Data Retrieval for Marine Species Distribution Modelling

Rutendo Musimwa, Willem Boone and Johannes Nowe

2024-09-24

Contents

Install packages and setup environment **1**

1. OCCURENCE DATA **1**

2. ENVIRONMENTAL DATA **4**

Discover more **6**

Install packages and setup environment

```
#Using the pak R package, multiple packages can be downloaded easily
if(!require('pak'))install.packages('pak')
# install and load packages
pckgs <- c("arrow", "tidyverse", "doParallel", "rasterVis", "mapview",
           "ENMeval", "dynamicSDM", "gridExtra", "raster", "spThin",
           "BiocManager","Rarr","worrms","stars","foreach","terra",
           "formatR","ggplot2","rnaturalearth","rnaturalearthdata","ggspatial")

pak::pkg_install(pckgs)

invisible(lapply(pckgs, library, character.only = TRUE))
rm(pckgs)
```

1. OCCURENCE DATA

This document demonstrates how to retrieve occurrence data for marine species from the EDITO platform.

Step 1: Establish data lake connection (S3F)

use `S3FileSystem` from package `arrow` to connect tot the data lake.

```
data_lake <- S3FileSystem$create(anonymous = TRUE,
                                 scheme = "https",
                                 endpoint_override = "s3.waw3-1.cloudferro.com")
```

Step 2: Access EurOBIS occurrence data stored in the parquet file.

The EurOBIS data is stored at following location in the data lake:

`emodnet/biology/eurobis_occurence_data/eurobisgeoparquet/eurobis_no_partition_sorted.parquet`

Using this address, you can open the dataset using `arrow::open_dataset`

```
path_to_eurobis = file.path("emodnet",
                            "biology",
                            "eurobis_occurence_data",
                            "eurobisgeoparquet",
                            "eurobis_no_partition_sorted.parquet")

eurobis <- arrow::open_dataset(data_lake$path(path_to_eurobis))
```

Step 3: Filter the occurrences

You can filter the data set on following criteria:

- `aphiaidaccepted`: int
- `latitude` & `longitude`: int or float
- `Date`: string formatted date `YYYY-MM-DD`

Some example Aphia IDs

Species	Aphia ID
Atlantic mackerel	127023
Atlantic herring	126417
European seabass	126975

More Aphia IDs can be found using the `worrms` package to query the worms database. Searching on scientific name or common name:

```
# Searching on scientific name
worrms::wm_name2id("Clupea harengus")
```

```
## [1] 126417
```

```
# Searching on common name
worrms::wm_records_common("Atlantic herring")|> dplyr::select("AphiaID","scientificname")
```

```
## # A tibble: 1 x 2
##   AphiaID scientificname
##     <int> <chr>
## 1  126417 Clupea harengus
```

Define your parameters here: For example for herring:

```
aphia_ID = 126417
sel_longitude = c(-15, 30)
sel_latitude = c(35, 65)
start_date = "2010-01-01"
end_date = "2020-12-31"
```

Perform the selection:

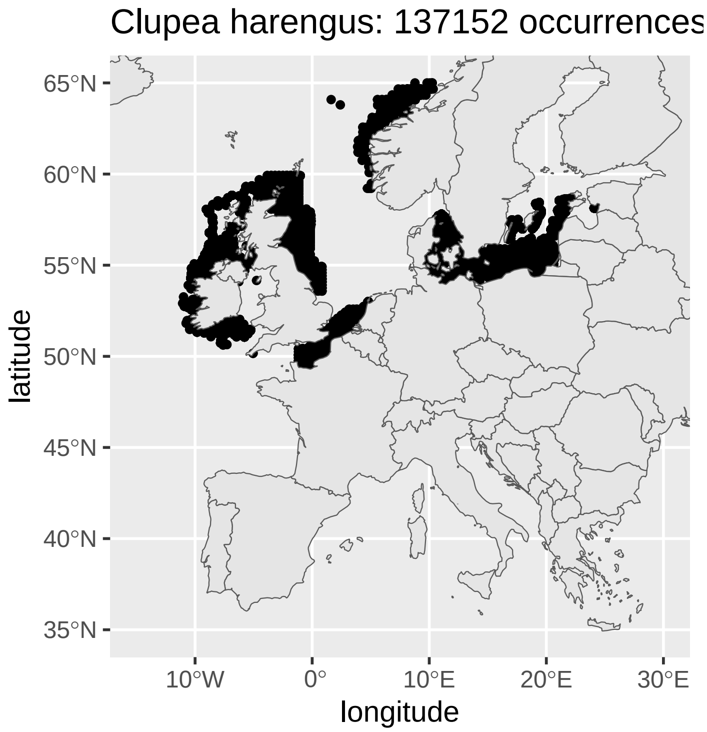
```
my_selection <- eurobis |>
  filter(aphiaidaccepted == aphia_ID,
         longitude > sel_longitude[1],
         longitude < sel_longitude[2],
         latitude > sel_latitude[1],
         latitude < sel_latitude[2],
         observationdate >= as.POSIXct(start_date),
         observationdate <= as.POSIXct(end_date)) |>
  collect()
```

Inspect the selection, for herring you should have `137.152` occurrences between 2010-2020 in region of interest (lat(-20; 40), lon(30; 65)).

Step 4: Visualize your selection on a map

Plot the data on a map using the `ggplot` package. Notice that plotting might take some time, depending on the number of records in the dataset. An interactive plot can be made using the `mapview` package.

```
#for an interactive plot:
#mapview(my_selection$longitude, my_selection$latitude, crs = "epsg:4326")
europe <- ne_countries(continent="europe",scale="medium")
ggplot()+
  geom_point(data=my_selection,aes(x=longitude,y=latitude),size=1)+
  geom_sf(
    data = europe
  )+
  coord_sf(sel_longitude, sel_latitude)+
  labs(title= paste0(my_selection$scientificname[1],
                     ": ",nrow(my_selection),
                     " occurrences"))
```



2. ENVIRONMENTAL DATA

Step 1: Source editoTools

Source this R script, it contains several useful functions to interact with the data lake.

```r
source("editoTools.R")
```

Step 2: inspect EDITOSTAC

`EDITOSTAC` is a data frame from `editoTools` and contains a library of all available datasets.

Look at the data frame in your environment to see which datasets are available.

Step 3: set search parameters

Following parameters are used:

- `variable`: String (e.g. `thetao`, `so`, `zooc`, `phyc`)
- `StacCatalogue`: use `EDITOSTAC` which is a data frame created in `editoTools`.
- `lon_min`: int or float.
- `lon_ma`: int or float.
- `lat_min`: int or float.
- `lat_max`: int or float.
- `requestedTimeSteps`: can be one of `NA 86400000 21600000 3600000 10800000 900000 604800000`

- `date`: single datestring
- `select_layers`: NULL or int, if multiple layers available it will ask which one you want, can pre-select the first, second, … layer

```
variable = "thetao"
stacCatalogue = EDITOSTAC
lon_min = -12
lon_max = 10
lat_min = 48
lat_max = 62
requestedTimeSteps = 3600000
date = "2020-09-01"
```

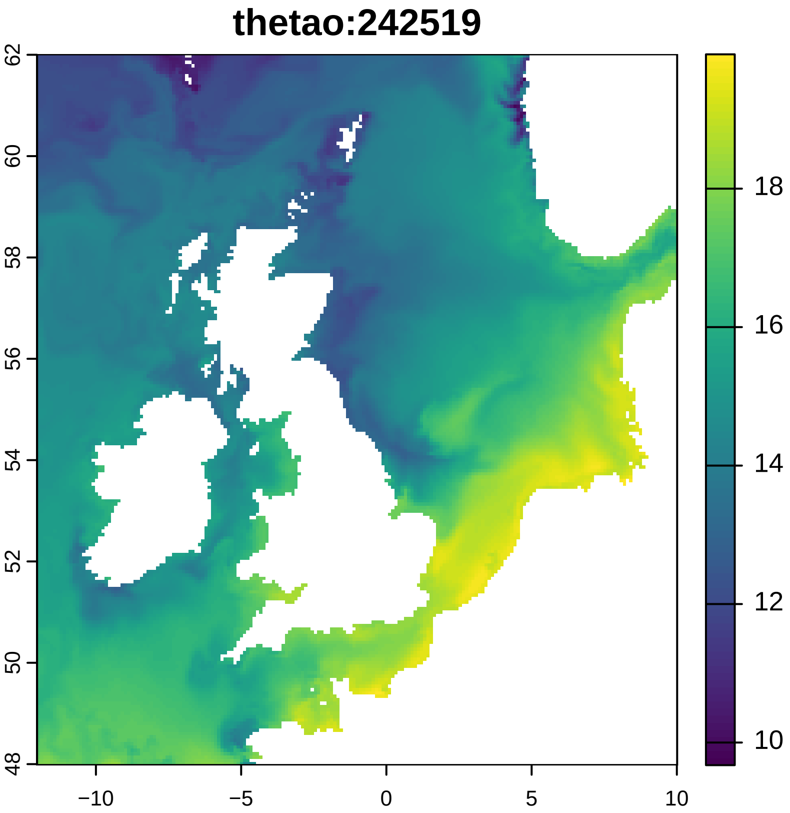
Step 3: request raster values

This function might give you some warnings, but it runs just fine.

```
raster_example <- getRasterSlice(variable,
                                   stacCatalogue = EDITOSTAC,
                                   lon_min = lon_min,
                                   lon_max = lon_max,
                                   lat_min = lat_min,
                                   lat_max = lat_max,
                                   requestedTimeSteps = requestedTimeSteps,
                                   date = date,
                                   select_layers = NULL)
```

Step 4: plot the raster as a map

```
#for an interactive plot: mapview(raster_example)
terra::plot(raster_example,main=names(raster_example))
```



Discover more

`EditoTools.R` contains several other useful functions such as `enhanceDF` which are demonstrated in the tutorial "*Using EditoTools for predictive modelling of Atlantic herring larvae in the North Sea*"**.**

» Go a head and explore this notebook