# Occupancy Models for African Bird Atlas Data

**Res Altwegg**

*res.altwegg@uct.ac.za*

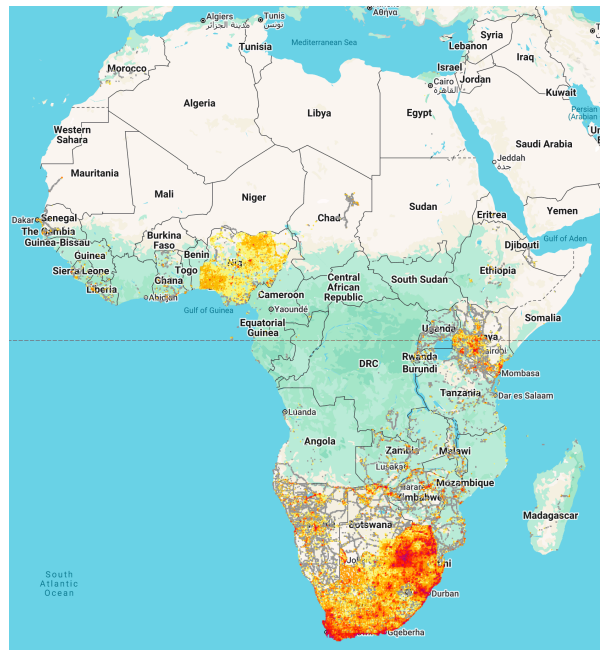
*Centre for Statistics in Ecology, Environment and Conservation, University of Cape Town*

September 3, 2026



## African Bird Atlas Project

- 2007 to now
- $\sim \frac{1}{2}$ million checklists
- $>$ 1 million records per year
- full protocol vs adhoc
- pentad $5^\circ \times 5^\circ$

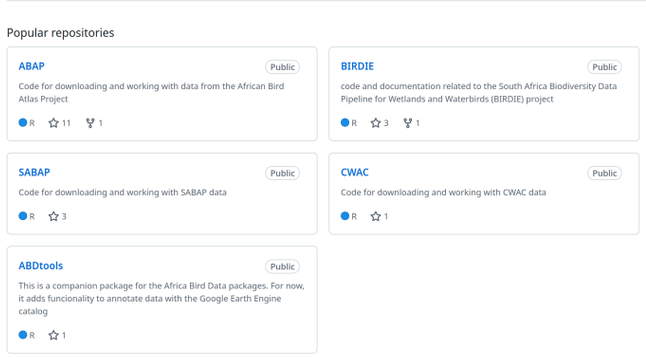





## AfricaBirdData – GitHub Organisation

**AfricaBirdData**

https://github.com/Afric




BIRDIE project

Cervantes et al 2024, Ostrich 95: 312-320



# Getting started...

## Download and install ABAP package:

```
install.packages("remotes")
remotes::install_github("AfricaBirdData/ABAP")
```

## Load ABAP package and some other useful ones:

```
library(ABAP)
library(sf)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2)
```



# Getting started

## Find species code

```
# We can search for all starling species
starlings <- searchAbapSpecies("Starling")
starlings$Common_species[26:35]
```

```
 [1] "Miombo Blue-eared" "Narrow-tailed"     "Neumann's"
 [4] "Pale-winged"       "Pied"              "Principe"
 [7] "Purple"            "Purple-headed"     "Red-winged"
[10] "Rosy"
```

```
# Then we can extract the code we are interested in
starlings[starlings$Common_species == "Pied", "Spp"]
```

```
# A tibble: 1 × 1
  Spp
  <chr>
1 746
```

Photo: Warwick Tarboton



# Getting the bird data

## Decide on scope of study

```
spp_code   <- 746 # pied starling
region     <- "Western Cape"
region_type <- "province"
years <- c(2023,2024)
```

## Then download ABAP data

```
my_Abap_data <- getAbapData(
  .spp_code     = spp_code,
  .region_type  = region_type,
  .region       = region,
  .years = years
)
```

Photo: Warwick Tarboton



# Format of the downloaded bird data

```
print(my_Abap_data, width = Inf)
```

```
# A tibble: 8,978 × 24
   CardNo                    StartDate  EndDate    StartTime Pentad    ObserverNo TotalHours Hour1 Hour
   <chr>                     <date>     <date>     <chr>     <chr>          <int>      <int> <int> <int
 1 3415_2135_020269_20230101 2023-01-01 2023-01-01 08:37:59  3415_2135      20269          2    12     2
 2 3420_2135_020269_20230101 2023-01-01 2023-01-01 09:54:48  3420_2135      20269          2    16     2
 3 3405_1855_014447_20230101 2023-01-01 2023-01-01 06:09:39  3405_1855      14447          2    27     3
 4 3400_1830_014447_20230101 2023-01-01 2023-01-01 11:48:16  3400_1830      14447          2    41     5
 5 3400_1845_014447_20230101 2023-01-01 2023-01-01 07:51:44  3400_1845      14447          2    30     4
 6 3405_1830_014447_20230101 2023-01-01 2023-01-01 12:11:55  3405_1830      14447          2    40     4
 7 3405_1845_014447_20230101 2023-01-01 2023-01-01 07:51:21  3405_1845      14447          5    46     5
 8 3355_2235_016065_20230101 2023-01-01 2023-01-01 08:43:50  3355_2235      16065          6    10     1
 9 3420_1905_020033_20230101 2023-01-01 2023-01-02 11:49:05  3420_1905      20033          3    33     4
10 3400_1850_000142_20230102 2023-01-02 2023-01-02 11:11:47  3400_1850        142          2    13     1
# ℹ 8,968 more rows
```

```
length(unique(my_Abap_data$Pentad)) # How many unique pentads?
```

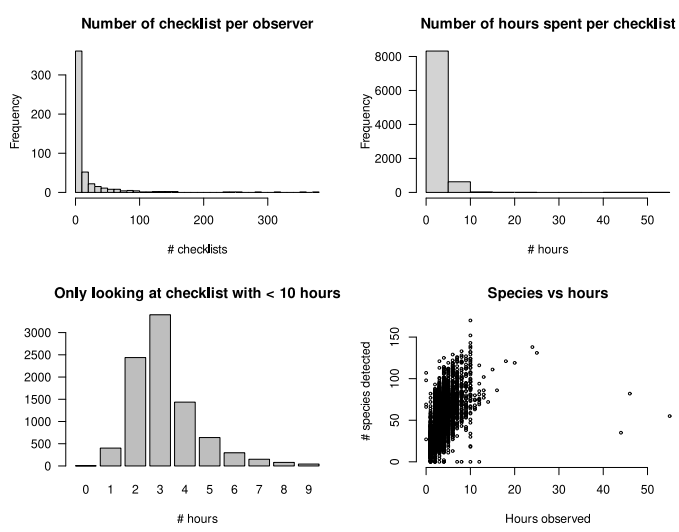
```
[1] 1129
```

```
length(unique(my_Abap_data$ObserverNo)) # How many unique observers
```

```
[1] 507
```



# Explore the data



# Some data cleaning

```
# Eliminate checklists with <2 hours effort
my_Abap_data <- my_Abap_data[my_Abap_data$TotalHours>1,]

# Eliminate checklists with 0 species detected
my_Abap_data <- my_Abap_data[my_Abap_data$TotalSpp>0,]
```



# Getting the pentad data

## Get the penda mask for the chosen region

```
WC_pentads <- getRegionPentads(.region_type = region_type, .region = region)
```

## Then join pentad info with bird data

```
visit_data <- my_Abap_data %>%
  left_join(WC_pentads, by = c("Pentad" = "pentad")) %>%
  st_as_sf()
```



# The data now look like this:

```
head(visit_data)
```

```
Simple feature collection with 6 features and 30 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: 18.5 ymin: -34.4 xmax: 21.7 ymax: -34
Geodetic CRS:  WGS 84
# A tibble: 6 × 31
  CardNo                    StartDate  EndDate    StartTime Pentad    ObserverNo TotalHours Hour1 Hour2 Hour
  <chr>                     <date>     <date>     <chr>     <chr>          <int>      <int> <int> <int> <int
1 3415_2135_020269_20230101 2023-01-01 2023-01-01 08:37:59  3415_2135      20269          2    12    21
2 3420_2135_020269_20230101 2023-01-01 2023-01-01 09:54:48  3420_2135      20269          2    16    20
3 3405_1855_014447_20230101 2023-01-01 2023-01-01 06:09:39  3405_1855      14447          2    27    32
4 3400_1830_014447_20230101 2023-01-01 2023-01-01 11:48:16  3400_1830      14447          2    41    50
5 3400_1845_014447_20230101 2023-01-01 2023-01-01 07:51:44  3400_1845      14447          2    30    49
6 3405_1830_014447_20230101 2023-01-01 2023-01-01 12:11:55  3405_1830      14447          2    40    47
```

- How many checklists do I have per pentad?
- Where has the species been seen?

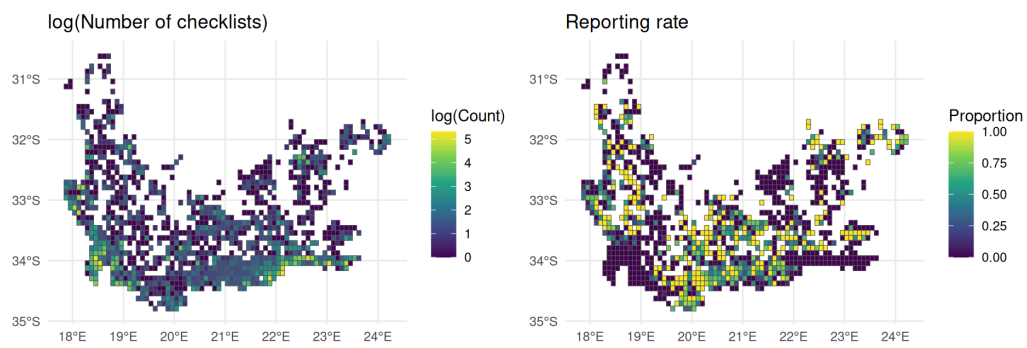
# Explore the raw data

```
# aggregate pdata er pentad
visit_data_ag <- visit_data %>%
  group_by(Pentad) %>%
  summarise(
    totalLists = n(),
    detected = sum(Spp==spp_code)
   )

# calculate reporting rate
visit_data_ag$ReportingRate <- visit_data_ag$detected/visit_data_ag$totalLists
```

$$\text{Reporting rate} = \frac{\text{Number of detections}}{\text{Number of checklists}}$$



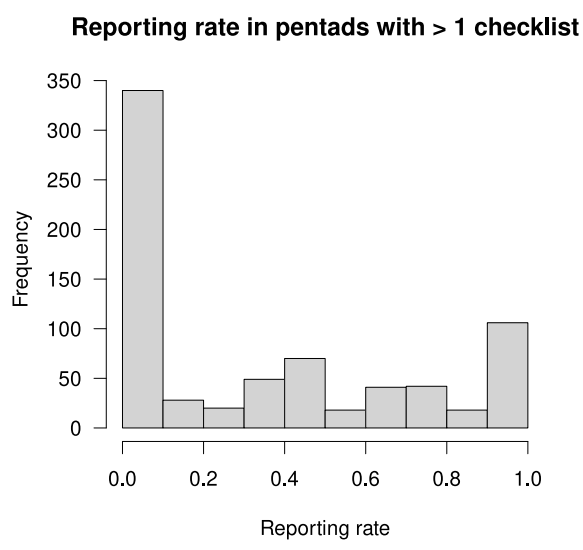
# Plot number of checklists and reporting rates in space





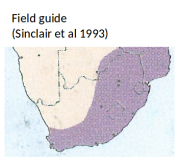

# Data exploration

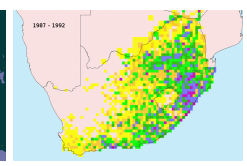
Photo: Warwick Tarboton



# Occupancy

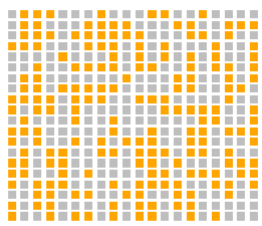
- Where a species occurs; which of a set of suitable patches are occupied; what determines where a species can live…
- (Metapopulation) ecology, conservation, red-listing





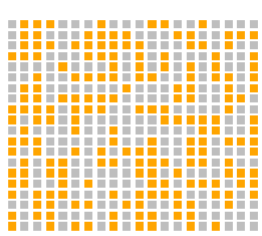



# Occupancy: the proportion of sites occupied by a species





# Occupancy: the proportion of sites occupied by a species



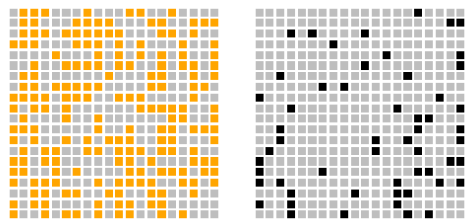
Occupancy:

$$\Psi = \frac{occupied}{total}$$

$$logit(\Psi) = f(covariates)$$

# The species is not detected in all occupied cells

Detection probability $p < 1$



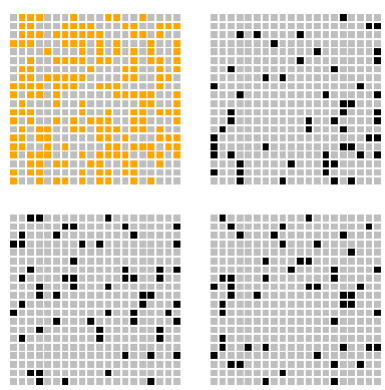
`Naive approach':

$$\Psi \times p = \frac{occupied}{total}$$

$$logit(\Psi \times p) = f(covariates)$$



# The species is not detected in all occupied cells



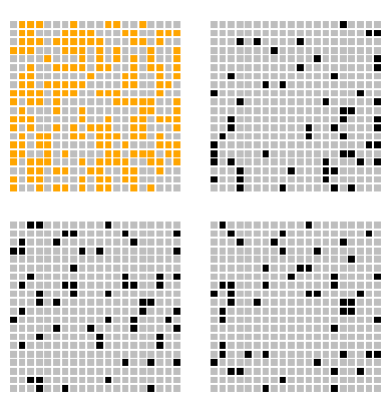
Repeated sampling

Assumptions:

- Closure (no colonisation or exctinction)
- No false detections



# The species is not detected in all occupied cells



Survey histories:

1 = detected

0 = not detected

- (1,1) 000
- (1,2) 001
- (2,2) 000

# Hierarchical model

- $\Psi_i =$ probability that site $i$ is occupied
- $z_i =$ true occupancy at site $i$: 1 = occupied, 0 = not occupied
- $p_{ij} =$ prob of detecting the species at site $i$ during survey $j$
- $y_{ij} =$ detection (1) or non-detection (0) at site $i$ during survey $j$

$$z_i|\Psi_i \sim Bern(\Psi_i)$$

$$y_{ij}|z_i, p_{ij} \sim Bern(z_i p_{ij})$$



# Covariate modelling

How do occupancy and detection vary among sites, $i$, and visits, $j$?

$$logit(\Psi_i) = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_U x_{iU}$$

$$logit(p_{ij}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_U x_{iU} + \beta_{U+1} y_{ij1} + \ldots + \beta_{U+V} y_{ijV}$$

- $U$ site-level covariates: $x_{i1}, \ldots, x_{iU}$
- $V$ survey-specific covariates: $y_{ij1}, \ldots, y_{ijV}$



# A closer look at assumptions

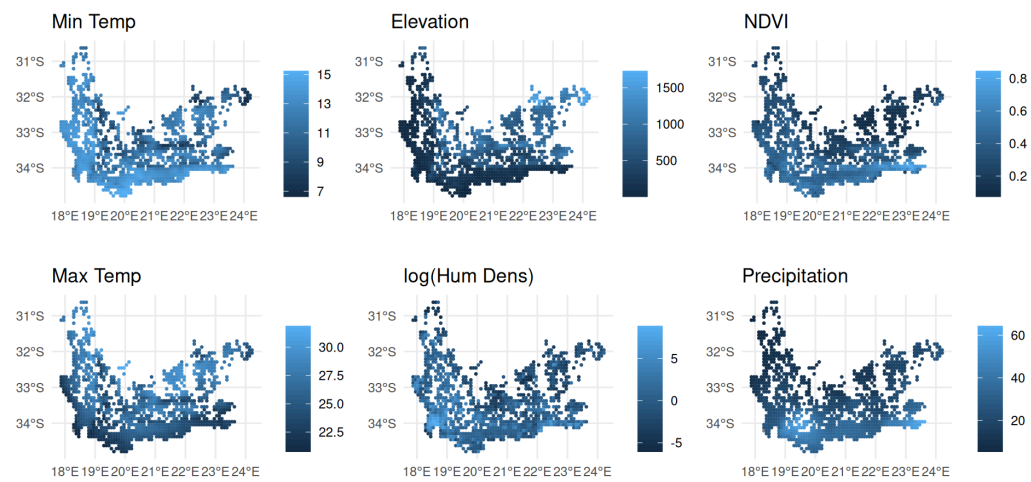
- Closure (no colonisation or extinction)
- No false detections
- No unmodelled heterogeneity in occupancy
- No unmodelled heterogeneity in detection
- Independent detections



# Site covariates

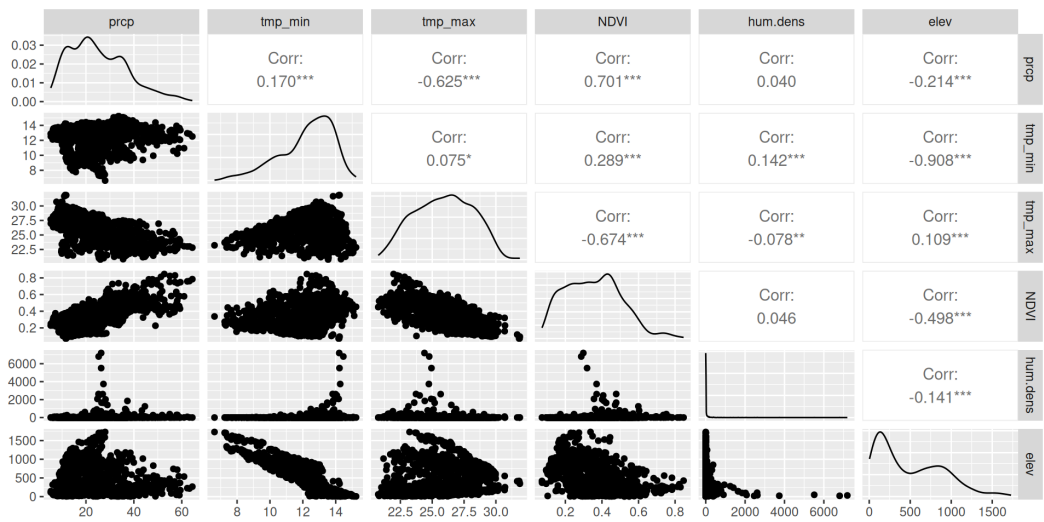
```r
site_covs <- read.csv("site_covs.csv")

# this drops any pentads not in the abap data of interest, and ensures the order matches
pentad_order_abap <- sort(unique(visit_data$Pentad))
site_covs_WC <- site_covs[match(pentad_order_abap, site_covs$pentad),]
```

# Site covariates






# Site covariates






# Get the data into the format `unmarked` needs

```
source("00_helper_functions_dataprep.R")

#check
stopifnot(identical(site_covs_WC$pentad, pentad_order_abap))

#need to drop the identifier as unmarked does not use an ID column
site_covs_WC <- site_covs_WC[, setdiff(names(site_covs_WC), "pentad"), drop = FALSE]

mySpecies_um <- abapToUnmarked_single2(
  abap_data = my_Abap_data,
  pentads   = WC_pentads,
  cap_visits = 50,       # caps visits at 50
  siteCovs = site_covs_WC,
  siteCovs_clean = TRUE  # selects complete cases
)
```

# Structure of `unmarked` object

```r
str(mySpecies_um)
```

```
Formal class 'unmarkedFrameOccu' [package "unmarked"] with 4 slots
  ..@ y       : int [1:1099, 1:50] 0 0 0 0 0 0 0 0 0 0 ...
  ..@ obsCovs :'data.frame':    54950 obs. of  3 variables:
  .. ..$ hours   : num [1:54950] 10 NA NA NA NA NA NA NA NA NA ...
  .. ..$ jday    : num [1:54950] 65 NA NA NA NA NA NA NA NA NA ...
  .. ..$ totalspp: int [1:54950] 18 NA NA NA NA NA NA NA NA NA ...
  ..@ siteCovs:'data.frame':    1099 obs. of  10 variables:
  .. ..$ x       : num [1:1099] 18.5 18.6 18.7 18.4 18.5 ...
  .. ..$ y       : num [1:1099] -30.6 -30.6 -30.6 -30.8 -30.8 ...
  .. ..$ lon     : num [1:1099] 18.5 18.6 18.7 18.4 18.5 ...
  .. ..$ lat     : num [1:1099] -30.6 -30.6 -30.6 -30.8 -30.8 ...
  .. ..$ prcp    : num [1:1099] 10.6 11.6 11.25 9.05 11.7 ...
  .. ..$ tmp_min : num [1:1099] 11.4 10.9 10.9 12.1 11.7 ...
  .. ..$ tmp_max : num [1:1099] 28.1 27.5 27.6 28.9 28.6 ...
  .. ..$ NDVI    : num [1:1099] 0.223 0.238 0.215 0.209 0.221 ...
  .. ..$ hum.dens: num [1:1099] 0.00614 0.00625 2.82966 0.00523 0.01921 ...
  .. ..$ elev    : num [1:1099] 725 876 869 438 594 ...
  ..@ obsToY  : num [1:50, 1:50] 1 0 0 0 0 0 0 0 0 0 ...
```



# Fit model using `unmarked`

```r
library(unmarked)

m1 <- occu(~ log(hours) ~ NDVI + tmp_max + tmp_min + log(hum.dens), data = mySpecies_um)
summary(m1)
```

```
Call:
occu(formula = ~log(hours) ~ NDVI + tmp_max + tmp_min + log(hum.dens),
    data = mySpecies_um)

Occupancy (logit-scale):
              Estimate     SE     z     P(>|z|)
(Intercept)    -11.714 2.3781 -4.93 0.000000840
NDVI            -5.966 1.3863 -4.30 0.000016790
tmp_max          0.488 0.0938  5.20 0.000000203
tmp_min          0.289 0.0713  4.05 0.000050961
log(hum.dens)   -0.299 0.0643 -4.65 0.000003377

Detection (logit-scale):
            Estimate     SE     z                P(>|z|)
(Intercept)    -0.94 0.1036 -9.07 0.0000000000000000012
log(hours)      0.65 0.0847  7.68 0.0000000000001578280
```



# Goodness-of-fit

... for another talk...

Photos: Warwick Tarboton

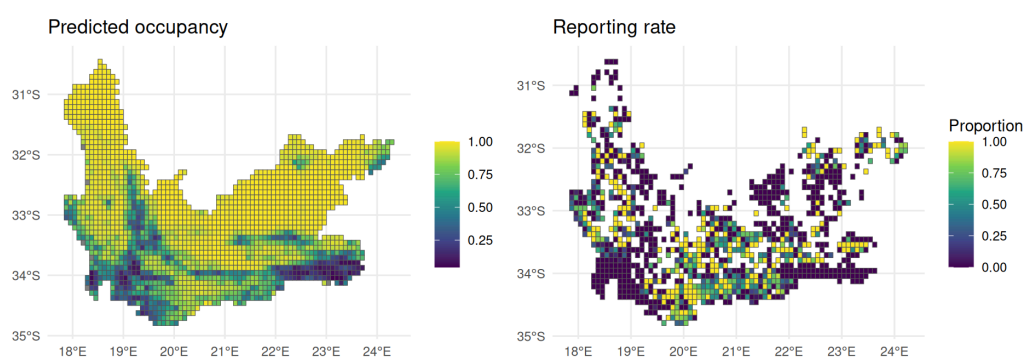


# Predictions

```r
WC_pentad_predictions <- WC_pentads %>%
  left_join(
    site_covs, by = c("pentad" = "pentad")) %>%
  st_as_sf()

pr <- predict(m1, type='state', newdata=as.data.frame(WC_pentad_predictions[,c("NDVI","tmp_min","tmp_ma

WC_pentad_predictions <- cbind(WC_pentad_predictions, pr)
```

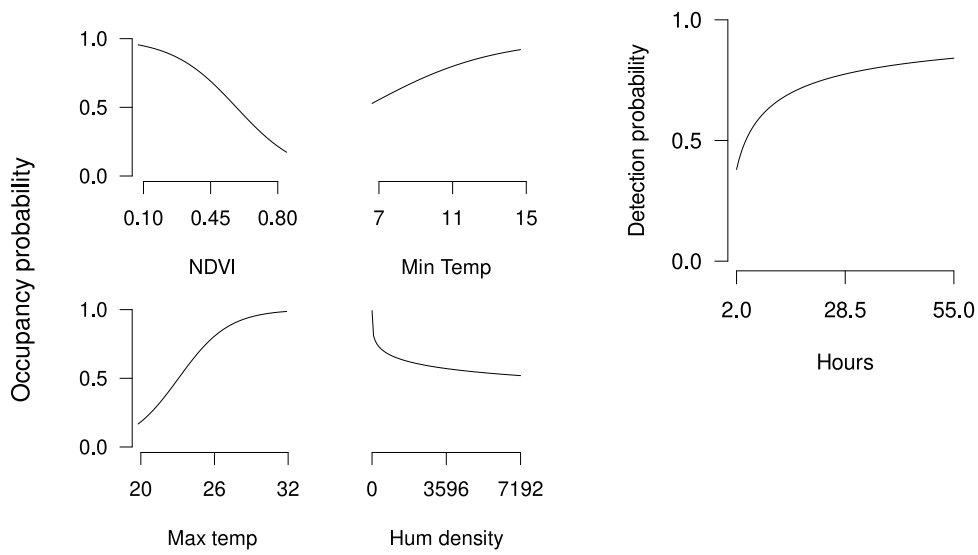


# Predictions






# Fitted covariate relationships



# Next steps

- land-use variables?
- interactions?
- quadratic effects?
- splines?

Warwick Tarboton



# Reading

The ABAP R packages


- Cervantes F, Lee AT, Henry DA, Brooks M, Visser V, Cloete A, et al. The Africa Bird Data packages: African citizen science bird data in R. Ostrich. 2024 Dec 20;95(4):312–20. doi:10.2989/00306525.2024.2423652


The ABAP project – description of protocol and data set


- Brooks M, Rose S, Altwegg R, Lee AT, Nel H, Retief E, et al. The African Bird Atlas Project: a description of the project and BirdMap data-collection protocol. Ostrich. 2022;93:223–32.


Occupancy models – the bible


- MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press; 2017.


Occupancy models for atlas data


- Altwegg R, Nichols JD. Occupancy models for citizen-science data. Methods Ecol Evol. 2019;10(1):8–21. doi:10.1111/2041-210X.13090