

SEEC stats toolbox seminar series:

# Building and using SDM ensembles: An introduction

**Mzabalazo Z. Ngwenya**

Centre for Statistics in Ecology, Environment and Conservation (SEEC)



Department of Statistical Sciences  
University of Cape Town



## 1 Introduction

- Species distribution modeling
- Ensemble SDMs

## 2 Creating a SDM ensemble from a single data set

- Data
- Modeling

## 3 Using multiple data sources to create SDMs:

- Setting
- Properties

# Introduction

## 1.1 Species distribution modeling

Species Distribution Modeling (SDM): Synonyms and related methods

- climate envelope modeling
- habitat modeling
- environmental/ecological niche modeling

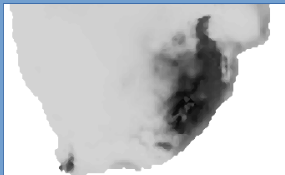
## Objectives for SDM:

- a) Inference and explanation
- b) Mapping and interpolation
- c) Forecasting

# SEEC Stats Toolbox introducing Species Distribution Modeling

## SEEC Stats Toolbox

Species distribution modelling in R





[Home](#) > [Stats Toolbox](#) > [Spatial and Species Distribution Toolboxes](#) > [Species Distribution Modelling](#)

Experimental and survey design  
Classification and regression trees  
Generalised Linear Mixed Models  
Generalised additive models (GAMs)  
Data exploration  
Analyses Toolboxes  
R packages and R related Toolboxes  
**[Spatial and Species Distribution Toolboxes](#)**  
**[Species Distribution Modelling](#)**  
SDMs - using spatial information to supplement biased occurrence data  
Occupancy models  
Distance sampling  
Handling Spatial Data  
Spatial capture-recapture (SCR) modelling  
Animal movement modelling with moveHMM  
Time-to-detection occupancy models  
Spatial occupancy models  
Single-season occupancy models using a Bayesian approach  
Spatial Interpolation

## Stats Toolbox Seminars

### Species Distribution Modelling

Species distribution modelling (SDM) is a burgeoning area of research in fields such as ecology, conservation, phylogeography and invasion biology. Simply put, SDMs use spatial occurrence data together with broadscale environmental data to predict spatial patterns of environmental suitability for species.

In our inaugural Stats Toolbox Seminar, Vernon Visser provided a brief introduction to SDMs. Below you can find the lecture slides and R script from this seminar. Provided in these materials is:

- A step-by-step guide to running your own SDM
- Suggestions for best practices
- References that can help provide more detail on the methods
- An R script that is annotated to make its understanding and adaptability easier

These materials should give you a major head start to running your own SDMs!

#### Presentation slides

**R scripts** (these are in a zip file. One file is the main script. The other file, "Zurell et al. 2012. SI R functions. DDI.txt" must be placed in the same directory as which you place the main script and which you will set as your working directory.

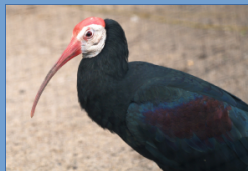
Share on



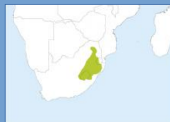
[Back to Top](#)

# Eight steps to your own SDM

1. Occurrence data
2. Environmental data
3. Background samples
4. Study extent
5. Data cleaning
6. Modelling
7. Checking your model
8. Projecting your model



Southern Bald Ibis  
*Geronticus calvus* (Boddaert, 1783)



[www.hbw.com](http://www.hbw.com)

## 1.2 Ensemble SDMs

### What are ensembles?

Ensemble modeling is the combining of models and model output from various algorithms.<sup>1</sup> The fitted models may be trained using various algorithms and or data sets.

### Benefits of ensembles

- Ensemble models generally have better predict performance than what would be obtained from a single model
- Can lead to better estimation of model parameters

---

<sup>1</sup> Ensemble modeling is closely related to model averaging where various models are also combined. The main distinction between ensembles and model averaging is that in ensembles models are built using various algorithms and/or data. In model averaging models are built using one algorithm and/or data set.



## Scenarios in which one may wish to you use SDM ensembles:<sup>2</sup>

### 1. Improve accuracy of predictions and obtain more robust forecasts:

- \* Elith, J. et al. 2006. Novel Methods Improve Prediction of Species' Distribution from Occurrence Data. *Ecography* 29: 129–51.
- \* Araujo, Miguel B., and Mark New. 2007. Ensemble Forecasting of Species Distributions. *Trends in Ecology and Evolution* 22: 42–47.

### 2. Use multiple data sets in order to create SDMs:

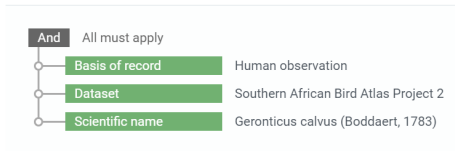
- \* Fletcher, R. J. et al 2019. A practical guide for combining data to model species distributions. *Ecology* 100:e02710
- \* Miller, D. A. W. et al 2019. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* 10:22–37.

---

<sup>2</sup>To enhance the assesment of the accuracy estimated (regression) parameters and the effect of predictors one would use model averaging instead of ensembles.

## Examples:

### 1. Scenario 1



### 2. Scenario 2



# Creating a SDM ensemble from a single data set

## 2.1 Data

Occurrence data:

[www.gbif.org](http://www.gbif.org) (Global Biodiversity Information Facility)

The collage consists of four overlapping screenshots from the GBIF website:

- Top-left:** The GBIF homepage with the header "From open access to biodiversity data" and statistics: Occurrence records: 1,665,550,073; Datasets: 57,331; Publishing institutions: 1,637; Peer-reviewed papers using data: 6,631.
- Top-right:** A species profile for *Geronticus calvus* (Boddaert, 1783), showing its classification and a distribution map.
- Bottom-left:** A search results page for *Geronticus calvus*, displaying a list of datasets and a map of its distribution in East Africa.
- Bottom-right:** A detailed search results table for *Geronticus calvus*, showing columns for dataset, occurrence ID, date, and coordinates.

| Dataset                 | Occurrence ID      | Date       | Coordinates          |
|-------------------------|--------------------|------------|----------------------|
| GBIF:100000000000000000 | 100000000000000000 | 2010-01-01 | 10.000000, 35.000000 |
| GBIF:100000000000000000 | 100000000000000000 | 2010-01-01 | 10.000000, 35.000000 |
| GBIF:100000000000000000 | 100000000000000000 | 2010-01-01 | 10.000000, 35.000000 |
| GBIF:100000000000000000 | 100000000000000000 | 2010-01-01 | 10.000000, 35.000000 |
| GBIF:100000000000000000 | 100000000000000000 | 2010-01-01 | 10.000000, 35.000000 |

Get data

How-to

Tools

Community

About

adw001

Occurrences

Administrative data management

Coordinate uncertainty in meters

Year

Month

Dataset

Search

☐ Southern African Bird Atlas Project 2 8,163
 ☐ Southern African Bird Atlas Project 3,748
 ☐ EOD - eBird Observation Dataset 2,123
 ☐ SAFRING: Historical Bird Ringing Records (2... 370
 ☐ Observation.org: Nature data from around th... 175
 ☐ iNaturalist Research-grade Observations 79
 ☐ iNaturalist 12
 ☐ eBird: eBird Count 9
 ☐ Xeno-canto: Bird sounds from around the world 1
 ☐ Nigerian Conservation Foundation (NCF) Data... 1

Country or area

<https://www.gbif.org/occurrence/search?basis=of record>

HUMAN OBSERVATION: Iason key: 2/60703

SEARCH OCCURRENCES | 14,883 RESULTS

TABLE

GALLERY

MAP

TAXONOMY

METRICS

DOWNLOAD

| Scientific name                           | Country or area | Coordinates  | Month & year  | Basis of record   |
|---|-----------------|--------------|---------------|-------------------|
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 29.25, 29.91 | 2021 February | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 33.86, 27.81 | 2020 January  | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 27.85, 30.91 | 2020 January  | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 28.25, 28.31 | 2020 January  | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 28.25, 28.31 | 2020 January  | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    | 27.58, 28.91 | 2020 January  | Human observation |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    |              |               |                   |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    |              |               |                   |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    |              |               |                   |
| <i>Geronticus calvus</i> (Boddaert, 1783) | South Africa    |              |               |                   |

Get data

How-to

Tools

Community

About

adw001

Running

CANCEL

FILTER APPLIED 23 MARCH 2021

RETURN QUERY

The download has been started and is currently being processed.

Please expect up to 3 hours for the download to complete. Most downloads will complete within 15 minutes.

A notification email with a link to download the results will be sent to the following address once ready: [mtzabalazo.ngwenya@gmail.com](mailto:mtzabalazo.ngwenya@gmail.com)

**Citation:** GBIF.org (23 March 2021) GBIF Occurrence Download <https://doi.org/10.15468/dl.jhmt3>

**License:** Unspecified

Make sure to read the [data user agreement](#) and [citation guidelines](#).

And

All must apply

Basis of record

Human observation

Dataset

Southern African Bird Atlas Project 2

Scientific name

*Geronticus calvus* (Boddaert, 1783)

## Reading in and cleaning occurrence data:

```
# Importing data, view and edit data file as needed
> SABAP2raw <- read.delim("FILEPATH")
> recs_SABAP2 <- SABAP2raw[,c( "decimalLongitude", "decimalLatitude", "species",
  "countryCode")]

# Cleaning and cross-checking
> library(CoordinateCleaner)
> subset(recs_SABAP2, !is.na(decimalLatitude))
> cl_recSABAP2 <- clean_coordinates(recs_SABAP2, lon="decimalLongitude",
  lat="decimalLatitude", countries="countryCode", tests=c("centroids", "outliers"))
> recs_SABAP2 <- recs_SABAP2[cl_recSABAP2$.summary,]
> head(recs_SABAP2)
```

|   | decimalLongitude | decimalLatitude | species           | countryCode |
|---|------------------|-----------------|-------------------|-------------|
| 1 | 28.20792         | -28.70792       | Geronticus calvus | ZA          |
| 2 | 30.12458         | -27.37458       | Geronticus calvus | ZA          |
| 3 | 28.54125         | -28.37458       | Geronticus calvus | ZA          |
| 4 | 30.54125         | -29.54125       | Geronticus calvus | ZA          |
| 5 | 28.45792         | -28.37458       | Geronticus calvus | ZA          |
| 6 | 29.45792         | -29.79125       | Geronticus calvus | ZA          |

## Viewing data:

```
#View these data on a map
> library(maptools)
> data(wrld_simpl) #Get the world map
> sa = wrld_simpl[wrld_simpl$ISO2%in%c('ZA','NA','BW','ZW','MZ','LS','SZ'),]
> plot(sa) #Plot southern African countries
> points(recs_SABAP2$decimalLongitude, recs_SABAP2$decimalLatitude, col='red')
```



# Environmental data:

<https://www.worldclim.org/> ( )



Home

They are coded as follows:

- BIO1 = Annual Mean Temperature
- BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- BIO3 = Isothermality (BIO2/BIO7) ( $\times 100$ )
- BIO4 = Temperature Seasonality (standard deviation  $\times 100$ )
- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter
- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

[Historical climate data](#)  
[Historical monthly weather data](#)  
[Future climate data](#)

## Bioclimatic variables:

```
# Downloading data
> library(raster)
> ext <- extent( c(0,60,-40,0))    #defining extent
> clim_curr <- getData("worldclim", var="bio", res=5, download=T)
> clim_curr <- crop(clim_curr,ext)

# Testing for collinearity
> library(usdm)
> cortest2 <-vifcor(clim_curr,0.7)
> cortest2

# Dropping collinear predictors as identified by vifcor()
> climSABAP2 <- exclude(clim_curr,cortest2)
> names(climSABAP2) #Final set of predictors

[1] "bio2" "bio3" "bio5" "bio8" "bio9" "bio15" "bio18" "bio19"
```



## Joining occurrence and climate data:

```
# Joining species and climate data
> data_SABAP2 <- cbind(recs_SABAP2, extract(x = climSABAP2, y =
  data.frame(recs_SABAP2[,c('decimalLongitude', 'decimalLatitude')]))))
> final_SABAP2 <- data_SABAP2[,c("decimalLongitude", "decimalLatitude", "species",
  "bio2", "bio8", "bio9", "bio15", "bio18", "bio19")]
> final_SABAP2$species <-1
> head(final_SABAP2)
```

|   | decimalLongitude | decimalLatitude | species | bio2 | bio3 | bio8 | bio9 | bio15 | bio18 | bio19 |
|---|------------------|-----------------|---------|------|------|------|------|-------|-------|-------|
| 1 | 28.20792         | -28.70792       | 1       | 147  | 54   | 198  | 89   | 62    | 333   | 42    |
| 2 | 30.12458         | -27.37458       | 1       | 138  | 57   | 173  | 88   | 70    | 369   | 38    |
| 3 | 28.54125         | -28.37458       | 1       | 149  | 55   | 179  | 77   | 61    | 325   | 42    |
| 4 | 30.54125         | -29.54125       | 1       | 125  | 57   | 215  | 141  | 60    | 368   | 56    |
| 5 | 28.45792         | -28.37458       | 1       | 150  | 54   | 181  | 77   | 61    | 318   | 41    |
| 6 | 29.45792         | -29.79125       | 1       | 146  | 55   | 182  | 85   | 76    | 509   | 44    |

## Joining occurrence and climate data (cont.):

```
# Making sure objects are spatial objects
> class(final_SABAP2)
[1] "data.frame"
> coordinates(final_SABAP2) <- ~decimalLongitude + decimalLatitude
> class(final_SABAP2)
[1] "SpatialPointsDataFrame"
attr(,"package")
[1] "sp"

# Checking and fixing projection issues
> projection(final_SABAP2)
[1] NA
> proj4string(final_SABAP2) <- projection(climSABAP2)
> projection(final_SABAP2)
[1] "+proj=longlat +datum=WGS84 +no_defs"
```

## 2.2 Modeling

### Model fitting and assessment:

```
> library(sdm)
> # installAll() use only if running sdm for the first time

> SABAP2.sdm <- sdmData(species~., train = final_SABAP2, predictors = climSABAP2,
  bg = list(n=1000))

# Fitting models via four algorithms;
# getmethodNames() displays all available algorithms
> fit_SABAP2 <- sdm(species~., SABAP2.sdm, methods=c("gam","svm","rf","mars"),
  replication=c("boot"), n=10)
> fit_SABAP2
> roc(fit_SABAP2)
```

```

class                                : sdmModels
=====
number of species                    : 1
number of modelling methods          : 4
names of modelling methods           : gam, svm, rf, mars
replicate.methods (data partitioning) : bootstrap
number of replicates (each method)   : 10
total number of replicates per model : 10 (per species)
-----
model run success percentage (per species) :
-----
method      species
-----
gam         :      100   %
svm         :      100   %
rf          :      100   %
mars        :      100   %

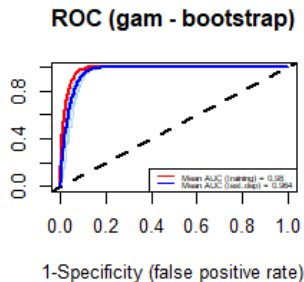
#####
model Mean performance (per species), using test dataset (generated using partitioning):
-----

## species : species
=====

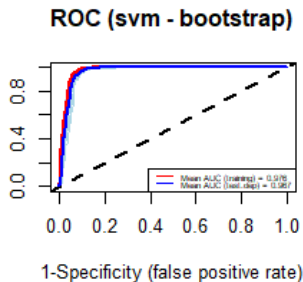
methods :      AUC   |      COR   |      TSS   |      Deviance
-----
gam      :      0.98   |      0.94   |      0.94   |      0.33
svm      :      0.98   |      0.95   |      0.94   |      0.26
rf       :      0.99   |      0.96   |      0.96   |      0.16
mars     :      0.98   |      0.94   |      0.93   |      0.24

```

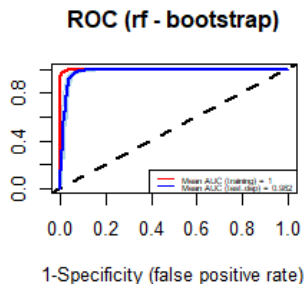
Sensitivity (true positive rate)



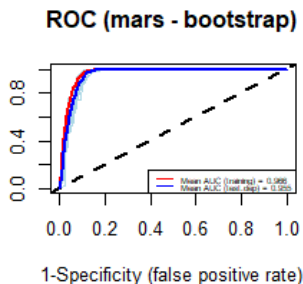
Sensitivity (true positive rate)



Sensitivity (true positive rate)



Sensitivity (true positive rate)



## Making and using ensembles

```
# Current species distribution
```

```
> SABAP2curr <- ensemble(fit_SABAP2, climSABAP2, filename = 'ens.img',  
  setting = list(method='weighted', stat='AUC')) #creating ensemble  
> plot(SABAP2curr)  
> points(final_SABAP2)
```

```
# Future species distribution
```

```
> clim_fut <- raster::getData("CMIP5", var='bio', res=5, model='AC', rcp=85, year=70)  
> names(clim_fut) <- names(clim_curr)  
> clim_fut <- exclude(clim_fut, cortest2)  
> clim_fut <- crop(clim_fut, ext)
```

```
# Ensemble prediction
```

```
> SABAP2fut <- ensemble(fit_SABAP2, clim_fut, 'ensf.img', setting =  
  list(method='weighted', stat='AUC'))  
> plot(SABAP2fut)  
> points(final_SABAP2)
```

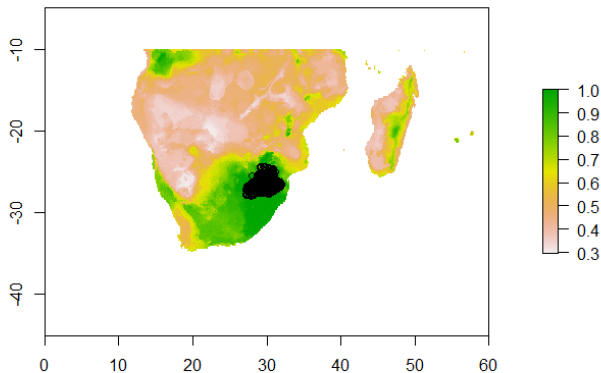


Figure 1: Current predicted distribution of *Geronticus calvus*.

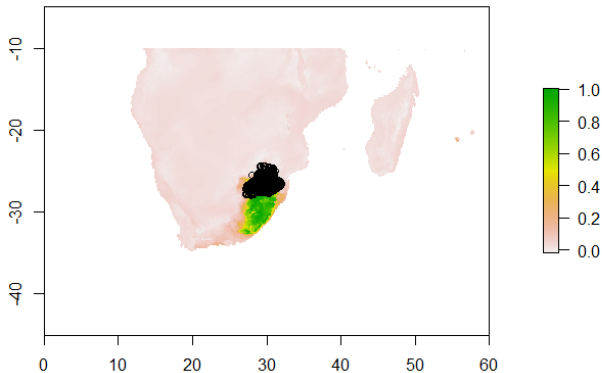
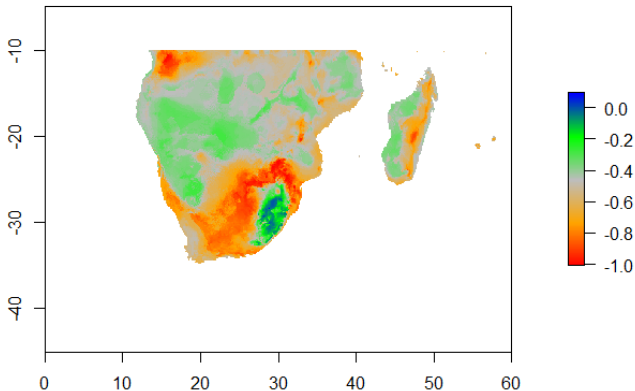


Figure 2: Forecasted distribution of *Geronticus calvus*.



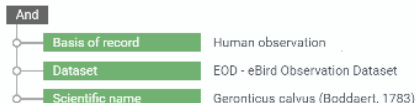
## Inference on range changes:

```
# Compare current and future distributions  
> rangeChange <- SABAP2fut - SABAP2curr  
> clz <- colorRampPalette(c('red', 'orange', 'gray', 'green', 'blue'))  
> plot(rangeChange, col=clz(200))
```



# Using multiple data sources to create SDMs

## 3.1 Setting



## 3.2 Properties

TABLE 1. Some characteristics of different approaches for combining data.

| Characteristic  | Simple pooling | Independent models | Auxiliary data | Informed priors | Integrated models |
|---|----------------|--------------------|----------------|-----------------|-------------------|
| Can account for different sampling issues                           | No             | Yes                | Yes            | Yes             | Yes               |
| Can account for variation in spatial or temporal support among data | No             | Yes                | No             | Yes             | Yes               |
| Can account for uncertainty from both data sources                  | No             | No                 | No             | Yes             | Yes               |
| Can allow for different predictors for each data source             | No             | Yes                | Yes            | Yes             | Yes               |
| Sequential vs. simultaneous modeling of data sources                | Simultaneous   | Sequential         | Sequential     | Sequential      | Simultaneous      |

Fletcher et al (2019)

# Additional references



Araujo M.B., Anderson R.P., Marcia Barbosa A., Beale et al (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, **5**.



Báez J.C., Barbosa A.M., Pascual P., Ramos M.L., Abascal F. (2020). Ensemble modeling of the potential distribution of the whale shark in the Atlantic Ocean. *Ecology and Evolution*, **10**, 175-184.



Naimi B. and Araujo M.B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, **39**, 368-375.



Thuiller W., Guéguen M., Renaud, J. et al. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, **10**, 1446.