Valavi, R., G. Guillera-Arroita, J. Lahoz-Monfort, and J. Elith. 2021. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs.

Data S1

Data and codes for model fitting and evaluation

Authors

Roozbeh Valavi
School of Biosciences, University of Melbourne, Parkville, VIC 3010, Australia
rvalavi@student.unimelb.edu.au and valavi.r@gmail.com

Gurutzeta Guillera-Arroita
School of Biosciences, University of Melbourne, Parkville, VIC 3010, Australia
gurutzeta.guillera@unimelb.edu.au

José J. Lahoz-Monfort
School of Biosciences, University of Melbourne, Parkville, VIC 3010, Australia
jose.lahoz@unimelb.edu.au

Jane Elith
School of Biosciences, University of Melbourne, Parkville, VIC 3010, Australia
j.elith@unimelb.edu.au

File list

The `background_50k` folder contains 50,000 random background sample we used for modelling the species in each region. It includes the following files that has the background samples corresponding to each region name.

- `AWT.csv` e.g., the 50k random background samples for species from the **AWT** region
- `CAN.csv`
- `NSW.csv`
- `NZ.csv`

- SA.csv
- SWI.csv

The modelling_codes folder contains codes for fitting all the models and evaluate the predictions. Each of the following .R files has the code for fitting the corresponding modeling method.

- biomod.R
- BRT.R
- cforest_w.R
- cforest.R
- Ensemble.R
- GAM.R
- GLM_unw.R
- GLM.R
- IWLR_GAM.R
- IWLR_GLM.R
- Lasso.R
- MARS.R
- MaxEnt_tuned.R
- MaxEnt.R
- MaxNet.R
- RF_downsampled.R
- RF.R
- Ridge_regression.R
- SVM_w.R
- SVM.R
- XGBoost.R

The two prediction_helper.R and vars.R codes provide variable and functions for several of the other models including: GAM.R, GLM_unw.R, GLM.R, IWLR_GAM.R, IWLR_GLM.R, Lasso.R, Ridge_regression.R, MaxEnt_tuned.R,and MaxEnt.R

The model_evaluation.R file is the code for evaluation of all the models. It calculates, AUC_{ROC}, AUC_{PRG} and COR for all species in all the models.

The `prediction_functions.R` file contains the functions for predicting with *e1071* (SVMs) and *glmnet* (Lasso and ridge regression) packages to raster layers for predicting spatial distribution of species.

Description

Notice that each csv file in the background folder has the exact column names and the same order from the presence-only training data in the *disdat* R package. This data can be directly used in the modelling code explained above.

For reproducing modelling, the code should be organized in the … folder and an output directory should be set; in the current code is specified as `models_output`, but you can change it to whatever you wish. Just remember to change the same directory in the evaluation codes.

Note that Ensemble model is an average of five models including BRT, GAM, Lasso, MaxEnt, and RF down-sample. For predicting with this model, you need to fit all the component models first, then run `Ensemble.R` code.

For model evaluation, all the models should be fitted first, then run model evaluation code (`model_evaluation.R)` to produce the result.