Occupancy modelling using hmmTMB

Théo Michelot

2022-12-15

In the following, we use hmmTMB to analyse a crossbill occupancy dataset from the R package unmarked, which is described by Schmid, Zbinden, and Keller (2004) and Kéry, Guillera-Arroita, and Lahoz-Monfort (2013).

```
# Load packages and color palette
library(ggplot2)
theme_set(theme_bw())
library(hmmTMB)
pal <- hmmTMB:::hmmTMB_cols
```

1 Data preparation

The dataset included in unmarked has one row for each survey site, with columns for:

- `ele`: elevation of the survey site
- `forest`: forest cover at survey site
- `surveys`: number of surveys carried out (either 2 or 3)

It also has many columns filled with 0s and 1s, depending on whether crossbills were detected at each site and each survey event, named with the following convention:

- `detxyz`: number of detections in year xy and survey z

```
# Load data from unmarked package
data(crossbill, package = "unmarked")

crossbill[1:6, 1:8]
```

```
  id  ele forest surveys det991 det992 det993 det001
```

```
1  1  450      3       3     0      0      0      0
2  2  450     21       3     0      0      0      0
3  3 1050     32       3    NA     NA     NA      0
4  4  950      9       3     0      0      0      1
5  5 1150     35       3     0      0      0      1
6  6  550      2       3    NA     NA     NA      0
```

For the HMM analysis, we need to change this dataset to a “long” format, where each year of survey is on a different row. The resulting data set has the following columns:

- `ID`: identifier for survey site
- `year`: survey year
- `elev` and `forest` are the environmental covariates
- `surveys`: number of surveys for this site and this year
- `y2`: count of detections for years with 2 surveys (either 0, 1, or 2)
- `y3`: count of detections for years with 3 surveys (0, 1, 2, or 3)

The last two columns are separate because the two variables need to be modelled separately by the HMM. Specifically, we assume that these variables follow binomial distributions, where the size parameter is either 2 (for `y2`) or 3 (for `y3`).

```r
# Get into format for hmmTMB
nsites <- nrow(crossbill)
data <- data.frame(ID = rep(1:nsites, each = 9),
                   year = rep(1:9, nsites),
                   elev = rep(crossbill$ele, each = 9),
                   forest = rep(crossbill$forest, each = 9),
                   surveys = rep(crossbill$surveys, each = 9))
y <- apply(as.matrix(crossbill[,5:31]), 1, FUN = function(x) {
  tapply(as.numeric(x), rep(1:9, each = 3), FUN = function(r) {sum(r, na.rm = TRUE)})
})
y <- as.numeric(y)
data$y2 <- ifelse(data$surveys == 2, y, NA)
data$y3 <- ifelse(data$surveys == 3, y, NA)
data$forest <- as.numeric(data$forest)
data$elev <- as.numeric(data$elev)
```

```
head(data)
```

```
  ID year elev forest surveys y2 y3
1  1    1  450      3       3 NA  0
2  1    2  450      3       3 NA  0
3  1    3  450      3       3 NA  0
4  1    4  450      3       3 NA  0
5  1    5  450      3       3 NA  0
6  1    6  450      3       3 NA  0
```

2 Model specification

In this application, the observation is the number of detections in a given site and year, and the hidden state is the occupancy of the site (i.e., either “occupied” or “not occupied”). The occupancy is modelled as a Markov chain, and one aim is to estimate the transition probabilities between those two states, i.e.,

- Pr(occupied → not occupied), the extinction probability, and
- Pr(not occupied → occupied), the colonisation probability.

In the following, we define state 1 as “not occupied” and state 2 as “occupied”.

We denote as $y_t^{(2)}$ and $y_t^{(3)}$ the number of detections in year t for years with 2 and 3 survey events, respectively. The observation model is

$$y_t^{(k)} | S_t = j \sim \text{binomial}(\text{size} = k, \text{prob} = p_j^{(k)})$$

where S_t is the hidden state. The parameter $p_j^{(k)}$ is the detection probability in state j at a site which was surveyed k times. We don’t actually need to estimate all those parameters, because we know that $p_1^{(k)} = 0$, i.e., the probability of detection is zero if the animal is not present.

We first create an R object from the `MarkovChain` class, for the hidden state model, indicating that we need a 2-state model. We use `initial_state = "stationary"` to fix the initial distribution of the Markov chain to the stationary distribution of the transition probability matrix (rather than estimating it for each survey site).

```
# Define hidden state model
hid1 <- MarkovChain$new(data = data, n_states = 2,
                        initial_state = "stationary")
```

The second step is to create an object of class Observation, to specify the observation model of the HMM. In particular, this requires passing a list of observation distributions (here, “binom” for both observed variables), and a list of initial parameter values. The size parameter is the same in both states: either 2 (for sites with 2 surveys) or 3 (for sites with 3 surveys). The detection probability parameter is 0 in state 1 for both variables, and we choose some plausible value for state 2 (here, $p = 0.5$).

```
# Define observation model
dists <- list(y2 = "binom", y3 = "binom")
par0 <- list(y2 = list(size = c(2, 2), prob = c(0, 0.5)),
             y3 = list(size = c(3, 3), prob = c(0, 0.5)))

obs1 <- Observation$new(data = data, dists = dists,
                        n_states = 2, par = par0)
```

We can now create an HMM object, which combines the hidden state and observation components. We use the argument fixpar to indicate that the detection probabilities in state 1 do not need to be estimated (and should be kept fixed to their initial value, zero). To find the name of the parameter that needs to be fixed, we can use the following command:

```
obs1$coeff_fe()
```

```
                            [,1]
y2.size.state1.(Intercept)     2
y2.size.state2.(Intercept)     2
y2.prob.state1.(Intercept)  -Inf
y2.prob.state2.(Intercept)     0
y3.size.state1.(Intercept)     3
y3.size.state2.(Intercept)     3
y3.prob.state1.(Intercept)  -Inf
y3.prob.state2.(Intercept)     0
```

The -Inf entries are the ones we want to keep fixed, so we can specify fixpar as shown below. It is defined as a list, in which the element obs is a named vector where the fixed parameters are set to NA.

```
# Fix detection prob to zero in state 1
fixpar <- list(obs = c("y2.prob.state1.(Intercept)" = NA,
                       "y3.prob.state1.(Intercept)" = NA))
```

Finally, we create and fit the model, which takes a few seconds.

```
# Create and fit model
hmm1 <- HMM$new(obs = obs1, hid = hid1, fixpar = fixpar)
hmm1$fit(silent = TRUE)
```

We can see the estimated parameters using `hmm1$par()`. Note that, if covariate effects were included on the observation parameters, then this would return estimated parameters *for the first time step*.

```
hmm1$par()
```

```
$obspar
, , 1

        state 1   state 2
y2.size       2 2.0000000
y2.prob       0 0.1675157
y3.size       3 3.0000000
y3.prob       0 0.5319512



$tpm
, , 1

          state 1   state 2
state 1 0.8656540 0.1343460
state 2 0.2093719 0.7906281
```

The estimates that the detection probability was $p_2^{(2)} = 0.17$ in years with 2 surveys, and $p_2^{(3)} = 0.53$ in years with 3 surveys. The colonisation probability over one year was about 0.13, and the extinction probability about 0.20.

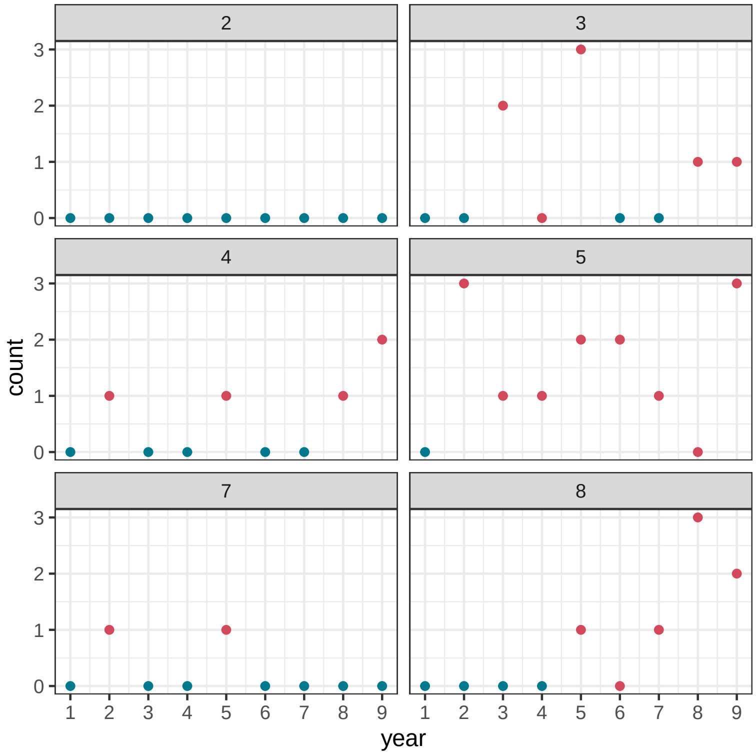
We can also obtain the most likely state sequence using the `viterbi()` function (after the Viterbi algorithm), for example to plot the observed time series coloured by estimated states. The code below creates such a plot for a few chosen sites.

```
# Get most likely state sequence
data$viterbi <- factor(hmm1$viterbi())

# Select sites to keep
ID_to_keep <- c(2, 3, 4, 5, 7, 8)
```

```
ind_to_keep <- which(data$ID %in% unique(data$ID)[ID_to_keep])

ggplot(data[ind_to_keep,], aes(year, y3, col = viterbi)) +
    geom_point() +
    facet_wrap("ID", ncol = 2) +
    ylab("count") +
    scale_color_manual(values = pal, guide = "none") +
    scale_x_continuous(breaks = 1:9)
```



3 Adding covariates

The dataset includes two environmental covariates: elevation, and forest cover. In this application, it might be of interest to investigate whether those affect the transition probabilities (i.e., colonisation and extinction probabilities). In this section, we focus on the effect of elevation. It could be included as a linear effect but we don't want to assume any parametric form for the relationship. We therefore use a smooth function, following formula syntax from the mgcv package.

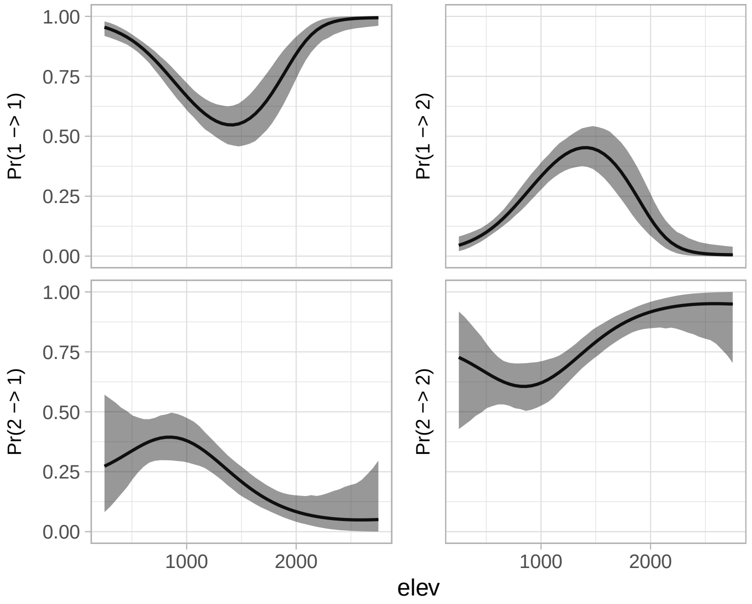
```
# Define hidden state process with elevation covariate
hid2 <- MarkovChain$new(data = data, n_states = 2,
                        formula = ~ s(elev, k = 5, bs = "ts"),
                        initial_state = "stationary")
```

The rest of the model specification is unchanged. This time, fitting the model takes a few minutes, as the non-parametric relationship between elevation and transition probabilities needs to be estimated.

```
obs2 <- Observation$new(data = data, dists = dists,
                       n_states = 2, par = par0)
hmm2 <- HMM$new(obs = obs2, hid = hid2, fixpar = fixpar)
hmm2$fit(silent = TRUE)
```

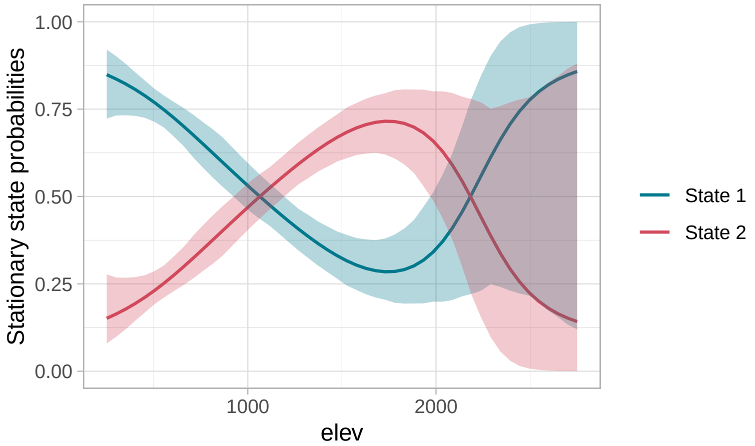
We can visualise the results by plotting the transition probabilities against the elevation covariate. The plot suggests that the colonisation probability is highest for an intermediate range of elevation values, roughly between 1000 and 1700m. On the other hand, it looks like the extinction probability decreases consistently with elevation.

```
hmm2$plot("tpm", var = "elev")
```



Another option is to plot the stationary state probabilities, i.e., the probabilities of being in each state in the long run, for a range of covariate values. This output can be easier to interpret in some cases than the transition probabilities themselves. The plot suggests that the probability of being in state 2 (occupied) is highest for elevations between 1000-2000m, although there is a lot of uncertainty for larger elevation values.

```
hmm2$plot("delta", var = "elev")
```



References

Kéry, Marc, Gurutzeta Guillera-Arroita, and José J Lahoz-Monfort. 2013. "Analysing and Mapping Species Range Dynamics Using Occupancy Models." *Journal of Biogeography* 40 (8): 1463–74.

Schmid, Hans, Niklaus Zbinden, and Verena Keller. 2004. "Überwachung Der Bestandsentwicklung häufiger Brutvögel in Der Schweiz." *Swiss Ornithological Institute Sempach Switzerland*.