

BioScore: User-manual

Introduction

BioScore is a biodiversity assessment model which fits species distribution models (SDM) for each species. For each species we fit an SDM based on three modeling techniques (generalized linear model, GLM; generalized additive model, GAM; boosted regression trees, BRT). We fit the SDMs with the BIOMOD2 package using default settings (ver. 3.3–7; Thuiller et al. 2016).

BioScore Structure

The BioScore script has a core which contains all functions used in the model, a total of 16 core functions. It further can be ran for 2 ecosystems namely terrestrial and wetlands. For each of these modules, model settings, system setting and BioScore model initiation are specific to these ecosystems (Table 1). These scripts contain the information which can alternate between runs. Systems configuration is the script required for the system settings storing all the pathways and the system requirements. The model configuration script is the script which contains all the specifications required by the model. By running the script `InitiateBioScore.R` the model can be ran.

Table 1: Table with scripts and their description

SCRIPTS	DESCRIPTION
CORE	
00. Bioscore	Script which combines all the below mentioned core functions in order to run the BioScore model
01. Model Set up	Functions to set up the model with the given model settings and system settings
02. Create log file	Functions to initiate a log file to log all the steps and errors in the BioScore run
03. Load and format environment data	Functions to load and format the environmental data
04. Species selection	Functions to select species based on the preferred habitats
05. Observation selection	Functions to select presence and absence data
06. Format input data	Functions to format the data to meet the biomod function input
07. Fit SDMs	Functions to fit the SDMs with three different algorithms
08. Evaluate SDMs	Functions to evaluate the fitted SDMs

09. Variable importance	Functions to calculate the variable importance
10. Calculate response curves	Functions to calculate the response curves and the modelled niche optima from the response curves
11. Project SDMs	Functions to project the SDMS for the given scenarios
12. Save Stats	Functions to save the species specific statistics (one file per species)
13. SDM Function	Functions to run the fitting of the SDM function for all the species defined
14. Save All Metrics	Functions to aggregate all files with species specific statistics into one file
15. Run SDM Function	Functions to run the SDM function, in parallel or not
Ecosystem: Terrestrial/Wetlands	
Bioscore system settings	Parameters to save all the system settings defined per user, model run and ecosystem
Bioscore model settings	Parameters to save all the model settings per user, model run and ecosystem
Initiate BioScore	Script to initiate the full run of BioScore model

System requirements

BioScore runs in RStudio (version 2023.12.0+369). The BioScore script is usually run in parallel on a machine with 32 CPUs and 512 RAM using 15 CPUs to run the model. Alternatively a machine with 24 CPUs and 672 RAM can be used.

Model Overview and usage

To run Bioscore follow the steps below.

Step 1: Setting up file pathways

Create the following directories (Fig1) at your required drive. The structure and names of the directories need to be exactly as shown in the figure to avoid errors. These file paths should be created within the base_dir-directory (table 2).

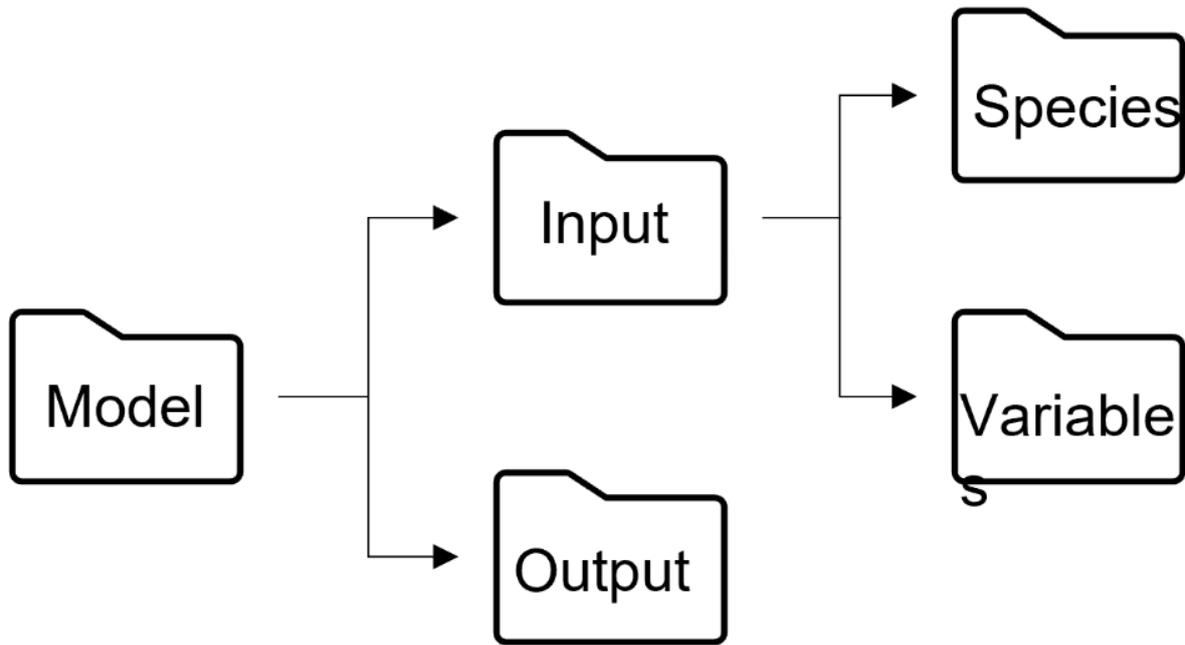


Figure 1: Directory names and structure for BioScore run

Step 2: Save the Input data to the file pathways

Input data has species input data (a file with species per plot and a file with all plot locations) retrieved from the European Vegetation Archive (EVA; Chytrý et al. 2016) with presence and absence data. It also contains a characteristic species list of distinct habitat types based on the EUNIS habitat type classification. Variables folder contains rasters of all the environmental variables which you want to use to fit the SDMs, e.g. climate, soil, land cover and pollution (nitrogen deposition) and other ecosystem specific variables. This also contains all the variables for the projection of various scenarios. The folder also contains OverviewVariables file which contains the names of the variables (e.g. Precipitation), the names of the respective files used to model the current (e.g. Precipitation_mean_1990-2010.tif) and, if relevant, the scenario distribution ranges (e.g. Precipitation_mean_2070-2100.tif). For administrative/archiving purposes only, further information can be added to the table, such as a reference to the source of the variables (Figure 2). The table should be such that it contains the variables names corresponding to the scenarios.

SharedNameVariable	current	source_current
Precipitation	mean_Prec_1990_2018.tif	"/mnt/data/hendriksm/species_composition/Marion/results/3_variables/CreateVariables_211014_Current"
TSUM	mean_TSUM_1990_2018.tif	"/mnt/data/hendriksm/species_composition/Marion/results/3_variables/CreateVariables_211014_Current"

Figure 2: Table format for OverviewVariable table

It is important to ensure that the resolution, projection and the extent of all the rasters is same and that the units of the environmental data in different scenarios is

identical. It is also important to ensure that projection of the rasters and the species data match. Before starting the model run, ensure these are similar by checking for units, resolution, projection and extent for species and environmental data.

Step 3: Set up the system and model settings

1. Set directories in Initiate Bioscore-file

Set the file paths in the Initiate Bioscore-file (Table 2).

Table 2: Table with directories and their description

Directory	Description
user_dir	directory within which all subdirectories are located
base_dir	subdirectory of user.dir, where all input data is stored and all output data can be written.
github_dir	directory of your local github repository of the BioScore scripts.
BS_module	Either 'terrestrial' or 'wetlands'

2. System settings

BioScore system configuration script contains all the pathways and directories for the BioScore script to run and find the input data (Table 3).

Table 3: Table with directories and their description

Directory	Description
var_fit_dir	subdirectory of base.dir, with rasters of environmental variables to be used for the fitting and projection of the regression models (Fig1).
species_in_dir	file located in a subdirectory of base.dir, with .csv file species observations per plot (Fig1).
plots	file located in a subdirectory of base.dir, with all plots derived from EVA-database. This database is used in combination with the files in species_in_dir to determine the absence values (Fig1).
species_out_dir	subdirectory of base.dir, in which all output files are stored. Always include the name of the script followed by date in the name of the output directory. For e.g., BioScore_Core_240122. Do not forget to add the date. (Fig1)
species_special	file located in a subdirectory of base.dir, with list of characteristic/special group of species.

TableWithOverviewVariables file located in var_fit_dir with the OverviewVariables file.

3. Model settings

BioScore model settings contains all the model parameters which determine the parameters for the model runs (Table 4).

Table 4: Table with parameters and their description

Parameter	Description
taxo_group	Taxon group used to select species, name should matche the taxon groups mentioned in the species input data
SaveIntRes	logical. save intermediate biomod2-results. default is FALSE.
min_obs_per_var	minimum number of observation per variable for the SDMs to be fitted. either 5 or 10.
Name	name of the person who runs the script. This is saved to the logfile.
DropHabitats_plots	list of characters referring to habitat types which should be removed from the plots. e.g.inland waters, fresh water, marine water and man-made
DropHabitats_species	list of characters referring to habitat types for which the characteristic species should be removed, e.g.inland waters, fresh water, marine water and man-made
MinYear	first year of the period for which you want to select the plots
VIFthreshold	maximum VIF-value, in order to decrease correlation between variables
PerformanceMetric	Metric used to weigh the range maps per algorithm into an ensemble map. Either "ROC" or " TSS".
modeltesting	Do you want to run the model for testing purposes only? Setting this to TRUE will only project SDMs for Northern Italy and will select a smaller set of absences, which will speed up the running time, but will also alter model results.
ProjectScenarios	names of scenarios to be run. Scenario names should be identical to the names in Input/Variables/OverviewVariables.csv, but can be a selection of scenarios in this table.
MIVCalc	logical, set to TRUE if you want to calculate and save MIVs and response curves
AllMetricsSave	logical, set to TRUE if you want to save the all metric file or not

Removeoriginal	logical, set to TRUE if you want to remove the species specific stats files
RunInParallel	logical, set to TRUE if the model should be run in parallel
CPU	number of CPU's to use when running the model in parallel

Step 4: Run the initiate BioScore script

The IntiateBioScore script runs the full BioScore model. The full BioScore model follows the steps mentioned below:

1. The model set loads the libraries required with their specific versions. It creates the directories required further to store the output from the model run. And further loads all the core functions required to run the BioScore model.
2. The create log file has functions to open a logfile and write in the logfile.
3. Loading and formatting of the data is done with functions for loading and formatting the environmental data, including the following functions which make a plot selection, load environmental raster with an ecosystem specific list of rasters to load and extract environmental values at locations of plots with an ecosystem specific list of rasters which are then written as table to hard disk. Finally it also performs VIF analysis such that it keeps only environmental variables in table which have a VIF < 5 or 10.
4. Species selection is performed to select species for specified habitats and taxon group.
5. Then presences and absences data for selected species is performed from the input species data table and images of plot location is created and saved to the hard disk.
6. Species presence and absence data is formatted for fitting the SDMs and to create projections.
7. Then the SDMs are fitted twice using the biomod2 package, once for cross validation (80% for fitting, 20% for testing) and once with all data (100% for fitting).
8. The fitted, cross validated SDMs are then evaluated by calculating the optimal TSS, AUC or MCC value. The algorithms are weighted with a metric of choice, set with the BinarizationMetric parameter. Either MCC, TSS or ROC.
9. The variable importance is calculated based on the model all data.
10. The response curves and the modelled indicator values (MIVs) are calculated, if MIVCalc is set to TRUE.
11. Projections of the SDMs are made (based on the SDMs fitted with all data) and a map with the ensembled probability of occurrence is also made. Four cut-off values are extracted (if available in a previously calculated AllMetrics

file, see Table 4) or the PoO cut off values for binarization are calculated. The metrics used to calculate the optimal cut off values are maximising true skill statistic (TSS), minimising the difference between sensitivity and specificity (DSS), maximising Matthew's correlation coefficient (MCC), and maximising F-measure (F).

12. All the stats specific to the species are saved to the hard disk.
13. Steps 5 to 13 are repeated for each species with help of the SDM function. This can be done in parallel.
14. After the whole run, all the species specific metrics are saved in one Allmetrics file and the species for which the SDMs were not fitted are written in the log file.

Output data

Various outputs are created and saved on the hard disk and saved within the base_dir-directory in the folder Model/Output. Below you will find a description of the outputs in the table below (Table 5)

Table 5: Table with output files/folders and their description

File/Folder	Description
CorrelationPressures	A file with correlations between the environmental variables.
Logfile_date&time	A log file containing log on the model run.
ImagesOfRangeMaps	A folder containing all the images of the range maps created per scenario.
PlotLocations	A folder containing all the images of the plot locations per species which were included to fit the SDMs.
RangeMaps	A folder containing all the range maps created per scenario.
ResponseCurves	A folder containing all the response curves from the fitted SDMs.
SpeciesMetrics	A folder containing AllMetrics file. This is a metric file containing all the metrics from the fitted SDMs for all the species for which the SDMs were run.

References

Chytrý, M., et al. (2016). "European Vegetation Archive (EVA): an integrated database of European vegetation plots." *Applied Vegetation Science* 19(1): 173-180. <https://doi.org/10.1111/avsc.12191>

Thuiller, W., et al. (2016). "biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7.1. <https://CRAN.R-project.org/package=biomod2>."