

# Preparing data for ENM/SDM

Species and climate data

Emilio Berti

# Data types

- ▶ Species occurrence (0/1; absences/presence) data. Usually a `data.frame`.
- ▶ Environmental (climatic) data. Usually a `SpatRaster`.

Species data

# Species occurrence data

*Podarcis muralis* (Laurenti, 1768) from GBIF:

<https://doi.org/10.15468/dl.x74f4b>. Filters:

- ▶ The coordinate uncertainty of the records must be  $\leq 5km$ .
- ▶ The year of the record must be  $\geq 1970$  and  $\leq 2000$ .



Figure 1: The common wall lizard *Podarcis muralis*.

# Species occurrence data

Load GBIF data into R.

```
library(terra)

gbif <- read.csv(
  "../data/0002051-260120142942310.csv",
  sep = "\t"
)
```

## **i** Note

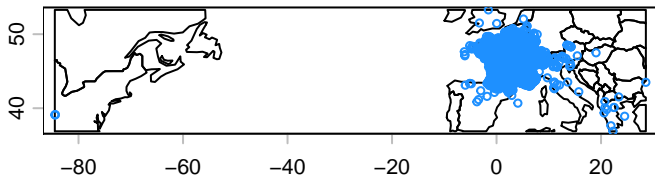
`sep = "\t"` specifies that the separator of the columns is a TAB, which is the standard used by GBIF.

This data frame has many columns that we do not need.

## Drop duplicates

Keep only long/lat columns and drop duplicate coordinates.

```
gbif <- gbif[, c("decimalLongitude", "decimalLatitude")]  
gbif <- unique(gbif)
```



### Warning

This is an European species.

*George Rau, a boy [...] returned from a family vacation to northern Italy.*

## Check for coordinate issues

The package `CoordinateCleaner` performs several quality checks on GBIF data and flags potential inaccuracies.

```
library(CoordinateCleaner)

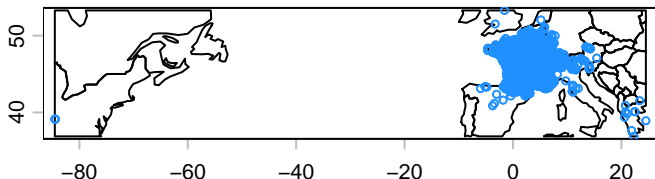
flags <- clean_coordinates(
  gbif,
  species = NULL,
  tests = c(
    "capitals", "centroids", "equal", "gbif",
    "institutions", "seas", "zeros"
  )
)
```

The data frame `flags` contains the column `.summary` with value `TRUE/FALSE`.

## Remove coordinates with issues

Retain only GBIF records that have `.summary = TRUE`.

```
gbif <- gbif[flags$.summary, ]
```



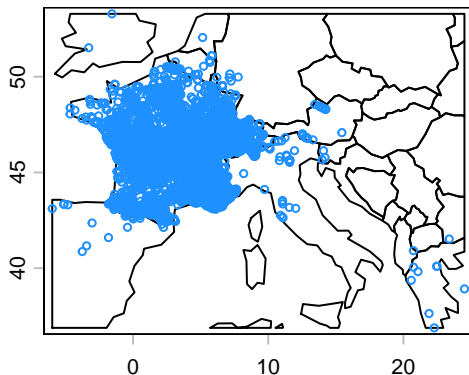
There are still points in the USA, which we want to remove manually.



## Remove coordinates manually

We can remove inaccuracies manually.

```
gbif <- gbif[gbif$decimalLongitude >= -20, ]
```



We have now a data frame of cleaned occurrences from GBIF.

# Pseudo-absences

- ▶ Generally, we need both presences and absences of species for ENM.
- ▶ We need to generate some *pseudo-absences*, i.e. simulated absences, and add them to the data frame.
- ▶ We do not want absences in the same location (grid cell) of presences.

# Pseudo-absences

Define the sampling region.

```
# data frame as SpatVector
gbif <- vect(
  gbif,
  geom = c("decimalLongitude", "decimalLatitude"),
  crs = "EPSG:4326"
)

# (convex) hull inscribing all known occurrences
hull <- convHull(gbif)
```

## Pseudo-absences

Rasterize GBIF data to a 0 (at least one record) 1 (no records) layer.

```
# load one climate layer as template of the grid cell
grid <- rast("../data/wc2.1_10m_bio_1.tif") |> crop(hull)

# create a raster with
# - 0 if there is a gbif record in that cell
# - 1 if not
# - NA for sea cells
r <- rasterize(gbif, grid, fun = \(x) 0, background = 1)
r[is.na(grid)] <- NA

# remove areas outside the polygon inscribing all GBIF records
r <- mask(r, hull)
```

# Pseudo-absences

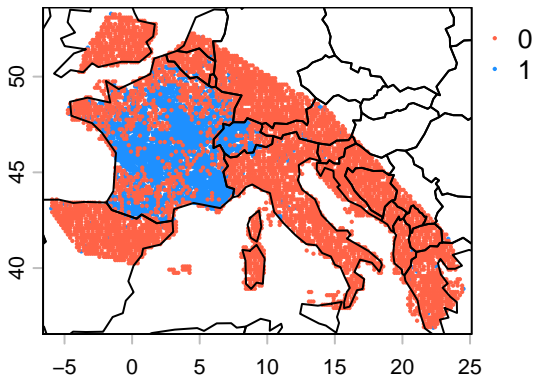
Sample within the area excluding cells with at least one presence.

```
# sample absences
abs <- spatSample(
  r,
  length(gbif),      # n(abs) = n(pres)
  as.points = TRUE,  # return a SpatVector
  method = "weights", # trick to remove cells with a record
  values = FALSE      # we do not care about the values of
)
```

## Finalize species occurrence data

Stitch the two SpatVector together.

```
gbif$occ <- 1 # presence  
abs$occ <- 0 # absence  
  
# combine into one SpatVector  
p <- rbind(gbif, abs)
```



Climate data

# Climate data

- ▶ WorldClim bioclimatic variables.
- ▶ Many others are available, but WorldClim is easy to start with.
- ▶ Bioclimatic variables are derived temperature and precipitation variables that have the strongest influence on species.
- ▶ A list of all of them can be found at <https://www.worldclim.org/data/bioclim.html>.



# Climate data

Load eight bioclimatic variables.

```
# list of files of bioclimatic variables
ff <- list.files(
  "../data",          # where the files are
  pattern = "wc2.1",  # wc = WorldClim
  full.names = TRUE   # full path
)
```

```
# load them into memory
climate <- rast(ff)
climate
```

```
class       : SpatRaster
size        : 1080, 2160, 8  (nrow, ncol, nlyr)
resolution  : 0.1666667, 0.1666667  (x, y)
extent      : -180, 180, -90, 90  (xmin, xmax, ymin, ymax)
coord. ref. : lon/lat WGS 84 (EPSG:4326)
sources     : wc2.1 10m bio 1.tif
```

## Extract climate

Extract the values of climate at the species occurrence locations using `p`.

```
d <- extract(climate, p, ID = FALSE, cell = TRUE)
```



`cell = TRUE` return also the ID of the cell of the raster where the records are found. This is useful to keep only one record per grid cell.

## Assign species occurrences

Assign the occurrence status (presence/absence) to this data frame.

```
d$occ <- p$occ
```

# Drop duplicates

Drop duplicate records, i.e. multiple records for the same grid cell.

```
# drop rows with duplicated cells
d <- d[!duplicated(d$cell), ]

# drop the `cell` column
d <- d[, -which(names(d) == "cell")]
```

## Balance absences and presences

Finally, make sure to have more or less the same number of presences and absences.

```
table(d$occ) # not balanced
```

```
      0      1  
5016 1644
```

```
# subsample  
n <- table(d$occ)[["1"]]  
index_pres <- which(d$occ == 1)  
index_abs <- which(d$occ == 0)  
d <- d[c(index_pres, sample(index_abs, n)), ]  
table(d$occ) # balanced
```

```
      0      1  
1644 1644
```

# Workflow

## Occurrence data

1. Download GBIF data.
2. Filter and clean GBIF data.
3. Sample absences (if needed).

## Climate data

1. Download climate data.
2. Extract values for occurrence data.
3. Drop duplicates.

Occurrence + climate data  $\rightarrow$  ENM.