Statistical and Technical Methods in BIRDIE (DRAFT)

Francisco Cervantes

01 September 2021

Introduction

This document provides details on the statistical models and used in the data pipeline and their implementation.

Abundance and population trends

State-space models

State-space models are fitted to CWAC data to decouple the underlying process of interest, namely, the population size of waterbird species, from observational artifacts that translate into counting errors.

State-space models are use to describe and understand dynamic systems that may not be perfectly observed. Within this framework, we consider a process of interest that evolves with time and which we may observe at certain occasions. However, our observations may be distorted by some imperfect observation process. For example, we might be interested in estimating the number of birds present at a certain site, but often field observers will miss some birds. By taking repeated measurements (counts in our example) over time, and assuming that the process of interest should evolve slowly compared to observation error, we may be able to disentangle these two processes.

In the particular case of CWAC data, we let the population size of a certain species at a given site at time t be denoted by μ_t. Actually, μ_t is the log of the population size (I should explain this better). Observers count the number of birds at this site at times $t = 1, 2, ..., N$, for a total of N measurements. Note that for simplicity we are assuming that sampling occasions are regularly spaced, although they need not be, as we will see later. Within the CWAC programme framework, there are yearly counts and therefore, t indexes the different years. Then,

$$
\begin{aligned}
\mu_t &= \mu_{t-1} + \alpha_t,\ \alpha_t \sim N(0, \sigma_\alpha^2) \\
y_t &= \mu_t + e_t,\ e_t \sim N(0, \sigma_e^2),
\end{aligned}
\tag{1}
$$

where μ_t represents the number of birds present at the site at time t and y_t represents the number of birds counted at the site at time t. The terms α_t and e_t denote stochastic changes in abundance and stochastic errors in the measurements respectively. Note that the state at time t, μ_t depends on the state at time $t - 1$, therefore there should be a temporal structure in the process μ_t that is not present in the observations y_t, if we condition on μ_t.

We can now extend the model to account for the fact that many sites host migrant species that arrive during the summer months. Therefore, we would expect an influx of birds to the population that we will denote by β_t. A proportion of these birds, stay during the winter months, while others leave to their breeding grounds in the North. The change in population size during the winter months due to the wintering migrants is denoted by λ_t. Then,

$$
\begin{aligned}
\mu_t &= \mu_{t-1} + \beta_{t-1}\text{summer}_t + \lambda_t\text{winter}_t + \omega_t,\ \omega_t \sim N(0, \sigma_\omega^2) \\
\beta_t &= \beta_{t-1} + \zeta_t,\ \zeta_t \sim N(0, \sigma_\zeta^2) \\
\lambda_t &= \lambda_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma_\epsilon^2) \\
y_t &= \mu_t + \alpha_t\text{summer}_t + e_t\text{winter}_t,\ \alpha_t \sim N(0, \sigma_\alpha^2),\ e_t \sim N(0, \sigma_e^2)
\end{aligned}
\tag{2}
$$

As previously, μ_t denotes the size of the population at time t. The variables $summer_t$ and $winter_t$ are indicators of summer period and winter period respectively (i.e., $summer_t$ equals 1 in summer and 0 in winter, and $winter_t$ is the opposite). The summer population at time t, μ_t, is calculated by adding a quantity β_{t-1} to the population at time $t-1$, μ_{t-1}. Then, the winter population at time t is calculated by adding a quantity λ_t, representing the difference between the summer and winter population at time t, to the summer population of that same time (year) t.

(There is something strange here: β_{t-1} actually is the difference between the summer population and the previous winter, when a portion of the previous summer migrants was still present (let's call these "wintering" birds). Or are we assuming that wintering birds left? Is it important at all?)

As we can see both μ_t and λ_t change over time and depend on their value at the previous time. This specification correspond to parameters that perform a random walk over time.

Implementation:

At the moment this state-space model is fitted in `R` with the package `jagsUI` that uses `rjags`. We have written a few functions in `BIRDIE` to make fitting and plotting output from these models easy. The basic workflow to fit this model to a species at a certain site (we use Barberspan as an example) is the following:

```r
# Load data (Barberspan example)
counts <- BIRDIE::barberspan

# Select a species (ADU code)
sp <- 83

# Prepare data to fit an SSM
ssmcounts <- BIRDIE::prepSsmData(counts, species = sp)

# Fit 2-season dynamic trend model
fit_dyn <- BIRDIE::fitCwacSsm(ssmcounts, mod_file = BIRDIE::writeJagsModelFile(),
                              param = c("beta", "lambda", "sig.zeta", "sig.w",
                                        "sig.eps", "sig.alpha", "sig.e", "mu_t", "mu_wt"))
```

We can then create plots and prepare data to export to use in the dashboard by doing:

```r
# Plot
pers_theme <- ggplot2::theme_bw()
p <- BIRDIE::plotSsm2ss(fit = fit_dyn, ssm_counts = ssmcounts, dyn = TRUE,
                        plot_options = list(pers_theme = pers_theme,
                                            colors = c("#71BD5E", "#B590C7")))

plot(p$plot)
```

There is also a function that allows us to fit a model and produce output for all the species detected at a certain site at once to facilitate automation:

```r
# Load data (Barberspan example)
counts <- BIRDIE::barberspan
```

```
# Fit all species and save data outputs and plots to "analysis/out_nosync/".
BIRDIE::loopSsmAllSpp(barberspan, data_outdir = "analysis/out_nosync/",
                      plot_outdir = "analysis/out_nosync/",
                      param = c("beta", "lambda", "sig.zeta", "sig.w", "sig.eps",
                                "sig.alpha", "sig.e", "mu_t", "mu_wt"),
                      jags_control = list(ncores = 3))
```

Species distribution

Occupancy modelling

Occupancy models are fitted to detection/non-detection data from the Southern Africa Bird Atlas Project (SABAP) to delineate the distribution of waterbird species and its dynamics over time. Within the SABAP framework observers visit pentads and make a list of the bird species detected during the visit. Detection data is assumed to work perfectly (i.e., there are no identification errors), but non-detections may be caused by either the species not being present in the pentad or by the observers not being able to detect it, although it was present. Therefore, occupancy models describe two processes simultaneously: i) the underlying occupancy of the sites (pentads), and ii) the observation process whereby species present might or might not be observed.

More precisely, we define z_{it} to be the true occupancy of site i in year t, which can be 1 (if species present) or 0 (if species absent) and has distribution:

$$z_{it}|\psi_{it} \sim \text{Bernoulli}(\psi_{it}), \tag{3}$$

where ψ_{it} is the occupancy probability at site i and year t. The logit transformation of ψ_{it} can be modelled as a linear combination of covariates and smooth functions of covariates, such that:

$$\text{logit}(\psi_{it}) = \boldsymbol{x}_{it}^{\intercal}\boldsymbol{\beta} + \sum_{k=1}^{K} f_k(u_{ik}) \tag{4}$$

where u_{ik} is a smooth function of the covariate u_k, which is defined as

$$f_k(u_{ik}) = \sum_{j=1}^{J} \text{B}_j(u_{ijk})\gamma_{ijk} \tag{5}$$

where the smooth function f is represented by a set of basis functions B_j evaluated at the value of the covariates associated with site i at time t. A notable case for the use of such functions is to incorporate temporal and spatial effects.

Then, the probability of detection of a species that is present in site i on visit j is denoted by p_{ij}. Following the same logic as for the probability of occupancy, the logit transformation of p is modelled as a linear combination of covariates and smooth functions:

$$\text{logit}(p_{ij}) = \boldsymbol{w}_{ij}^{\intercal}\boldsymbol{\alpha} + \sum_{h=1}^{H} f_h(v_{ih}), \tag{6}$$

Finally, the likelihood of observation y_{ij} can be written as:

$$y_{ij}|z_{it}, p_{ij} \sim \text{Bernoulli}(z_{it}p_{ij}) \tag{7}$$

Implementation

We use the functionality provided by the `R` package developed by Richard Glennie `occuR`. We have also created functions in the `BIRDIE` package to simplify this process and facilitate automation. The basic workflow entails:

- Defining a region and a species of interest we want to estimate occupancy for.

```r
library(BIRDIE)

# SABAP code for the species of interest
sp_sel <- 6

# Region of interest
region <- "South Africa"
```

- Next we need to extract all pentads in the region and annotate them with covariates. Currently we are using climatic covariates provided by the TerraClimate data set and the surface water dataset provided by the European Commission Joint Research Centre.

```r
# Set up multicore
future::plan("multisession", workers = 6)

sites <- prepOccSiteData(region = region,
                         years = 2008:2011,
                         clim_covts = c("prcp", "tmax", "tmin", "aet", "pet"),
                         covts_dir = "analysis/downloads/",
                         file_fix = c("terraClim_", "_03_19"),
                         savedir = "analysis/data/pentads_sa.rds")

future::plan("sequential")
```

- Next, we need to prepare occupancy records for the selected pentads. For this, we download data from the SABAP project and annotate this records with covariates.

```r
# Set up multicore
future::plan("multisession", workers = 6)

pa_dat <- prepOccVisitData(region = "South Africa",
                           sites = sites,
                           species = sp_sel,
                           years = 2008:2011,
                           clim_covts = c("prcp", "tmax", "tmin", "aet", "pet"),
                           covts_dir = "analysis/downloads/",
                           file_fix = c("terraClim_", "_03_19"),
                           savedir = "analysis/data/pentads_sa.rds")

future::plan("sequential")
```

- With these two elements we are ready to fit an occupancy model with `occuR`:

```r
# First, we need to give data the correct format
visits <- pa_dat %>%
    distinct(Pentad, year) %>%
    mutate(keep = 1)

sites <- unique(pa_dat$Pentad)
```

```
site_data <- sites %>%
    dplyr::select(-id) %>%
    sf::st_drop_geometry()  %>%
    tidyr::pivot_longer(cols = -c(Name, lon, lat, water)) %>%
    tidyr::separate(name, into = c("covt", "year"), sep = "_") %>%
    pivot_wider(names_from = covt, values_from = value) %>%
    mutate(year = as.numeric(year)) %>%
    left_join(visits, by = c("Name" = "Pentad", "year" = "year")) %>%
    filter(keep == 1) %>%
    dplyr::select(-keep) %>%
    arrange(lat, lon) %>%
    group_by(Name) %>%
    mutate(site = cur_group_id()) %>%
    ungroup() %>%
    group_by(year) %>%
    mutate(occasion = cur_group_id()) %>%
    ungroup() %>%
    data.table::as.data.table()

visit_data <- pa_dat %>%
    filter(year %in% unique(site_data$year)) %>%
    ungroup() %>%
    dplyr::left_join(site_data %>%
                         dplyr::select(Name, site) %>%
                         distinct(),
                     by = c("Pentad" = "Name")) %>%
    left_join(site_data %>%
                  dplyr::select(year, occasion) %>%
                  distinct(),
              by = "year") %>%
    group_by(site, occasion) %>%
    mutate(visit = row_number()) %>%
    ungroup() %>%
    data.table::as.data.table()


# Smooth for spatial effect on psi
fit <- fit_occu(list(psi ~ 1 + prcp + log(water+0.1) +
                         t2(lon, lat, occasion, k = c(25, 3), bs = c("ts", "cs"), d = c(2,1)),
                     p ~ 1 + log(TotalHours+0.1) + s(month, bs = "cs")),
                visit_data, site_data)
```