Chapter 3

sdmpredictors: an R package for species distribution modelling predictor datasets

Samuel Bosch[1,2], Lennert Tyberghein[1], Olivier De Clerck[2]

[1]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[2]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*

Unpublished manuscript.

Abstract

sdmpredictors is an open source R package which allows the end user to download terrestrial and marine environmental layers for the past, current and future climates. *sdmpredictors* contains metadata, statistics and pairwise correlations for the available datasets and layers. These correlations between predictors can be subsequently grouped and plotted. Currently *sdmpredictors* contains geophysical, biotic and climate data from WorldClim, ENVIREM, Bio-ORACLE and MARSPEC at 5 arcmin resolution and in the Behrmann equal area projection with a resolution of 7 kilometres.

Introduction

Species distribution modelling (SDM) is a commonly used tool in ecology and conservation biology. Correlative species distribution models relate species occurrences and (pseudo-)absence data to environmental predictor variables, based on statistically derived response surfaces (Guisan & Thuiller, 2005). Coinciding with the persistent interest in SDM, in the last 20 years numerous R packages related to SDM have been released. These include packages for downloading, checking and thinning occurrences (Aiello-Lammens et al., 2015; Chamberlain et al., 2016a; Provoost et al., 2016; Robertson et al., 2016), downloading the WorldClim environmental dataset (Hijmans et al., 2016; August et al., 2017), fitting models with various algorithms (Liaw & Wiener, 2002; Royle et al., 2012; Golding & Purse, 2016) and packages providing a fully integrated framework for SDM (Thuiller et al., 2009; Hijmans et al., 2016; Naimi & Araújo, 2016; August et al., 2017). With *sdmpredictors* we aim to complement these R packages by providing an easy to use interface for the acquisition of uniform and compatible terrestrial and marine predictors from different datasets for the past, current and future climate layers. It allows the end user to easily discover and use the different available layers from different predictor datasets.

Package description

sdmpredictors allows you to query the metadata for datasets ('list_datasets') and the environmental layers ('list_layers'). After selecting the required current climate layers they can be downloaded and loaded into the R session using the 'load_layers' function by providing the layer codes. Once layers are loaded into R they can be passed to all functions expecting a RasterStack with environmental data such as 'extract' from *raster* (Hijmans, 2016), 'BIOMOD_FormatingData' from *biomod2 (Thuiller et al., 2009)*, *LocalRaster* module in ZOON (August et al., 2017) and many more.

In order to load paleoclimatic and future climate layers a set of functions links current climate layers to past and future climate layers ('get_paleo_layers' or 'get_future_layers') or list out the available paleoclimatic and future climate layers ('list_layers_paleo' or 'list_layers_future'). After which the same 'load_layers' function can be used to actually download the data.

With the 'layer_stats' function various summarizing layer statistics like minimum, first quantile, median, third quantile, maximum, median absolute deviation, mean and standard deviation can be queried. The 'layers_correlation' function allows one

to query the Pearson correlation coefficient between two or more layers. Following the suggestion of Dormann et al. (2013), to avoid including heavily correlated predictors in one SDM, we provide the 'correlation_groups' function, which groups predictors based on their correlation. Correlations for cropped versions of the predictors or between externally sourced predictors can be calculated with the 'pearson_correlation_matrix' function.

Finally, citations for the used datasets and layers can be obtained with the 'dataset_citations' and 'layer_citations' functions, respectively.

Integrated datasets

Currently data layers are available both as Behrmann equal area projected rasters with a 7 km resolution and as 5 arcminutes unprojected rasters. For the terrestrial environment we added the WorldClim (Hijmans et al., 2005) and ENVIREM (Title & Bemmels, 2017) datasets and for the marine environment we included Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013). An overview of these datasets can be found in Table 1. The included datasets all represent multiyear aggregated data from interpolated *in situ* data and satellite observations of the Earth's surface. For all of datasets past and future climate data were added when available. This is by no means a fixed list and we encourage end users to suggest new datasets for inclusion in *sdmpredictors*.

Usage

In Supporting information we provide an example use case where *sdmpredictors* is used to provide environmental data for modelling the distribution of *Dictyota diemensis* Sonder ex Kützing, one of the species from the MarineSPEED benchmark dataset (Chapter 4). Additionally, the data provided by *sdmpredictors* can also be used for numerous other applications, including the generation of virtual species (Duan et al., 2015; Leroy et al., 2016), measuring niche overlap (Broennimann et al., 2012), linking the environment with species richness and biogeographic structure (Tittensor et al., 2010; Belanger et al., 2012) and modelling species abundance and population dynamics (Pearce & Boyce, 2006; Pratheepa et al., 2016). In summary, this package provides users with a set of functions for obtaining and using environmental predictor datasets for the past, current and future climate within R.

Table 1. Overview of the datasets included in *sdmpredictors*. For an up to date list use the function 'list_datasets'.

Dataset	Description
WorldClim	WorldClim is a set of global terrestrial climate layers. It has average monthly climate data for minimum, mean, and maximum temperature and for precipitation for 1960-1990. Additionally it contains a set of bioclimatic variables that are derived from the monthly temperature and rainfall values. They represent annual trends, seasonality and extreme or limiting environmental factors.
ENVIREM	The ENVIREM dataset is a set of 16 climatic and 2 topographic variables that can be used in modelling species' distributions. The strengths of this dataset include their close ties to ecological processes, and their availability at a global scale, at several spatial resolutions, and for several time periods. The underlying temperature and precipitation data that went into their construction comes from the WorldClim dataset (www.worldclim.org), and the solar radiation data comes from the Consortium for Spatial Information (www.cgiar-csi.org). The data are compatible with and expand the set of variables from WorldClim v1.4 (www.worldclim.org).
Bio-ORACLE	Bio-ORACLE is a set of GIS rasters providing marine environmental information for global-scale applications. It offers an array of geophysical, biotic and climate surface data derived from satellite data or interpolated from in situ data.
MARSPEC	MARSPEC is a set of high resolution climatic and geophysical GIS data layers for the world ocean. Seven geophysical variables were derived from the SRTM30_PLUS high resolution bathymetry dataset. These layers characterize the horizontal orientation (aspect), slope, and curvature of the seafloor and the distance from shore. Ten "bioclimatic" variables were derived from NOAA's World Ocean Atlas and NASA's MODIS satellite imagery and characterize the inter-annual means, extremes, and variances in sea surface temperature and salinity.

To cite sdmpredictors or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Bosch S., Tyberghein, L. and De Clerck, O. 2017. sdmpredictors: an R package for species distribution modelling predictor datasets. Version 0.

Acknowledgements

The research was carried out with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI.

Data accessibility

The *sdmpredictors* R package is available on CRAN and at https://github.com/lifewatch/sdmpredictors.

Supporting information

Here we detail a sample application where *sdmpredictors* is used to provide environmental data for modelling the distribution of *Dictyota diemensis* Sonder ex Kützing, one of the species from the MarineSPEED benchmark dataset (Chapter 4). The distribution of *D. diemensis* is restricted to Australia and New Zealand, but as only Australian distribution records are available we restricted ourselves for this use case to the Australian range (Womersley, 1987; Adams, 1994). We first start with exploring the available datasets and layers. Followed by the download of a set of 5 marine layers (salinity, sea surface temperature mean and range, bathymetry and shore distance) from Bio-ORACLE and MARSPEC. These are subsequently clipped with the shape of the Australian Exclusive Economic Zone using the *raster* and *mregions* packages (Chamberlain et al., 2016b; Hijmans, 2016). Secondly statistics and correlations for both the global and Australian data are inspected and visualized. For the correlation plot we additionally used the *ggplot2* and *cowplot* packages (Wickham et al., 2016; Wilke & Wickham, 2016). Since no predictors are grouped in a correlation group (Pearson correlation > 0.7) we used all predictors for building the SDM. We downloaded occurrences using *marinespeed* (Bosch et al., 2017), which are then used to create an SDM using ZOON (August et al., 2017). Finally the citations for the used layers are printed. For this application we used the Behrmann equal-area projected layers which required the projection of extents and occurrence points, avoiding oversampling of higher latitudes (Elith et al., 2011).

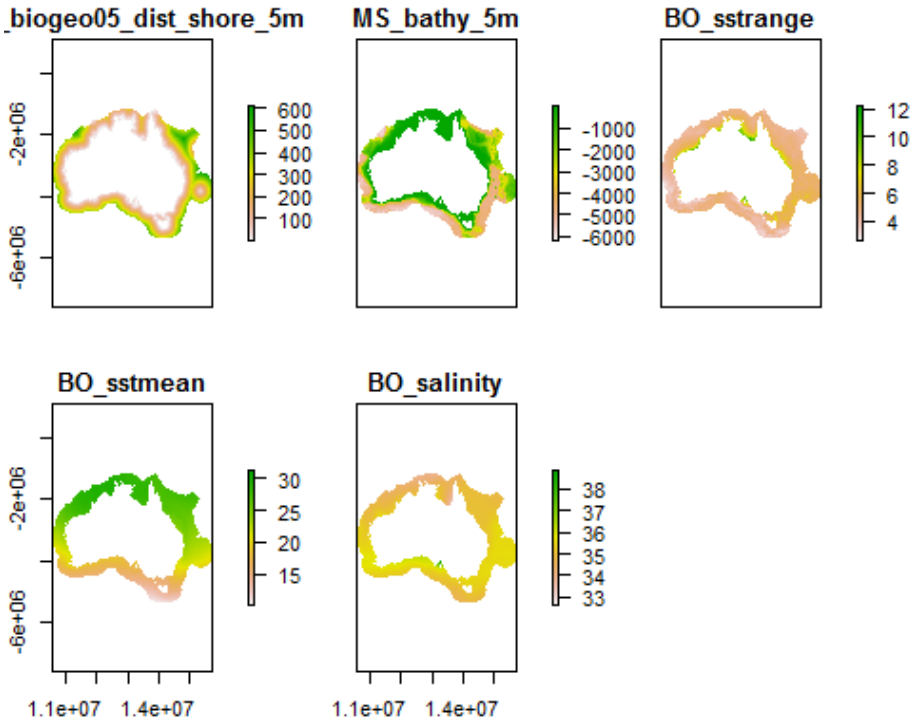
```
library(sdmpredictors)
library(mregions)
library(zoon)
# Inspect the available marine datasets and layers
datasets <- list_datasets(terrestrial = FALSE, marine = TRUE)
View(datasets[,c("dataset_code", "description")])
```

dataset_code	description
Bio-ORACLE	Bio-ORACLE is a set of GIS rasters providing marine environmental information for global-scale applications. It offers an array of geophysical, biotic and climate data at a spatial resolution 5 arcmin (9.2 km) in the ESRI ascii format.
MARSPEC	MARSPEC is a set of high resolution climatic and geophysical GIS data layers for the world ocean. Seven geophysical variables were derived from the SRTM30_PLUS high resolution bathymetry dataset. These layers characterize the horizontal orientation (aspect), slope, and curvature of the seafloor and the distance from shore. Ten "bioclimatic" variables were derived from NOAA's World Ocean Atlas and NASA's MODIS satellite imagery and characterize the inter-annual means, extremes, and variances in sea surface temperature and salinity. These variables will be useful to those interested in the spatial ecology of marine shallow-water and surface-associated pelagic organisms across the globe. Note that, in contrary to the original MARSPEC, all layers have unscaled values.

```
layers <- list_layers(datasets)
View(layers[1:2,c("dataset_code", "name", "description",
                  "primary_type")])
```

dataset_code	name	description	primary_type
Bio-ORACLE	Calcite (mean)	Calcite concentration indicates the mean concentration of calcite (CaCO3) in oceans.	Satellite (Aqua-MODIS), seasonal climatologies
Bio-ORACLE	Chlorophyll A (maximum)	Chlorophyll A concentration indicates the concentration of photosynthetic pigment chlorophyll A (the most common "green" chlorophyll) in oceans. Please note that in shallow water these values may reflect any kind of autotrophic biomass.	Satellite (Aqua-MODIS), monthly climatologies

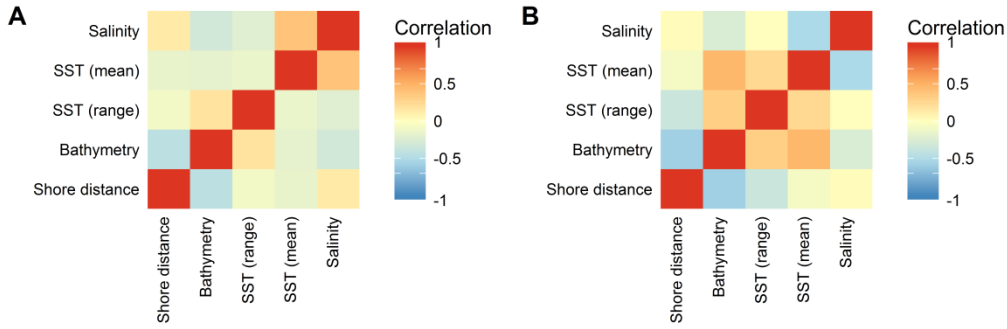
```
# Load equal area rasters, crop with the shape of the Australia EEZ
layercodes <- c("MS_biogeo05_dist_shore_5m", "MS_bathy_5m",
                "BO_sstrange", "BO_sstmean", "BO_salinity")
env <- load_layers(layercodes, equalarea = TRUE)
eez <- mregions::mr_shp("MarineRegions:eez", maxFeatures = NULL,
              filter = "Australian Exclusive Economic Zone")
eez <- sp::spTransform(eez, equalareaproj)
australia <- raster::crop(env, extent(eez))
australia <- raster::mask(australia, eez)
plot(australia)
```



```
# Compare statistics between original and Australian bathymetry
rbind(layer_stats("MS_bathy_5m"),
      calculate_statistics("Bathymetry Australia",
                           raster(australia, layer = 2)))
```

	layer_code	minimum	q1	median	q3	maximum	mad	mean	sd	moran	geary
171	MS_bathy_5m	-10493	-4865	-4082	-2984	-1	1313.584	-3661.049	1644.869	0.9728919	0.0096978
0%	Bathymetry Australia	-6163	-4377	-1868	-85	-1	2682.023	-2222.555	1987.391	0.9736722	0.0053917

```
# Compare correlations between predictors, globally and for Australia
prettynames <- list(BO_salinity="Salinity",
                    BO_sstmean="SST (mean)",
                    BO_sstrange="SST (range)",
                    MS_bathy_5m="Bathymetry",
                    MS_biogeo05_dist_shore_5m = "Shore distance")
p1 <- plot_correlation(layers_correlation(layercodes), prettynames)
australian_correlations <- pearson_correlation_matrix(australia)
p2 <- plot_correlation(australian_correlations, prettynames)
cowplot::plot_grid(p1, p2, labels=c("A", "B"), ncol = 2, nrow = 1)
```



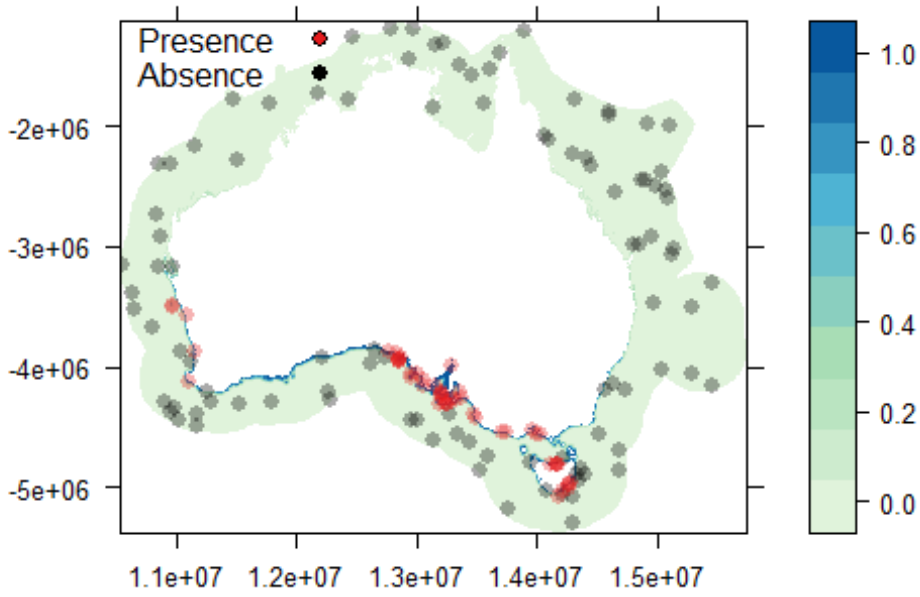

```
correlation_groups(australian_correlations, max_correlation = 0.7)

## [[1]]
##   MS_biogeo05_dist_shore_5m
## "MS_biogeo05_dist_shore_5m"
##
## [[2]]
##   MS_bathy_5m
## "MS_bathy_5m"
##
## [[3]]
```

```
##   BO_sstrange
## "BO_sstrange"
##
## [[4]]
##   BO_sstmean
## "BO_sstmean"
##
## [[5]]
##   BO_salinity
## "BO_salinity"
```

```
# Fetch occurrences and prepare for ZOON
occ <- marinespeed::get_occurrences("Dictyota diemensis")
points <- SpatialPoints(occ[,c("longitude", "latitude")],
                        lonlatproj)
points <- spTransform(points, equalareaproj)
occfile <- tempfile(fileext = ".csv")
write.csv(cbind(coordinates(points), value=1), occfile)
# Create SDM with ZOON
wf <- workflow(
  occurrence = LocalOccurrenceData(
    occfile, occurrenceType="presence",
    columns = c("longitude", "latitude", "value")),
  covariate = LocalRaster(stack(australia)),
  process = OneHundredBackground(seed = 42),
  model = LogisticRegression,
  output = PrintMap)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```



```r
# Layer citations
citations <- layer_citations(layercodes, astext=FALSE)
for(citation in citations) {
  print(citation, style="Bibtex")
}
```

```
## @Article{Bio-ORACLE,
##   author = {Lennert Tyberghein and Verbruggen Heroen and Klaas Pauly and
Charles Troupin and Frederic Mineur and Olivier {De Clerck}},
##   title = {Bio-ORACLE: a global environmental dataset for marine species
distribution modelling},
##   journal = {Global Ecology and Biogeography},
##   year = {2012},
##   volume = {21},
##   number = {2},
##   pages = {272-281},
##   doi = {10.1111/j.1466-8238.2011.00656.x},
## }
## @Article{MARSPEC,
##   author = {Elizabeth J. Sbrocco and Paul H. Barber},
##   title = {MARSPEC: ocean climate layers for marine spatial ecology},
##   year = {2013},
##   volume = {94},
##   number = {4},
##   pages = {979},
##   journal = {Ecology},
##   doi = {10.1890/12-1358.1},
## }
```