Species distribution modeling with three-step pseudo–absences

Maialen Iturbide

December 28, 2014

1 Introduction

This document provides an introduction to species distribution modeling (SDM) with three–step pseudo–absences.

Species distribution models (SDM) are statistical tools to predict the distribution of species in geographic space based on the relation of known species distribution to the environment. SDMs can be classified into *profile techniques* that only use distribution of presence data and *group discrimination techniques* that also require information of the environmental range where the species do not occur, that is, absence data. Due to the great effort involved in true absences sampling, most of the available datasets for predictive modeling are lacking in absence data (Zaniewski et al., 2002; Lobo et al., 2010), thereby some authors apply profile techniques such as ecological niche factor analysis (ENFA; i.e. Cianfrani et al., 2010; McKinney et al., 2012), Mahalanobis distance (MADIFA: i.e. Kuo, 2010; Martin et al., 2012) and environmental envelopes (BIOCLIM and DOMAIN: i.e. Giovanelli et al., 2010; Monk et al., 2010). However, given that group discrimination techniques generally perform better (Elith and et al, 2006; Engler et al., 2004; Chefaoui and Lobo, 2008), the most common methodological approach is to use group discrimination techniques relative to the available environment or background samples, also known as pseudo–absences, thus obtaining a representation of the environmental range in the region of study.

One of the most simple methods of generating pseudo–absences is to perform a random selection of the entire study area (Jiang et al., 2014; Carone et al., 2014; Sequeira et al., 2014). However, it rises the risk of introducing false absences into the model from locations that are suitable for the species. Faced with this problem, several authors employ a presence–only algorithm as a preliminary step to move pseudo–absences away in the environmental space (Zaniewski et al., 2002; Engler et al., 2004; Barbet-Massin et al., 2012; Liu et al., 2013).

The way of generating pseudo–absences strongly influences the results obtained (Lobo et al., 2010; Wisz and Guisan, 2009; Barbet-Massin et al., 2012; Hirzel et al., 2001), as well as the extent from which background is sampled, a constraint distribution of pseudo–absences around presence locations can lead to misleading models while the opposite, can inflate artificially test statistics and predictions, as well as potentially less informative response variables (VanDerWal and Shoo, 2009).

This document shows an example of a full Species distribution modeling process carried out with the `mopa` package in R. Pseudo–absences are generated in three-steps combining profiling techniques and background extent limitations.

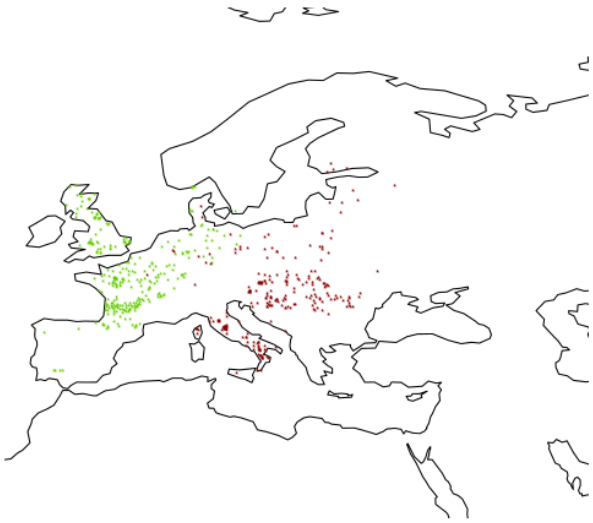
If you want to know more about SDM in R, you could consult, for example, documentation from package `dismo` made by Robert J. Hijmans and Jane Elith.

`NOTE:` Most functions presented are implemented in 'mopa'. In the few cases in which functions from other packages are used, the package name is always explicitly indicated.

2 Species occurrence data

Regarding presence data, Hernandez et al. (2006) suggested that research in environmental niche modeling should focus in broad distribution subunits that are based on distinct genetic linages, in this connection Gonzalez et al. (2011) demostrated that omission error is reduced when biologically meaningful data is modeled. Thus, functions in the `mopa` package are prepared to handle more than one group of presences at the same time (could be a list of either distribution subunits of a single species or distribution of multiple species), anyway, functions also perform with a single group or species (data frame). In this example we use a data set (list) available with the `mopa` package containing presence records of two phylogenetic groups (H11 and H5) of *Quercus* sp in Europe. This is, R–object `Oak_phylo2`, a modified subset of the `Quercus sp Europe Petit 2002` database (Petit and et al, 2002), which is available in the `Georeferenced Database of Genetic Diversity` or $(GD)^2$. To aid in map representation, a dataset called `wrld` containing a World map is also included in the package.

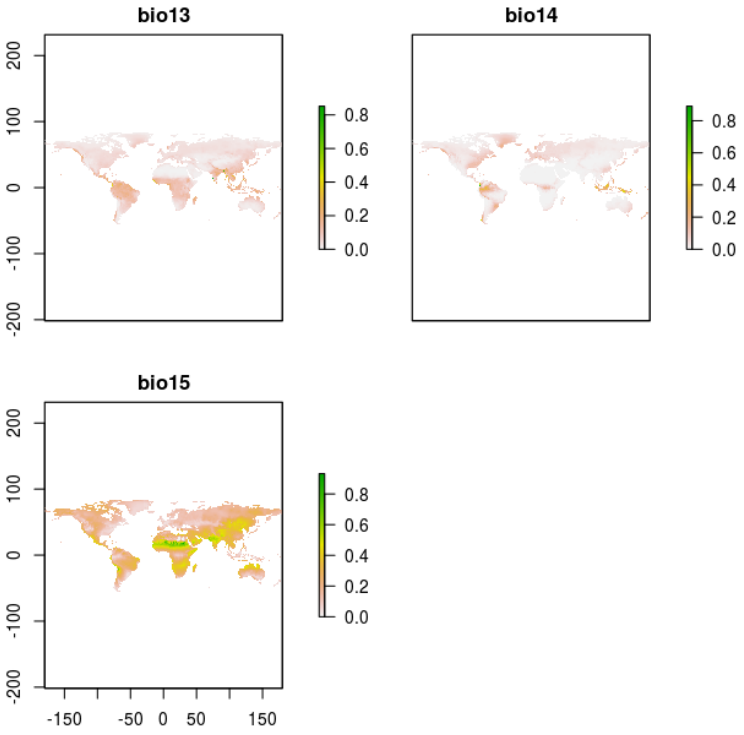
```
> library(mopa)
> data(wrld)
> data(Oak_phylo2)
> # Map
> plot(wrld, asp = 1, xlim= c(-10,50), ylim=c(40,60))
> for (i in 1:length(Oak_phylo2)) {
+   points(Oak_phylo2[[i]], pch = "*", cex = 0.5,
+          col = colors()[i*50])
+ }
```



3 Environmental variables

Predictor variables are typically organized as raster (grid) type files. The set of predictor variables (rasters) can be used to make a '`RasterStack`', which is a collection of '`RasterLayer`' objects (see `Raster-class` in the `raster` package for more info). `mopa` uses as input this type of raster objects and also provides the R–object `biostack`, which is a '`RasterStack`' of three bioclimatic variables.

```
> # RatserStack of envirinmental variables
> data(biostack)
> plot(biostack)
```



4 Creation of the background grid

The regular point grid which covers the continental area can be created with functions from the `raster` and `sp` packages as follows:

```
> library(raster)
> ac<-xyFromCell(biostack[[1]],  1:ncell(biostack[[1]]))
> ex<-extract(biostack[[1]], ac)
> sp_grid<-SpatialPoints(ac[-which(is.na(ex)),])
> projection(sp_grid)<-CRS("+proj=longlat +init=epsg:4326")
```

For easy of use, R–object `sp_grid` is available in `mopa`, covering the World at 10 km resolution.

5 Limit study area to the bounding boxes around presences

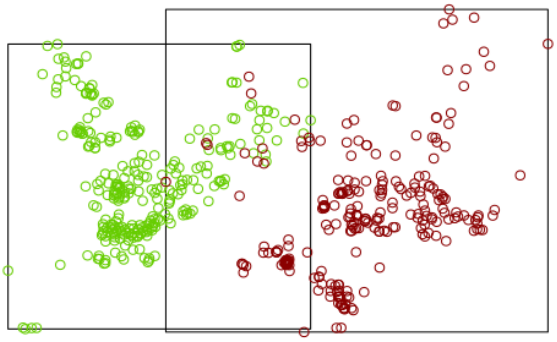
Function `boundingCoords` creates the matrix of bounding coordinates around point records (xy records). In this case, since `Oak_phylo2` object is a list of two groups of points, a list of two matrixes is created.

```
>  oak.extension <- boundingCoords(xy = Oak_phylo2)
```

Function `delimit` creates polygon shapes from bounding coordinates and limits `SpatialPoints` data (`sp_grid`) to the defined boundaries, in other words,

does the intersection of the background point grid with the bounding boxes. A list with two objects is obtained, (1)bbs: polygon shape of the bounding boxes and (2)bbs.grid: list of data frames of the background point grid limited by the bounding boxes.

```
> box.grid <- delimit(bounding.coords = oak.extension,
+                grid = sp_grid, names = names(Oak_phylo2))
> # Plot presences and bounding boxes
> plot(box.grid$bbs, asp = 1)
> for (i in 1:length(Oak_phylo2)){
+   points(Oak_phylo2[[i]], col = colors()[i*50])
+ }
```



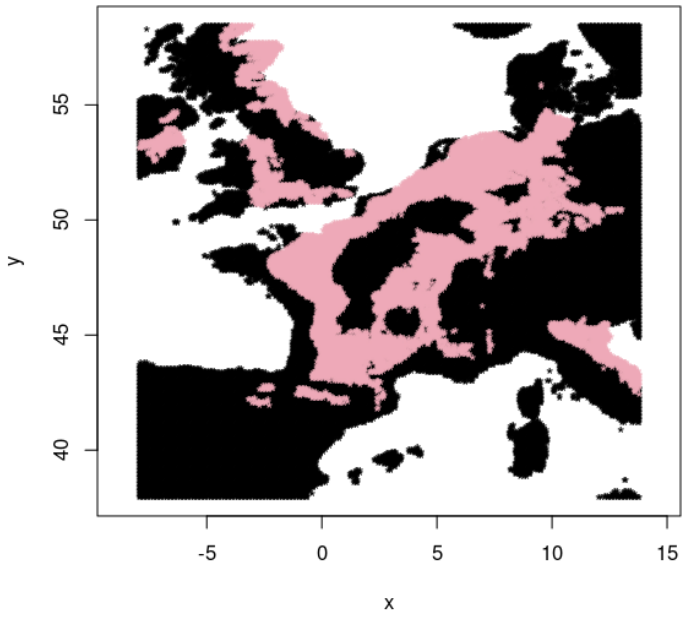
6 Three–step pseudo–absences generation

In this section we illustrate the steps to generate pseudo-absences in three-steps combining profiling techniques and background extent limitations.

6.1 STEP1: environmental profiling

The first step is the selection of the environmentally unsuitable areas with a presence only algorithm. function `OCSVMprofiling` uses One-class support vector machines (OCSVM: Scholkopf and Smola, 2001) to perform a preliminary binary classification of the study region (suitable/unsuitable) using as input the environmental conditions of the presence localities.

```
> unsuitable.bg <-OCSVMprofiling(xy = Oak_phylo2,
+                      varstack = biostack,
+                      bbs.grid = box.grid$bbs.grid)
> # Plot areas predicted as suitable (presence) and
> # unsuitable (absence) for group H11
> plot(unsuitable.bg$absence$H11, pch = "*", asp = 1)
> points(unsuitable.bg$presence$H11, pch = "*", col = "pink2")
```



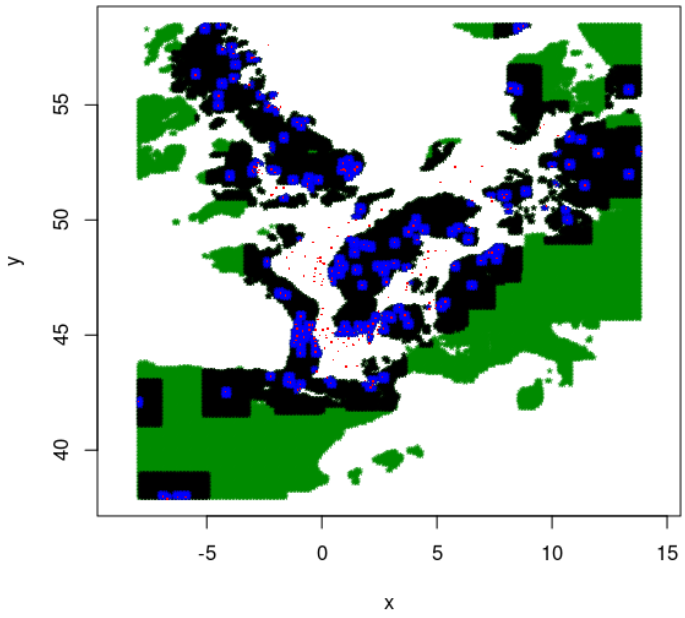
6.2 STEP2: SDM performing with pseudo–absences generated into different extents of the unsuitable background

In the second step, SDMs are performed with pseudo–absences generated into different extents of the unsuitable background. Several functions are involved in this step. Function `bgRadio` performs the partition of the background space considering multiple distance thresholds. In other words, it creates backgrounds of different spatial extent for each species/population. In the example below, extents are created for a sequence of 10 km between distances, from 20 km to half the length of the diagonal of the bounding box, as described in Sec. 2.4 of the manuscript. A list of matrices containing xy coordinates is returned, each matrix corresponding to a different background extent tested.

```
> ext <- bgRadio(xy = Oak_phylo2, bounding.coords = oak.extension,
+                bg.absence = unsuitable.bg$absence,
+                start = 0.166, by = 0.083,
+                unit = "decimal degrees")
```

```
[1] "creating background point-grids for species 1 out of 2"
[1] "creating background point-grids for species 2 out of 2"
```

```
> # Plot presences for group H11 and background extents of 20,
> # 120 and 520 km
> plot(ext$H11$km520, col = "green4", pch = "*", asp = 1)
> points(ext$H11$km120, pch = "*")
> points(ext$H11$km20, pch = "*", col = "blue")
> points(Oak_phylo2$H11, col = "red", pch = ".", cex = 1.5)
```



With function `PseudoAbsences`, you can create pseudo–absences either at random or with k–means clustering, by modifying argument `kmeans`. You can also set the prevalence (proportion of presences against pseudo–absences) and the exclusion buffer (minimum distance to be kept to presences without pseudo–absences).

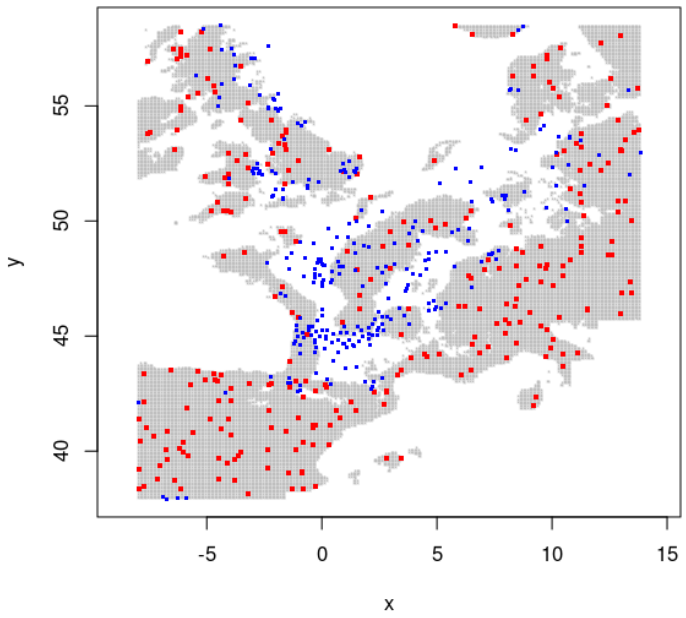
6.2.1 At random

In the example below, pseudo–absences are generated at random, in equal number to presences (prevalence) and keeping a 10 km distance to presences (exclusion buffer).

```
> pa_random <-PseudoAbsences(xy = Oak_phylo2, bg.grids = ext,
+                            exclusion.buffer = 0.083,
+                            prevalence = 0.5, kmeans = FALSE)
```

```
[1] "generating pseudo-absences for species 1 out of 2"
[1] "generating pseudo-absences for species 2 out of 2"
```

```
> # Plot presences/pseudo-absences for group H11 considering
> # the background extent of 520 km
> plot(ext$H11$km520, pch="*", col= "grey", cex=.5, asp=1)
> points(pa_random$H11$km520, col= "red", pch=".", cex=4)
> points(Oak_phylo2$H11, col= "blue", pch=".", cex=3)
```



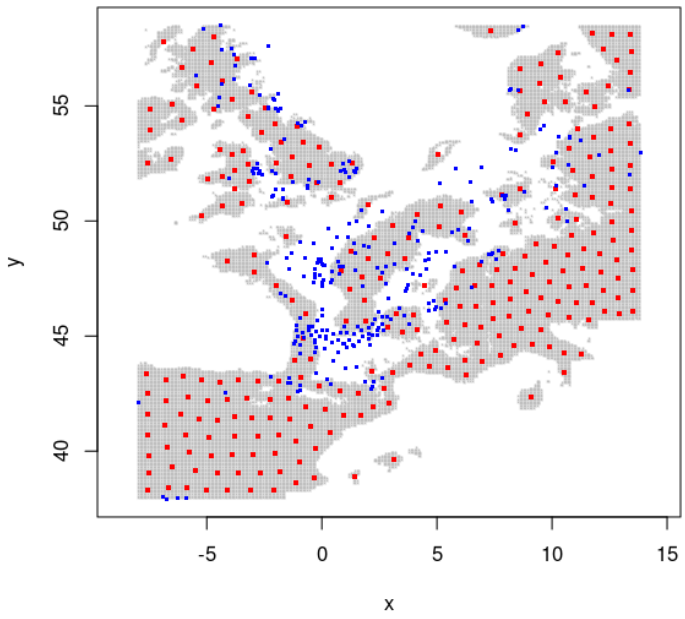
6.2.2 With k–means clustering

In the example below, pseudo-absences are generated with k-means clustering, in equal number to presences (prevalence) and keeping a 10 km distance to presences (exclusion buffer).

```
> pa_kmeans <-PseudoAbsences(xy = Oak_phylo2, bg.grids = ext,
+                            exclusion.buffer = 0.083,
+                            prevalence = 0.5, kmeans = TRUE,
+                            varstack = biostack)

[1] "generating pseudo-absences for species 1 out of 2"
[1] "generating pseudo-absences for species 2 out of 2"

> # Plot presences/pseudo-absences for group H11 considering
> # the background extent of 520 km
> plot(ext$H11$km520, pch = "*", col = "grey", cex = .5, asp = 1)
> points(pa_kmeans$H11$km520, col = "red", pch = ".", cex = 4)
> points(Oak_phylo2$H11, col = "blue", pch = ".", cex = 3)
```



Function `bindPresAbs` binds presence and absence data for each background extension.

```
> presaus <- bindPresAbs(presences = Oak_phylo2,
+                        absences = pa_random)
```

The `allModeling` function does the species distribution modelling and k-fold cross validation for a set of presence/absence data per species corresponding to a different background extent. Algorithms supported are "glm", "svm", "maxent", "mars", "randomForest", "cart.rpart" and "cart.tree".

In the example below, we do a 10–fold cross validation of the "mars" modelling algorithm.

```
> modirs <- allModeling(data = presaus, varstack = biostack,
+             k = 10, algorithm = "mars", destdir = getwd(),
+             projection = CRS("+proj=longlat +init=epsg:4326"))
```

Named Rdata objects are stored in the specified path. Each Object is given a name indicating the algorithm, background extent, and species in this order (if a single species is provided no name is given for de species). Character object with listed files is returned. Each Rdata consists of a list with six components:

(1)allmod: fitted model with all data for training, (2)auc: AUC statistic in the cross validation, (3)kappa: kappa statistic in the cross validation, (4)tss: true skill statistic in the cross validation, (5)mod: fitted model with partitioned data, (6)p: cross model prediction.

6.3 STEP3: selection of the optimum background extent and corresponding fitted model

In the third step, AUCs obtained and corresponding extents are fitted to a Michaelis-Menten model to extract the *Vm* coefficient (equation 1 in the manuscript). Then, the minimum extent at which the AUC surpasses the *Vm* value is selected as the threshold extent (see Figure 3 in the manuscript), being the corresponding fitted SDM the definitive to predict suitability probabilities in the study area.

We next indicate how to plot the results of the optimal spatial extent selection using the 'lattice' package. First we load the data generated with the `allModeling` function and extract the corresponding AUC values. Function `loadTestValues` loads and stores AUC data in a matrix:

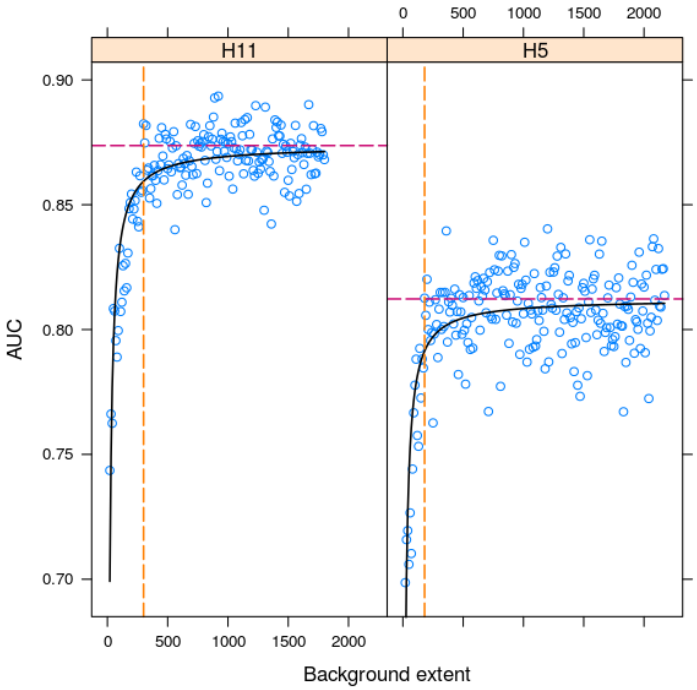
```
> auc_mars <-loadTestValues(data = presaus, test = "auc",
+                           algorithm = "mars")

[1] "loading values for species 1"
[1] "loading values for species 2"
```

Model fitting is done by function `indextent`, that internally uses the `nls` function of R package `stats`. An index of the threshold extents is obtained. A fitted model plot is also returned if argument `diagrams` is set to `TRUE` (the point crossed by the vertical line above the horizontal line corresponds to the smallest background extent at which the AUC overcomes the 'Vm' value (threshold extent)).

```
> ind <- indextent(testmat = auc_mars, diagrams = TRUE)
> ind

km300 km180
   29    17
```

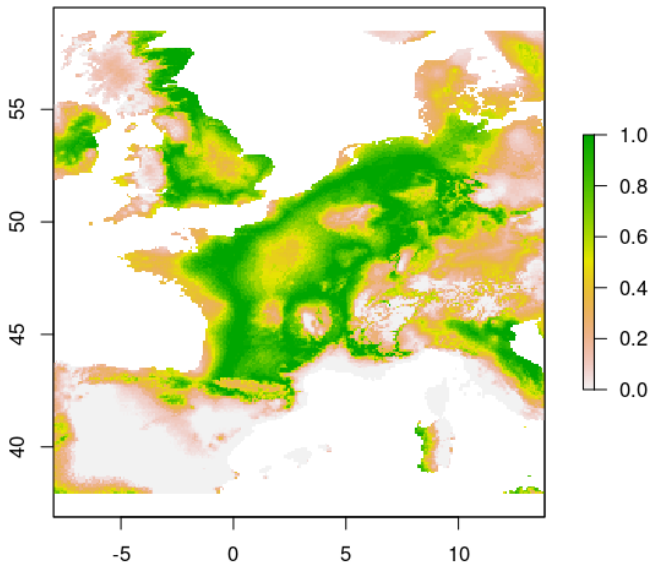


Thus, the `ind` object in this example gives the index of the background extents to be considered for each group/species and is used to extract definitive model components and data with function `loadDefinitiveModel` as follows.

```
> def <-loadDefinitiveModel(data = presaus, extents = ind,
+                         slot = "allmod", algorithm = "mars")
```

Once the optimal SDMs are chosen, we can generate the resulting suitability maps. In the example below we use function `biomat` for preparing a matrix with the variables for prediction in the study area. Then, the predictions are converted to a raster format with functions `SpatialPixelsDataFrame` (from the `sp` package) and `raster` (from the `raster` package).

```
> # Suitability map for the Oak group H11
> # Function 'biomat' prepares matrix with variables for
> # projection
> projectionland <- biomat(cbind(box.grid$bbs.grid$H11,
+                          rep(1, nrow(box.grid$bbs.grid$H11))),
+                          biostack)
> # Prediction
> p <- predict(def$H11, projectionland[ ,-1])
> p[which(p < 0)] <- 0
> p[which(p > 1)] <- 1
> # Convert prediction to a raster object
> spp <- SpatialPixelsDataFrame(box.grid$bbs.grid$H11,
+                                as.data.frame(p))
> ras <- raster(spp)
> plot(ras)
```



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.

Carone, M. T., Guisan, A., Cianfrani, C., Simoniello, T., Loy, A., Carranza, M. L., 2014. A multi-temporal approach to model endangered species distribution in europe. the case of the eurasian otter in italy. Ecological Modelling 274, 21–28.

Chefaoui, R. M., Lobo, J. M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling 210 (4), 478–486.

Cianfrani, C., Le Lay, G., Hirzel, A., Loy, A., 2010. Do habitat suitability models reliably predict the recovery areas of threatened species? Journal of Applied Ecology 47 (2), 421–430.

Elith, J., et al, 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41 (2), 263–274.

Giovanelli, J., de Siqueira, M., Haddad, C., Alexandrino, J., 2010. Modeling a spatially restricted distribution in the neotropics: How the size of calibration area affects the performance of five presence-only methods. Ecological Modelling 221 (2), 215–224, cited By (since 1996)28.

Gonzalez, S., Soto-Centeno, J., Reed, D., 2011. Population distribution models: Species distributions are better modeled using biologically relevant data partitions. BMC Ecology 11.

Hernandez, P. A., Graham, C. H., Master, L. L., Albert, D. L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29 (5), 773–785.

Hirzel, A. H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecological modelling 145 (2), 111–121.

Jiang, Y., Wang, T., De Bie, C., Skidmore, A., Liu, X., Song, S., Zhang, L., Wang, J., Shao, X., 2014. Satellite-derived vegetation indices contribute significantly to the prediction of epiphyllous liverworts. Ecological Indicators 38, 72–80.

Kuo, Y. L., 2010. Unexpected side-effects of winter feeding: Learning from mahalanobis distances factor analysis in the case of red-crowned cranes in hokkaido, japan. In: Modelling for Environment's Sake: Proceedings of the 5th Biennial Conference of the International Environmental Modelling and Software Society, iEMSs 2010. Vol. 1. pp. 112–116.

Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in victoria, australia. Ecological Modelling 249, 68–74.

Lobo, J. M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33 (1), 103–114.

Martin, J., Revilla, E., Quenette, P., Naves, J., Allainé, D., Swenson, J., 2012. Brown bear habitat suitability in the pyrenees: Transferability across sites and linking scales to make the most of scarce data. Journal of Applied Ecology 49 (3), 621–631.

McKinney, J., Hoffmayer, E., Wu, W., Fulford, R., Hendon, J., 2012. Feeding habitat of the whale shark rhincodon typus in the northern gulf of mexico determined using species distribution modelling. Marine Ecology Progress Series 458, 199–211.

Monk, J., Ierodiaconou, D., Versace, V., Bellgrove, A., Harvey, E., Rattray, A., Laurenson, L., Quinn, G., 2010. Habitat suitability for marine fishes using presence-only modelling and multibeam sonar. Marine Ecology Progress Series 420, 157–174.

Petit, R. J., et al, 2002. Chloroplast DNA variation in european white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. Forest Ecology and Management 156 (1âĂŞ3), 5–26.

Scholkopf, B., Smola, A. J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

Sequeira, A. M. M., Mellin, C., Fordham, D., Meekan, M., Bradshaw, C., 2014. Predicting current and future global distributions of whale sharks. Global Change Biology 20 (3), 778–789.

VanDerWal, J., Shoo, L. P., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling (4), 589–594.

Wisz, M. S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? an information-theoretic approach based on simulated data. BMC Ecology 9 (1), 8.

Zaniewski, A. E., Lehmann, A., Overton, J. M., 2002. Predicting species spatial distributions using presence-only data: a case study of native new zealand ferns. Ecological Modelling 157 (2), 261–280.