Biomod2EZ Manual and Tutorial

May 23, 2017

Type - Script Suite

Title - Report Generation for Species Distribution Modeling with Biomod2

Version - 1.0

Date - 2016-08-03

Authors - Paul R. Sesink Clee and Stephen Woloszynek

Maintainer - Paul R. Sesink Clee <psesinkclee@drexel.edu>

Contact - Paul R. Sesink Clee <psesinkclee@drexel.edu >

Description - Set of scripts to simplify the use of Biomod2 for species distribution modeling and ensemble creation with addition summary report generation features that are absent in the standard Biomod2 release.

Depends - R (>=3.2.1), biomod2, stats, utils, sp, raster, parallel, reshape, ggplot2, rgdal, base, rpart, rpart.plot, rattle, readr, pander, stringr, rmarkdown

I. INTRODUCTION

The goal of this script suite is to simplify the use of Biomod2 (Thuiller, 2016) for creating ecological niche models and ensembles for a given study taxa. Biomod2 can be used to create and project distribution models, but it falls short in its lack of report generation to concisely display results in an easily digestible format. If lengthy models are run without saving the results from the R console, users can lose information such as environmental predictor contributions and model performance scores. Additionally, saving results manually can become tedious when running multiple models in series. Biomod2EZ includes scripts that simplify and thoroughly explain all steps in the modeling process for new users, while incorporating a new report generation feature that provides users with an exported file that summarizes model results, ensemble projections, model performance scores, and environmental predictor contributions.

Biomod2EZ is distributed as a compressed file that should be extracted into a user's working directory. Biomod2 is a product of Thuiller et al. (2016) and is implemented unaltered in this script suite to simplify use and generate a visual report of results. This README should be read carefully before using the script suite for model production.

II. SCRIPT FUNCTIONS

1_DataInput_ModelInitiation.R

Presence/Absence Data

The table of presence and absence coordinates should have the following columns: (1) X coordinate column consisting of decimal degree longitude designation; (2) Y coordinate column consisting of decimal degree latitude designation; (3) species column used to identify points as presence ("1") or absence ("0") of your study taxa. This file must include a header row with column labels. For an example file, see: *Biomod2_EZ_SampleData/presabs.csv*

Environmental Variables

Map layers for the target study region should have identical extents and spatial resolution. Biomod2 can read different raster formats, but using ASCII grid (**.asc*) formatted files is suggested for universal compatibility between different packages/programs. It is good practice to keep all layers organized in the same folder to simplify the input of filepaths.

If you want to project your model to different layers or under climate change scenarios, you must uncomment the bottom section that contains a template raster stack, and enter filepaths to these projection layers.

DedupingPresAbs.R

This supplemental script can be used to remove duplicate presence and absence points (independently) from a presence/absence table file. This is a useful step to clean up your dataset and can help models run quickly while reducing biased sampling.

InstallingPackages.R

This script can be called directly from the first script (1*_DataInput_ModelInitiation.R*). It contains loops that check for, install, and load a variety of packages that are used by this script suite to accomplish data extraction, model building, and report generation.

ParametersAndReportGeneration.R

This script can be called directly from the first script (1*_DataInput_ModelInitiation.R*), but it is advised that you instead run it in blocks while becoming familiar with this script suite. This will help you pinpoint issues with your datasets, should any errors occur. For an exhaustive description, see the following section and detailed script annotation.

Biomod2_Report.Rmd

This *rmarkdown* file is used to combine exported model results/maps/tables into a rendered report HTML file.

Maxent.jar

This java executable is used by Biomod2 to run Maxent models (Phillips et al., 2006).

III. MODEL PREPARATION AND GENERATION

Loading Data

BIOMOD_FormattingData()

Formatting user-inputting data for use by Biomod2. Actions include: assigning presence and absence designation to *resp.var*, species coordinates to *resp.xy*, species name to *resp.name*, and raster stack of environmental layers to expl.var. These objects are used to format the user-inputted data for Biomod2.

BIOMOD_ModelingOptions()

Setting parameters for individual models. This block is mostly commented out by default, but it is suggested that you research each model, and determine parameters that will best suit your dataset.

Running Models

BIOMOD_Modeling()

Setting parameters for individual models. Options include: deciding which of the available models you want to run (*GLM* , *GAM* , *GBM* , *SRE*, *CTA*, *ANN* , *FDA*, *MARS*, *RF*, *Maxent.Phillips*, *Maxent.Tsuruoka*), number of replicates for each model (*NbRunEval*), splitting your dataset for training/testing (*DataSplit*), and evaluation metrics for models (*models.eval.meth* = *TSS*, *ROC*, *KAPPA*) among other parameters that are detailed in script annotations.

Creating Ensemble

BIOMOD_EnsembleModeling()

Setting parameters for ensemble forecasting. Options include: choosing which models to include in the ensemble (*chosen.models*), evaluation metrics (*eval.metrics*), and a set of options that involve weighting contributing models.

BIOMOD_Projection()

Setting parameters for projecting individual models. Options include: setting a new raster stack for projecting under climate scenarios or to a different area (*new.env*), selecting which models to project (*selected.models*), and output projection format (*output.format*).

BIOMOD_EnsembleForecasting()

Projecting ensemble model defined by BIOMOD_EnsembleModeling(). Additionally, the projected ensemble map is converted it to an ASCII grid raster and exported to the working directory for use in external mapping software.

Variable Decision Tree

This section uses the user-inputted presence/absence data to create a simple decision tree based on underlying environmental variables. This is performed by converting coordinates to the *SpatialPoints* format to ease the process of extracting environmental data at each point. These extracted variables are saved as a table to the working directory (*xy_value_extract.csv*). Finally, the *rpart* package is used to build a tree which is plotted using *rpart.plot*::*fancyRpartPlot*() and saved to the /Plots folder in the working directory as a PNG.

Generating Report

This block is used to export data structures and images for generating the final report using *rmarkdown*. An included function loops through a list of all model names and checks to see if each model was run successfully. If a model worked, the map is exported to the /Plots folder in the working directory as a PNG. If a model failed, it copies a blank placeholder image for display in the report.

Portions of the BiomodModelOut and BiomodModelEval objects are saved as R objects (RDS). These include: evaluation methods for all models and iterations that were run, variable importance for each iteration, and evaluation of the ensemble. Finally, *rmarkdown* is used to call and render Biomod2_Report.rmd to an html file that organizes all results and analyses. Note that there is also commented out code that can be used to instead export reports as PDFs (this requires a properly initialized LaTeX installation).

IV. REPORT COMPONENTS

Ensemble Projection

Displays map of the resulting ensemble model projection.

Testing Model Performance

Receiver Operating Characteristic (ROC, Hanley et al., 1982)

The area under the *ROC* curve is commonly used as a standalone measure of model performance. Values range from 0.5 to 1.0 where values low values reflect a model that is no better than a random association between species presence/absence and underlying environmental variables and high values close to 1.0 reflect a very strong signal of association.

True Skill Statistic (Hassen-Kruipers discriminant) (TSS, Monserud et al., 1992)

TSS compares the number of correct forecasts to a hypothetical set of perfect forecasts. Values range from -1 to +1, where values less than zero indicate that the model performs no better than random and values close to 1 perform very well. *TSS* is similar to *KAPPA*, but it is not affected by prevalence or by the size of the validation set.

Cohen's KAPPA (Allouche et al., 2006)

This metric uses the accuracy expected by chance to correct overall accuracy of a model. Values range from -1 to +1, where values less than zero indicate that the model performs no better than random and values close to 1 perform very well. It should be noted that *KAPPA* is criticized for being dependent on prevalence, which introduces bias to estimates of accuracy.

Individual Model Types

Generalized Linear Model (GLM)

Linear regression model that allows for non-linearity and is based on an assumed relationship between the response variable and predictor variables (Nelder et al., 1989).

Generalized Boosting Model (GBM)

A powerful machine-learning algorithm that can be used to fit regressions, perform classifications, and determine ranking. It applies boosting methods to regression trees by creating simple trees where each tree is based on the prediction residuals of the previous tree and each node is set on a binary decision. Each subsequent tree is used to find a new partition in the dataset that can further reduce error (Ridgeway, 2007).

Generalized Additive Model (GAM)

Combines aspects of both additive models and generalized linear models. They function like a *GLM* in that they can have different error structures and link functions, but instead of having an explicit functional form, the relationship uses non-parametric smoothers to describe the relationship. They are useful for distributions that have complex shapes (Hastie et al., 1990; Hastie, 2013).

Classification Tree Analysis (CTA)

A type of machine learning algorithm used for classifying remotely sensed data in support of land cover mapping and analysis. Classification trees structurally determine binary decisions to estimate the dependent variable (Therneau et al., 2010).

Artificial Neural Network (ANN)

A computational model based on the way that biological neural networks function. This type of model changes while information is fed through it and, in a sense, learns how to improve the model over subsequent iterations (Ripley et al., 2011).

Surface Range Envelope (SRE)

Same as *BIOCLIM* . Determines habitat suitability at each grid cell by comparing values of environmental variables there to a percentile distribution of variables at locations of known presence. The closer that habitats across the study region are to known suitable habitats (at locations of species presence), the more suitable the location is deemed (Busby, 1991).

Flexible Discriminant Analysis (FDA)

Used to predict a categorical dependent variable (ie. presence or absence) using one or more predictor variables. It is also known as 'pattern recognition', 'supervised learning', and 'supervised classification'. This differs from a cluster analysis, which is unsupervised. Objects with known groups are used to then determine the category that ungrouped objects fall in. This is done by identifying relationships among groups' covariance matrices to be able to discriminate between different groups. It is important to note that with N number of groups, the model requires $N - 1$ number of predictor variables (Febrero-Bande et al., 2012).

Multiple Adaptive Regression Splines (MARS)

Non-parametric regression technique that automatically models nonlinearities and interactions between variables. Starting with the mean of the response variables, the model finds basis functions that result in the smallest residual error. Each consecutive basis function consists of one term that is already in the model multiplied by a new hinge function (consisting of a variable and a knot). The *MARS* model, when creating new basis functions, must scour over all possible combinations of existing terms and all variables (Milborrow, 2013).

Random Forests (RF)

Non-parametric regression technique. Response is tested against predictor variables, and the model tries to split response variables into 2 groups that have the smallest amount of variation (presence vs. absence) in each part (this continues until it builds a full decision tree). Variables can show up in multiple locations in the tree. Pruning trees defines where to stop tree building (after how many levels). Each run randomizes presence/absence points used and environmental predictors used without using all of them at once. This allows the model to determine which variables make the model performance drop when removed (highly important variables) across many trees. With RF, you do not necessarily need to reduce the predictor set because it will only use the best variables for the final model (Liaw et al., 2002).

Maxent (MAXENT.Phillips)

Uses environmental data for known presence localities and for a large set of background points (or pseudoabsences) in a machine learning methodology using the principle of maximum entropy to model species distributions. This process chooses models with uniform/spread-out distributions while considering the study region as a density estimation of presence (Phillips et al., 2006).

Maxent (MAXENT.Tsuruoka)

Unlike the Phillips version that runs using java, the R package to implement a maximum entropy approach in species distribution modeling focuses on minimizing memory consumption for large datasets and is based on an efficient C++ implementation (Tsuruoka).

Ensemble Evaluation

Table showing the results of the model evaluation statistics from the list that is defined by *models.eval.meth*.

Presence/Absence Decision Tree

Environmental variables are extracted at each presence/absence point and are saved to a table in the working directory (*xy_value_extract.csv*). They are then used to create a Spatial Points object to create a simple classification tree using the rpart package (Therneau et al., 2010).

Variable Importance

Relative importance of each variable calculated for all models and runs.

Failed Models

If any models fail to complete properly, they will be listed here.

V. TUTORIAL

A sample dataset is also included in the download of the Biomod2EZ script suite (/Biomod2EZ_SampleData). This includes a presence/absence table and environmental layers, both of which are required input files to generate species distribution models and ensemble. This data was made with computing power in mind and is small enough to run quite quickly on a personal computer.

Input Files

presabs.csv
This table is made of the following columns: *X*, *X_WGS*84, *Y_WGS*84, and *TestTaxa*. *X* is for individual sample identifiers, *X_WGS*84 is the X coordinate for samples, *Y_WGS*84 is the Y coordinate for samples, and *TestTaxa* contains binary indicators of sample presence (1) and absence (0).

/environmental_layers
This folder contains ASCII rasters for 7 sample environmental layers that can be used to create tutorial species distribution models/ensemble and generate a summary report.

Data Input

1. Download and unzip the Biomod2EZ into your working directory.

2. In R, open 1_*DataInputModelInitiation.R*

3. In the "Data Setup" section, places where inputs that are needed should replace the existing *XXXXXXXX* placeholders. Each line is annotated with a comment that describes what data needs to be inputted.

   - setwd() is used to set your working directory. This line requires the full filepath to the folder that you are using for your working directory (ex. "D:/R/Biomod2_tutorial"). This folder should contain all of the unzipped Biomod2EZ files.

   - *prstbl* is the identifier for the presence table. Input the filepath to the presence table (presabs.csv) here (ex. "D:/R/Biomod2_tutorial/Biomod2EZ_SampleData/presabs.csv"").

   - *myRespName* is the identifier for the column in the presence/absence table that contains the species data (in this case, "*TestTaxa*").

   - *xname* is the identifier for the column in the presence/absence table that contains the X coordinate or longitude for samples (in this case, "*X_WGS*84").

   - *yname* is the identifier for the column in the presence/absence table that contains the Y coordinate or latitude for samples (in this case "*Y_WGS*84").

- *myExpl* is the identifier for a raster stack of all environmental layers used to create the species distribution models. Each line in this stack is for one ASCII raster from the /environmental_layers folder (ex. "D:/R/Biomod2_tutorial/Biomod2EZ_SampleData/environmental_layers/annualprecip.asc").
- *projname* is the identifier for a name for your project (ex. "Biomod2EZ_Tutorial")

Model Parameters, Projections, and Report Generation

1. In R, open */ParametersAndSettings/ParametersAndReportGeneration.R*

2. This script contains all internal settings for model/ensemble creation and preparing parts for the summary report generation. For more information about adjusting settings for Biomod2, see this Biomod Tutorial that uses the same object naming structure: http://www.will.chez-alice.fr/pdf/BiomodTutorial.pdf

3. The last 2 sections of this script deviate from the original Biomod2 framework.

   - Variable Decision Tree - extracts environmental data to points and builds a decision tree for presence/absence from it using the *rpart* package.

   - Generating Report - plots model projections and saves objects that are later included in the summary report.

4. Run all blocks of this script in chunks so that any potential errors will be not be missed. You will end with a report in your working directory called "Biomod2_Report.html" that displays all resulting maps and results and an ASCII raster called "MyEnsembleRaster.asc" that can be used to create maps in external mapping software.

REFERENCES

Allouche, O., Tsoar, A. and Kadmon, R. (2006) 'Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (tss)', *Journal of applied ecology,* 43(6), pp. 1223-1232.

Busby, J. R. (1991) 'Bioclim – a bioclimate analysis and prediction system', in Margules, C.R. & Austin, M.P. (eds.) *Nature conservation: Cost effective biological surveys and data analysis*: CSIRO, pp. 64–68.

Febrero-Bande, M. and Oviedo De La Fuente, M. (2012) 'Statistical computing in functional data analysis: The r package fda. Usc', *Journal of Statistical Software,* 51(4), pp. 1-28.

Hanley, J. A. and Mcneil, B. J. (1982) 'The meaning and use of the area under a receiver operating characteristic (roc) curve', *Radiology,* 143(1), pp. 29-36.

Hastie, T. (2013) 'Gam: Generalized additive models, r package, version 0.98', *URL http://CRAN. R-project. org/package= gam*.

Hastie, T. and Tibshirani, R. (1990) *Generalized additive models.* London: Chapman & Hall.

Liaw, A. and Wiener, M. (2002) 'Classification and regression by randomforest', *R news,* 2(3), pp. 18-22.

Milborrow, S. (2013) 'Derived from mda: Mars by trevor hastie and rob tibshirani. Earth: Multivariate adaptive regression spline models, 2011', *R package http://CRAN. R-project. org/package= earth. Cited on*, pp. 4.

Monserud, R. A. and Leemans, R. (1992) 'Comparing global vegetation maps with the kappa statistic', *Ecological Modelling,* 62(4), pp. 275-293.

Nelder, J. A. and Mccullagh, P. (1989) *Generalized linear models, 2nd edition. Monographs on statistics and applied probability*: CHAPMAN & HALL/CRC.

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006) 'Maximum entropy modeling of species geographic distributions', *Ecological Modelling,* 190(3), pp. 231-259.

Ridgeway, G. (2007) 'Generalized boosted models: A guide to the gbm package', *Update,* 1(1), pp. 2007.

Ripley, B. and Venables, W. (2011) 'Nnet: Feed-forward neural networks and multinomial log-linear models', *R package version,* 7(5).

Therneau, T. M., Atkinson, B. and Ripley, M. B. 2010. The rpart package.

Tsuruoka, Y. (2006) 'A simple c++ library for maximum entropy classification (2006)', *Software available at http://www-tsujii. is. su-tokyo. ac. jp/tsuruoka/maxent*.