Tutorial for SDMPlay: 1/ Compute Species Distribution Models

2021-02-01

Species distribution modelling (SDM) has been developed for several years to address conservation issues, to assess the direct impact of human activities on ecosystems and to predict the potential distribution shifts of invasive species (see Elith et al. 2006, Pearson 2007, Elith and Leathwick 2009). SDM relates species occurrences with environmental information and can predict species distribution on their entire occupied space. **This approach has been increasingly applied to Southern Ocean case studies, but requires corrections in such a context, due to the broad scale area, the limited number of presence records available and the spatial and temporal aggregations of these datasets.**

SDMPlay is a pedagogic package that will allow you to compute SDMs, to understand the overall method, and to produce model outputs. The package, along with its associated vignettes, highlights the different steps of model calibration and describes how to choose the best method to generate accurate and relevant outputs. SDMPlay proposes codes to apply a popular machine learning approach, BRT (Boosted Regression Trees) and introduces MaxEnt (Maximum Entropy). It contains occurrences of marine species and environmental descriptor datasets as examples associated with several vignette tutorials.

- **Tutorial #1/ Compute Species Distribution Models**
  Focusses on data structure, data preparation and general model computing.
- **Tutorial #2/ SDM outputs**
  Presents the main outputs you can generate with your SDM.
- **Tutorial #3/ Importance of model calibration**
  Highlights the procedure to accurately calibrate your model and proposes some methods to limit the influence of several biases.
- **Tutorial #4/ Spatial cross-validation**
  Cross-validation is a method to validate your model. When working with presence data spatially aggregated, the cross-validation procedure should be adapted. This tutorial provides some elements to apply this method (refering to Guillaumot et al. 2019).
- **Tutorial #5/ Spatial extrapolation**
  Models can extrapolate when projected on broad scale areas. This tutorial provides codes to calculate extrapolation scores and generate extrapolation maps that could be associated to SDM maps (refering to Guillaumot et al. 2020).

Data overview

Occurrence records

In the package, you can download the presence-only records of two echinoid species of the Kerguelen Plateau, *Brisaster antarcticus* and *Ctenocidaris nutrix* and the presence-only records of the sea stars *Odontaster validus* and *Glabraster antarctica*, distributed at the scale of the Southern Ocean. These species present contrasting ecological niches, with different feeding preferences and reproductive behaviours (David et al. 2005, Mah and Blake 2012). The complete dataset of Kerguelen echinoid species is available in Guillaumot et al. (2016), the complete dataset of Southern Ocean sea stars is available in the updated database of Moreau et al. (2018).

```r
library(SDMPlay)
data("ctenocidaris.nutrix") # Species distributed on the Kerguelen Plateau, table with
                            # longitude, latitude and several other columns
head(ctenocidaris.nutrix)
```

```
##    id     scientific.name scientific.name.authorship
## 56  1 Ctenocidaris_nutrix             (Thomson 1876)
## 57  2 Ctenocidaris_nutrix             (Thomson 1876)
## 58  3 Ctenocidaris_nutrix             (Thomson 1876)
## 59  4 Ctenocidaris_nutrix             (Thomson 1876)
## 60  5 Ctenocidaris_nutrix             (Thomson 1876)
## 61  6 Ctenocidaris_nutrix             (Thomson 1876)
##                           genus                        family
## 56 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
## 57 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
## 58 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
## 59 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
## 60 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
## 61 Ctenocidaris Mortensen 1910 Ctenocidarinae Mortensen 1928
##    order.and.higher.taxonomic.rank decimal.Longitude decimal.Latitude depth
## 56            Cidaroida Claus 1880          67.13167        -48.98500   315
## 57            Cidaroida Claus 1880          67.33167        -49.44167   301
## 58            Cidaroida Claus 1880          67.51167        -49.00500   206
## 59            Cidaroida Claus 1880          67.54167        -48.11667   365
## 60            Cidaroida Claus 1880          67.88500        -49.46667   191
## 61            Cidaroida Claus 1880          68.05833        -49.06667   178
##    year       campaign             reference          vessel
## 56 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
## 57 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
## 58 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
## 59 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
## 60 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
## 61 1975 MD04 (BENTHOS) De Ridder et al. 1992 Marion Dufresne
```

You can similarly load data for Southern Ocean distributed species

```r
data("Odontaster.validus") # Species distributed around the entire Southern Ocean, table
                           # with longitude and latitude only
head(Odontaster.validus)
```

```
##   longitude latitude
## 1  166.6492 -77.8504
## 2  166.6492 -77.8504
## 3  166.6492 -77.8504
## 4  166.6492 -77.8504
## 5  166.6492 -77.8504
## 6  166.4818 -77.4319
```

In which concerns environmental descriptors, two regions are presented in this package: the Kerguelen Plateau and the Southern Ocean. For the Kerguelen Plateau, the environmental dataset compiles 15 environmental descriptors, displayed in a raster format, for three time periods [1965-1974], [2005-2012], and for the future climatic scenario AIB (IPCC, 4th report 2007) for 2200. Grid-cell pixels are set at a 0.1° resolution and data were not interpolated (presence of N/A values in the area). Extra metadata and environmental layers are available in Guillaumot et al. (2016).

Environmental layers

Load the raster stacks

```r
library(raster)
data(predictors1965_1974)
data(predictors2005_2012)
data(predictors2200AIB)
```

Observe their content

```r
predictors2005_2012
```

```
## class      : RasterStack
## dimensions : 100, 179, 17900, 15  (nrow, ncol, ncell, nlayers)
## resolution : 0.1, 0.1  (x, y)
## extent     : 63, 80.9, -56, -46  (xmin, xmax, ymin, ymax)
## crs        : +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0
## names      :         depth, seasurfac//_2005_2012, seasurfac//_2005_2012, seafloor_//_2005_2012, sea
## min values : -4.977000e+03,          3.263100e-01,         -4.002820e+00,         -2.985100e-01,
## max values :    -1.0000000,             9.0129099,            -1.2873000,             4.8823700,
```

```r
names(predictors2005_2012)
```

```
##  [1] "depth"
##  [2] "seasurface_temperature_mean_2005_2012"
##  [3] "seasurface_temperature_amplitude_2005_2012"
##  [4] "seafloor_temperature_mean_2005_2012"
##  [5] "seafloor_temperature_amplitude_2005_2012"
##  [6] "seasurface_salinity_mean_2005_2012"
##  [7] "seasurface_salinity_amplitude_2005_2012"
##  [8] "seafloor_salinity_mean_2005_2012"
##  [9] "seafloor_salinity_amplitude_2005_2012"
## [10] "chlorophyla_summer_mean_2005_2012"
## [11] "geomorphology"
## [12] "sediments"
## [13] "slope"
## [14] "seafloor_oxygen_mean_2005_2012"
## [15] "roughness"
```

You can select only a part of the layers with the 'subset' function of the raster package.

```r
layer_ex <- raster::subset(predictors2005_2012, c(1:4))
plot(layer_ex, cex.axis=0.7, cex.main=0.8,legend.width=1, legend.shrink=0.5)
```

As you can notice (Fig. 1), particularly for seafloor layers, maps are incomplete and contain an important number of missing values (N/A), because data were not interpolated in space. You can interpolate your data if you want using functions provided in the raster package, but you need to be aware that it will complexify interpretation afterwards.

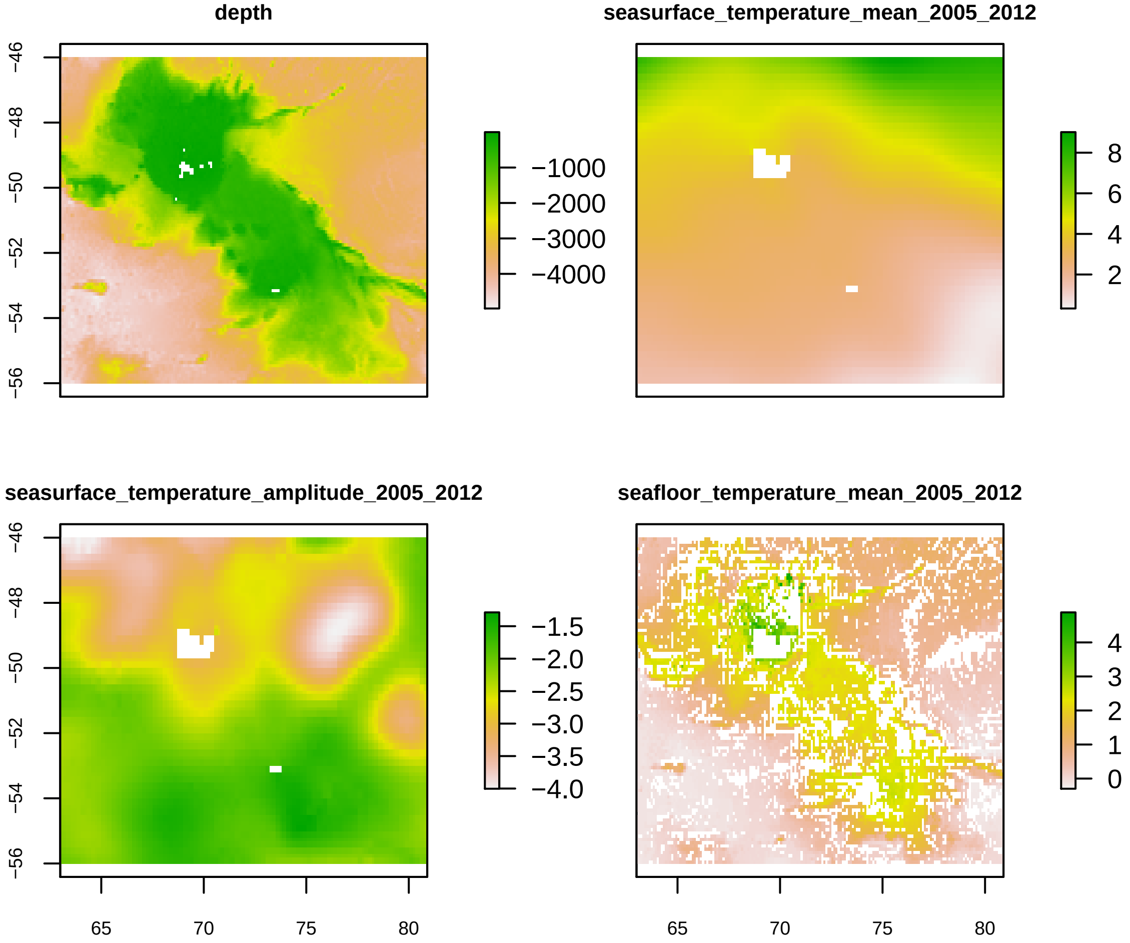
depth
seasurface_temperature_mean_2005_2012
seasurface_temperature_amplitude_2005_2012
seafloor_temperature_mean_2005_2012
−46
−48
−50
−52
−54
−56
−1000
−2000
−3000
−4000
8
6
4
2
−1.5
−2.0
−2.5
−3.0
−3.5
−4.0
4
3
2
1
0
65
70
75
80

Figure 1: Some environmental descriptors on the Kerguelen Plateau area

At the scale of the Southern Ocean, some layers are available on the AAD website (click to access to the link).

Download the following dataset, select, open and stack some layers according to your study species:

Three of these layers are provided in the internal data of the SDMPlay package, to help you immediately generate simple examples.

```r
data("depth_SO")
data("ice_cover_mean_SO")
data("seafloor_temp_2005_2012_mean_SO")

predictors_stack_SO <- stack(depth_SO,ice_cover_mean_SO,seafloor_temp_2005_2012_mean_SO)
names(predictors_stack_SO)<-c("depth","ice_cover_mean","seafloor_temp_mean")
```

If you have downloaded the environmental descriptors on your computer and stored them in a folder (called for example "environmental_layers"), you can do:

```r
library(ncdf4)
depth <- raster("environmental_layers/depth.nc")
seafloor_temp_mean <- raster("environmental_layers/seafloor_temp_2005_2012_mean.nc")

predictors_stack <- stack(depth,seafloor_temp_mean)
```

```r
# Plot the raster layers, create nice color palettes (Fig. 2)
library(RColorBrewer)
my.palette.oranges <- brewer.pal(n = 9, name = "Oranges")
my.palette.blue <- rev(brewer.pal(n = 9, name = "Blues"))

plot(raster::subset(predictors_stack_SO,1), col=my.palette.blue)
points(worldmap, type="l")
```

We also advice you to have a look at the SOmap (click to link) R package for nice plotting of Southern Ocean maps

```r
remotes::install_github("AustralianAntarcticDivision/SOmap")
library(SOmap)
SOmap()
# see the tutorial at the above website to customize your plot
```

Prepare model inputs

The first step after loading and checking your data is to adapt your dataset for modelling. Model algorithms require a table containing the environmental values associated with occurrence data.

Usually, SDMs are calibrated with presence and absence records. If you don't have access to absence records (or you don't rely in them!), it is necessary to sample in the projected area a set of points to define the background environmental conditions (termed background or pseudo-absence records)(Pearce and Boyce 2006). In this case, the SDM will be calibrated with presence AND background data instead of presences/absences.

Several background sampling methods exist (Phillips et al. 2009), and its choice depends on the sampling pattern of the presence-only records and on the scientific questions. In this first vignette, we will give

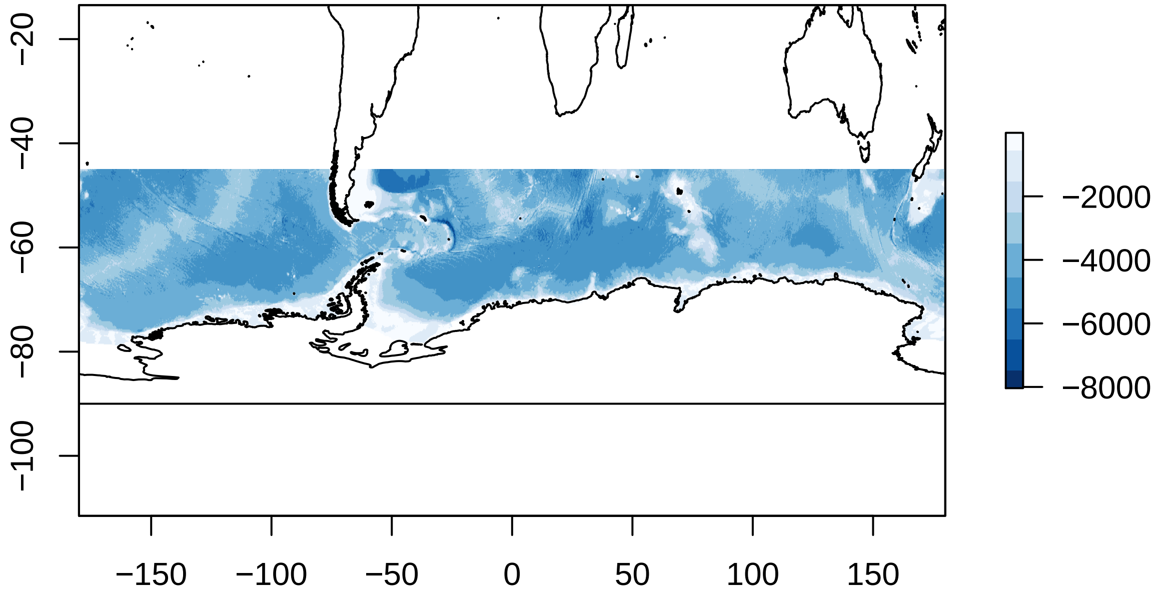

Figure 2: Bathymetry layer at the Southern Ocean extent

the example of a random sampling of the background records but see Tutorial #3/ “Importance of model calibration” for further implementations.

The first step before running a SDM is to create a table such as presented below. The `SDMtab` function can automatize its creation.

ID*	Longitude	Latitude	Depth	...	Temperature
1	63.33	-48.26	-480	...	1.4
1	64.13	-48.57	-104	...	1.2
...	...	...	...	...	...
0	67.32	-47.23	-1013	...	2.5
0	67.90	-55.45	-98	...	4.3

*ID corresponds to presence data (ID=1), or background data (ID=0) (or absences if you have absences !).

To create this table, apply the following steps:

Extract longitude and latitude values

You need to work with longitude and latitude data only -> clean your initial dataframe.

```
ctenocidaris.nutrix.occ <- ctenocidaris.nutrix[,c(7,8)] # longitude (first column),
                                                        # latitude (second column)
head(ctenocidaris.nutrix.occ)
```

```
##    decimal.Longitude decimal.Latitude
## 56          67.13167        -48.98500
## 57          67.33167        -49.44167
## 58          67.51167        -49.00500
## 59          67.54167        -48.11667
## 60          67.88500        -49.46667
## 61          68.05833        -49.06667
```

Create your SDMtab dataframe

```
SDMtable_ctenocidaris <- SDMPlay:::SDMtab(xydata=ctenocidaris.nutrix.occ,
       predictors=predictors2005_2012,
       unique.data=FALSE,
       same=TRUE)
```

When `unique.data= TRUE`, presence-only duplicates located on a same grid-cell pixel will be removed from the `xydata` variable.

`same` and `background.nb` functions refer to the sampling of background data: `background.nb`, indicates the specific number of background data to be sampled, while `same` is a shortcut that makes the number of background data similar to the number of presence-only data available. You can refer to Barbet-Massin et al. (2012) to choose the most appropriate number of background data to sample for your case study.

We can display the beginning and the end of the first columns of this new `SDMtab` object:

```
head(SDMtable_ctenocidaris[,c(1:5)])
```

```
##   id longitude latitude depth seasurface_temperature_mean_2005_2012
## 1  1     67.15   -48.95  -653                               4.24109
## 2  1     67.35   -49.45  -204                               3.89770
## 3  1     67.55   -49.05  -168                               4.06841
## 4  1     67.55   -48.15  -355                               4.65109
## 5  1     67.85   -49.45  -136                               3.86259
## 6  1     68.05   -49.05  -155                               4.03309
```

```
tail(SDMtable_ctenocidaris[,c(1:5)])
```

```
##     id longitude latitude depth seasurface_temperature_mean_2005_2012
## 245  0     71.25   -49.35  -380                              3.207595
## 246  0     66.65   -48.15  -599                              4.565590
## 247  0     70.15   -53.85 -3487                              2.440090
## 248  0     69.65   -47.95  -135                              4.675900
## 249  0     80.25   -52.65 -3428                              1.077700
## 250  0     65.25   -50.95 -3385                              3.428560
```

The dataframe combines environmental values of the 125 presence-only data available (ID=1) and environmental values associated with 125 background data randomly sampled in the area (ID=0).

You can display the sampled data on a map (Fig.3):

```
# nice colors
bluepalette<-colorRampPalette(c("blue4","blue","dodgerblue", "deepskyblue",
                                "lightskyblue"))(800)

# map
data("worldmap")
# Isolate depth layer from the environmental stack (Kerguelen data)
# Use it as a nice background for your figure
depth <- subset(predictors2005_2012,1)

# Extract background coordinates from SDMtable
background.occ <- subset(SDMtable_ctenocidaris,SDMtable_ctenocidaris$id==0)[,c(2,3)]

# plot the result (Fig.3)
plot(depth, col=bluepalette, cex=0.8,legend.width=0.5, legend.shrink=0.4,
     legend.args=list(text='Depth (m)', side=3, font=2, cex=0.8))
points(worldmap, type="l")
points(ctenocidaris.nutrix.occ, pch= 20, col="black")
points(background.occ, pch= 20, col="red")
legend("bottomleft", pch=20, col=c("black", "red"), legend=c("presence-only data","background data"),
       cex=0.6, bg="white")
```

You can assess the quality of your dataset with the `SDMdata.quality` function. This function estimates the percentage of presence-only records that fall on grid-cell pixels containing non-informative values (N/A). It estimates the quality of your dataset.

```
head(SDMdata.quality(SDMtable_ctenocidaris))
```

```
##                                             NA.percent (%)
```

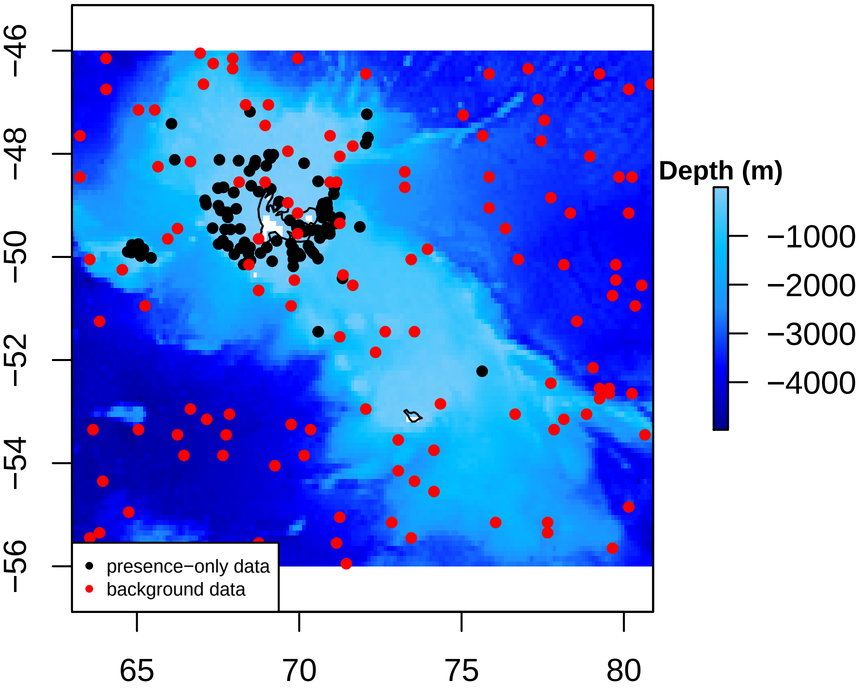

Figure 3: Presence data of Ctenocidaris nutrix and sampled background records on the Kerguelen Plateau

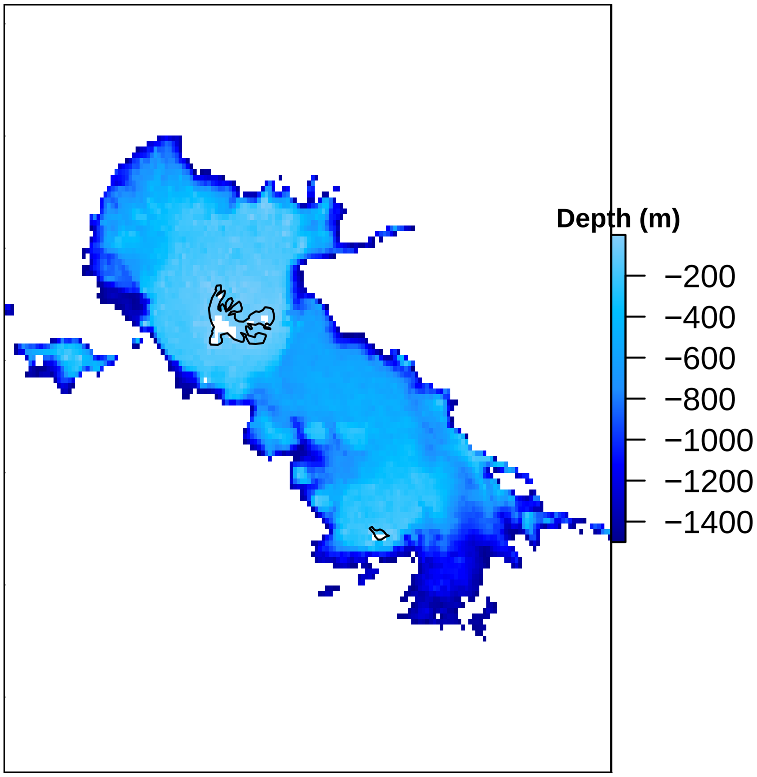

Figure 4: Restraining the area to 1,500m depth

```
## depth                                                  1.2
## seasurface_temperature_mean_2005_2012                  5.6
## seasurface_temperature_amplitude_2005_2012             5.6
## seafloor_temperature_mean_2005_2012                   32.0
## seafloor_temperature_amplitude_2005_2012              32.0
## seasurface_salinity_mean_2005_2012                     5.6
```

A last calibration step that you can perform before modelling is delineating the modelled area (Fig.4). The delim.area function can be used to restrict in geography and/or depth the environmental descriptor layers. This step can play an important role to enhance modelling performances by limiting the extent of extrapolation.

```
# restrict to 1500m depth
predictors2005_2012_1500m <- SDMPlay:::delim.area(predictors2005_2012, longmin=62, longmax=80,
                                                  latmin=-55, latmax=-45, interval=c(0,-1500))
# plot the new layer (Fig.4)
plot(subset(predictors2005_2012_1500m,1), col=bluepalette,legend.width=0.5, legend.shrink=0.4,
     legend.args=list(text='Depth (m)', side=3, font=2, cex=0.8))
points(worldmap, type="l")
```

You can focus your background sampling on this restrained environment (Fig.5). Run again the SDMtab code with these changes. The function will omit the N/A pixels when selecting the random background data.

```
SDMtable_ctenocidaris_1500 <- SDMtab(xydata=ctenocidaris.nutrix.occ,
       predictors=predictors2005_2012_1500m,
```

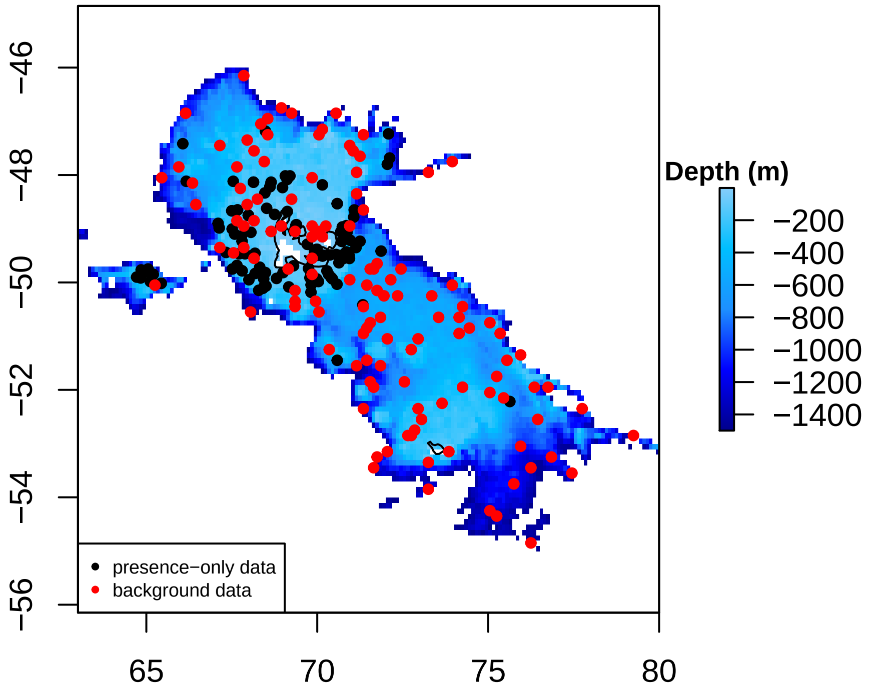

Figure 5: Restraining the area and the background sampling to 1,500m depth

```
          unique.data=FALSE,
          same=TRUE)
```

```
# Observe the changes (Fig.5)
background.occ_1500 <- subset(SDMtable_ctenocidaris_1500,SDMtable_ctenocidaris_1500$id==0)[,c(2,3)]
plot(subset(predictors2005_2012_1500m,1), col=bluepalette, cex=0.8, legend.width=0.5,
     legend.shrink=0.4,
     legend.args=list(text='Depth (m)', side=3, font=2, cex=0.8))
points(worldmap, type="l")
points(ctenocidaris.nutrix.occ, pch= 20, col="black")
points(background.occ_1500, pch= 20, col="red")
legend("bottomleft", pch=20, col=c("black", "red"), legend=c("presence-only data",
                                                               "background data"), cex=0.6)
```

Perform species distribution models

Once you have built your `SDMtab` dataframe, you can easily perform models using the `compute.brt` or `compute.maxent` functions.

```
compute.brt(x, proj.predictors, tc = 2, lr = 0.001, bf = 0.75,
           n.trees = 50, step.size = n.trees)
compute.maxent(x, proj.predictors)
```

The fonctions require two main parameters, `x` which correspond to the `SDMtab` object previously created and `proj.predictors` being the `RasterStack` containing the environmental descriptors on which you want to project your model. The other arguments aim at calibrating the model. You can refere to Elith et al. (2008) and Elith et al. (2011) to choose the parameters according to your dataset. BRT arguments are explained in the gbm package.

Example for BRT

Predict species distribution on the Kerguelen Plateau, for [2005-2012]

```
Cteno_model_2005_2012 <- SDMPlay:::compute.brt(x=SDMtable_ctenocidaris_1500,
                                               proj.predictors=predictors2005_2012_1500m,
                                               tc = 2, lr = 0.001, bf = 0.75, n.trees = 500)
```

While the function is uploading, you can observe that the `gbm` function, called by `SDMPlay`, calculates the regression trees until reaching the best estimation of the predicted deviance. See Tutorial #3/"Importance of model calibration" for more information about the choice and the influence of these parameters on model predictions. See also Elith et al. (2008) for details on these parameters.

Afterwards, different outputs can be produced (see Tutorial #2/ SDM outputs for detailed applications). Here we will just provide the example of the distribution map:

```
# display nice colors
palettecolor <- colorRampPalette(c("deepskyblue", "darkseagreen","lightgreen",
                                   "green","yellow","gold","orange", "red","firebrick"))(100)
# plot the results (Fig.6)
plot(Cteno_model_2005_2012$raster.prediction,col=palettecolor, main="Projection for [2005-2012]",
     cex.axis= 0.7,
     legend.width=0.5, legend.shrink=0.25,
     legend.args=list(text='Distribution probability', side=3, font=2, cex=0.8))
points(worldmap, type="l")
```

The output of your model cannot extrapolate on the grid-cell pixels from which it does not know environmental values (N/A pixels). Choose the option of interpolating your `RasterStack` layers before modelling or when projecting if you want to obtain smoother prediction maps. The map gives you the species distribution probabilities contained between 0 and 1.

Project on other time periods

If you want to project your model on another time period and infere your species distribution for other environmental conditions, you just need to change the `proj.predictors` in `compute.brt`, and replace it by a stack of future layers. The fonction will do the relationship between the environmental descriptors used for creating the model (the stack of predictors of present conditions that you have used to create the SDMtab matrtix) and the one for projecting (your future stack in this case). Be careful ! You must ensure that the extent, number, order and names of your future raster layers (stacked) are similar than the ones you have for the present time period.

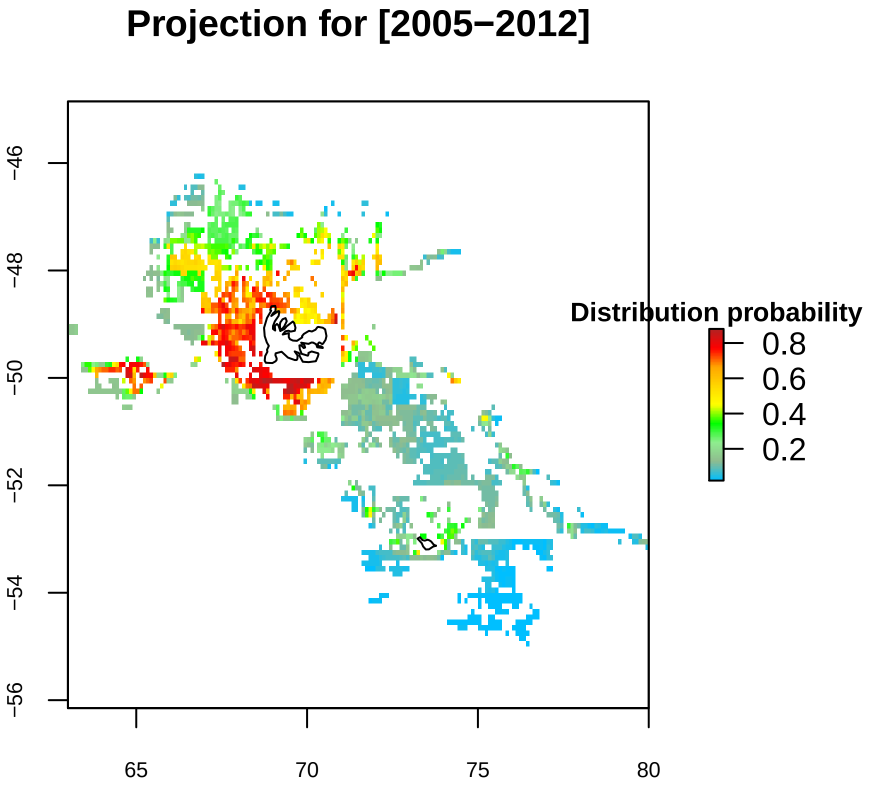

Figure 6: Projection for [2005-2012], Ctenocidaris nutrix predicted distribution, BRT

Example for MaxEnt

The procedure for MaxEnt algorithm is similar to BRT. `compute.maxent` uses the functionalities of the *dismo* `maxent` function. This function calls MaxEnt species distribution software, which is a java program that can be downloaded here (click to access to the link). In order to run `compute.maxent`, put the `maxent.jar` file downloaded at this address in the *java* folder of the *dismo* package (path obtained with the `system.file('java', package='dismo')` command). For issues with Java installation, consult dismo and rJava packages.

MaxEnt model outputs are similar to BRT, you can compute maps, response plots, environmental descriptor contributions. Refere to the example section of the function for more details.

Go further

SDMPlay provides extra fonctions to go further in your modelling work. You can perform null models with `null.model`, evaluate modelling performance and define probability threshold with `SDMeval`, calculate extrapolation and test different cross-validation procedures. See the following tutorials and the examples provided within the different functions for further details.

References

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2), 327-338.

David, B., Choné, T., Mooi, R. & de Ridder C. (2005). *Antarctic echinoidea* (Vol. 10). ARG Gantner.

Elith, J., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J Hijmans, R., Huettmann, F., … & A Loiselle, B. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), 129-151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802-813.

Elith, J. & Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677-697.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), 43-57.

Guillaumot, C., Martin, A., Fabri-Ruiz, S., Eléaume, M. & Saucède, T. (2016). Echinoids of the Kerguelen Plateau–occurrence data and environmental setting for past, present, and future species distribution modelling. *ZooKeys*, (630), 1.

Guillaumot, C., Artois, J., Saucède, T., Demoustier, L., Moreau, C., Eléaume, M. … & Danis, B. (2019). Broad-scale species distribution models applied to data-poor areas. *Progress in Oceanography*, 175, 198-207.

Guillaumot, C., Moreau, C., Danis, B. & Saucède, T. (2020). Extrapolation in species distribution modelling. Application to Southern Ocean marine species. *Progress in Oceanography*, 188, 102438.

Mah, C.L. & Blake, D.B. (2012). Global diversity and phylogeny of the Asteroidea (Echinodermata). *PloS One*, 7(4), e35644.

Moreau, C., Mah, C., Agüera, A., Améziane, N., Barnes, D., Crokaert, G., … & Jażdżewska, A. (2018). Antarctic and sub-Antarctic Asteroidea database. *ZooKeys*, (747), 141.

Pearce, J.L. & Boyce, M.S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43(3), 405-412.

Pearson, R.G. (2007). Species' distribution modeling for conservation educators and practitioners. Synthesis. *American Museum of Natural History*, 50.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), 181-197.