})$

or

$\text{logit}(p_{it}) = \ln(\frac{p_{it}}{1-p_{it}}) = \alpha_{k[i]} + \beta \times \text{survey duration}_{it}$

Covariates on occupancy probability, e.g.

$\psi_i = \text{logistic}(\alpha_\psi + \beta_\psi \times \text{population density}_i)$

Real-life example

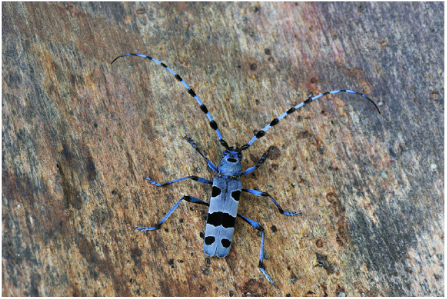


FIGURE 13.4 The remarkable "blue bug", the cerambycid beetle *Rosalia alpina*, Switzerland, 2009 (Photograph by T. Marent).

Bluebug *Rosalia alpina* from Kéry (2010)

The dataset

- 27 sites (woodpiles), 6 replicated counts for each.
- Covariates: `forest_edge` (edge or more interior), `date`, `hour` (date and hour of day, both of these are control variables)
- Detection at 10 of 27 woodpiles and from 1 to 5 times
- Questions:
  - Have some bluebugs been likely missed in some sites?
  - How many times should one visit a woodpile?
  - Effect of forest edge? (bluebugs are a typical forest species)