An example of species distribution modelling with ${\tt biomod2}$

 $\begin{array}{c} {\tt biomod2~version:1.0} \\ {\rm R~version~2.15.1~(2012\text{-}06\text{-}22)} \end{array}$

Damien Georges & Wilfried Thuiller

July 5, 2012

Contents

1	Introduction	3
2	Formatting the data	4
3	Modeling3.1 Building models	
4	Projection	17

1 Introduction

This vignette illustrates how to build, evaluate and project a single species distribution model using biomod2 package. The three main modeling steps, described bellow, are the following:

- 1. formatting the data
- 2. computing the models
- 3. making the projections

The example is deliberately simple (few technicals explanations) to make sure it is easy to transpose to your own data relatively simply.

NOTE 1:

Several other vignettes will be written soon to help you to go through biomod2 details and subtleties

2 Formatting the data

In this vignette, we will work (because it is the most common case) with:

- only presences data that we will be extracted from a raster
- environmental raster layers (e.g. Worldclim)

Let's import our data.

```
\_ R input _{	extstyle -}
# load the library
library(biomod2)
# load our species raster
# we consider only the presences of Myocastor coypus species
myResp.ras <- raster( system.file(</pre>
                          "external/species/Myocastor_coypus.img",
                          package="biomod2") )
# extract the presences data
# the name
myRespName <- 'Myocastor'</pre>
# the XY coordinates of the presence
myRespXY <- xyFromCell(object=myResp.ras,</pre>
                         cell=which(myResp.ras[]>0))
# and the presence data
myResp <- extract(x=myResp.ras, y=myRespXY)</pre>
# load the environmental raster layers (could be .img, ArcGIS rasters or any supported for
# Environmental variables extracted from Worldclim (bio_3, bio_4,
# bio_7, bio_11 & bio_12)
myExpl = stack( system.file( "external/climat/current/bio3.grd",
                               package="biomod2"),
                  system.file( "external/climat/current/bio4.grd",
                               package="biomod2"),
                  system.file( "external/climat/current/bio7.grd",
                               package="biomod2"),
                  system.file( "external/climat/current/bio11.grd",
                               package="biomod2"),
                  system.file( "external/climat/current/bio12.grd",
                               package="biomod2"))
```

NOTE 2:

You may have community or atlas data for which you have both presence and absence. In this case extract the presences and the absences points and code them by 0/1.

NOTE 3:

If your environmental data are in matrix/data.frame format, you have to give a species as vector (or a one column Spatial.points.data.frame) having a length that match with the number of rows of your environmental data. That implies to add NA's in all points where you do not have information on species presence/absence.

When your data are correctly loaded, you have to transform them in an appropriate biomod2 format. This is done using BIOMOD_FormatingData. As all models need both presences and absences to run, you may need to add some pseudo-absences (or background data) to your data. That is necessary in the case of presence-only, and may be useful in the case of insufficient absence data. 3 algorithms are now implemented to extract a range of pseudo-absence data: 'random', 'SRE' and 'disk'.

Here, we will create two sets of pseudo-absence data using the random algorithm.

NOTE 4:

If you have both presence-absence data and a large number of presence (not the case here), it's strongly recommended to split your data.frame into two pieces and to keep a part for evaluating all your models on the same data.set (i.e. eval.xxx args)

NOTE 5:

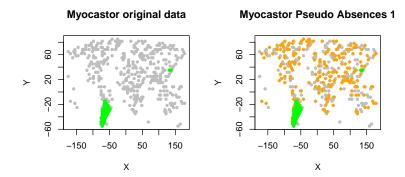
The PA.nb.absences arg represents the total number of pseudo-absence extracted for each set of extraction (true absences + selected PA). It must be then higher than the number of true absences (if any). If not, no pseudo-absences are selected.

```
R output _______ R output _______ ! No data has been set aside for modeling evaluation
```

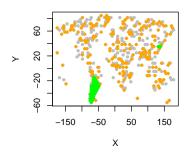
- > Pseudo Absences Selection checkings...
- > random pseudo absences selection
- > Pseudo absences are selected in explanatory variables

At this point, check whether the data are correctly formatted by printing and plotting the created object.

```
\_ R input \_
myBiomodData
_____ R output ______ R output _____
sp.name = Myocastor
       59 presences, 0 true absences and 383 undifined points in dataset
       5 explanatory variables
    bio_3 bio_4
                              bio_7
Min. : 9.95 Min. : 103 Min. : 54.4
Median: 41.93 Median: 5386 Median: 278.1
Mean :42.09 Mean : 6624
                           Mean :288.8
3rd Qu.:56.00 3rd Qu.:10811
                          3rd Qu.:393.5
Max. :91.43 Max. :21774
                          Max. :710.3
   bio_11
                 bio_12
Min. :-426.7 Min. : 4
1st Qu.:-151.9 1st Qu.: 259
Median : 63.9 Median : 620
Mean : 15.7
              Mean : 938
3rd Qu.: 200.4
              3rd Qu.:1319
Max. : 275.0
              Max. :5133
2 Pseudo Absences dataset available ( PA1 PA2 ) with 200 absences in each (true abs + pseudo Absences)
-----
                        \_ R input \_
plot(myBiomodData)
```



Myocastor Pseudo Absences 2



The colors for this plot match with...

- Presences
- Absences
- Pseudo Absences
- Remaining Backgroud

3 Modeling

3.1 Building models

This step may be considered as the core of the modeling procedure within biomod2. Here you have to choose between 10 different algorithms ('GLM', 'GBM', 'GAM', 'CTA', 'ANN', 'SRE', 'FDA', 'MARS', 'RF', 'MAXENT'). Before running the models, you can customize their set of parameters and options using BIOMOD_ModelingOptions. The created object is then given to BIOMOD_Modeling in the next step. For the sake of simplicity, we keep all default options.

NOTE 6:

A vignette on models' parametrization will be available soon

```
# 2. Defining Models Options using default options.

myBiomodOption <- BIOMOD_ModelingOptions()
```

We are now ready for running the set of models on our species. As we do not have evaluation data, we will make 1-fold cross-validation (number controlled by NbRunEval argument) of our models by randomly splitting our data set into 2 subsets: DataSplit % for calibrating and training the models and the remainder for testing them. Each model will be tested (and evaluated if any evaluation data is given) according to models.eval.meth evaluation metrics (chosen into 'KAPPA', 'TSS', 'ROC', 'FAR', 'SR', 'ACCURACY', 'BIAS', 'POD', 'CSI' and 'ETS'). To ensure our models will be comparable in term of scale, we decided to rescale them all with a binomial GLM (rescal.all.models). The VarImport argument corresponds to the number of resampling of each explanatory variable to measure the relative importance of each variable for each selected model.

NOTE 7:

No weights are given but some will be automatically generated. Indeed, in the particular case of pseudo-absence selection, we make sure the prevalence is kept to 0.5. It means that the presence data have the same weight than the pseudo-absence data, even if a large number of the latter has been extracted.

```
Loading required library...
```

Checking Models arguments...

```
Creating suitable Workdir...
----- Myocastor Modeling Summary ------
5 environmental variables (bio_3 bio_4 bio_7 bio_11 bio_12)
Number of evaluation repetitions: 2
Models selected : SRE CTA RF MARS FDA
Total number of model runs : 20
---- Run : Myocastor_PA1
----- Myocastor_PA1_RUN1
Model=Surface Range Envelop
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Classification tree
       5 Fold Cross-Validation
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Breiman and Cutler's random forests for classification and regression
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Multiple Adaptive Regression Splines
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Flexible Discriminant Analysis
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
-=-=- Myocastor_PA1_Full
Model=Surface Range Envelop
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Classification tree
       5 Fold Cross-Validation
       Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Breiman and Cutler's random forests for classification and regression
```

```
Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Multiple Adaptive Regression Splines
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Flexible Discriminant Analysis
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
-=-=- Run : Myocastor_PA2
----- Myocastor_PA2_RUN1
Model=Surface Range Envelop
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Classification tree
        5 Fold Cross-Validation
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Breiman and Cutler's random forests for classification and regression
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Multiple Adaptive Regression Splines
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Flexible Discriminant Analysis
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
-=-=- Myocastor_PA2_Full
Model=Surface Range Envelop
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
Model=Classification tree
        5 Fold Cross-Validation
        Evaluating Model stuff...
        Evaluating Predictor Contributions...
```

```
Model=Breiman and Cutler's random forests for classification and regression
      Evaluating Model stuff...
       Evaluating Predictor Contributions...
Model=Multiple Adaptive Regression Splines
       Evaluating Model stuff...
      Evaluating Predictor Contributions...
Model=Flexible Discriminant Analysis
      Evaluating Model stuff...
       Evaluating Predictor Contributions...
When this step is over, have a look at some outputs:
  • modeling summary
                       _____ R input _____
     myBiomodModelOut
    ______ R output ___
    BIOMOD.models.out
    Specie modelised : Myocastor
    Considered variables: bio_3 bio_4 bio_7 bio_11 bio_12
    Computed Models: Myocastor_PA1_RUN1_SRE Myocastor_PA1_RUN1_CTA Myocastor_PA1_RUN1_RF
    Failed Models : none
    -=-=-=-=-=-=-
   • models evaluations
     # get all models evaluation R input _
     myBiomodModelEval <- getModelsEvaluations(myBiomodModelOut)</pre>
     # print the dimnames of this object
     dimnames(myBiomodModelEval)
                        ____ R output __
    [[1]]
    [1] "TSS" "ROC"
    [[2]]
```

```
[1] "Testing.data" "Cutoff"
                             "Sensitivity"
[4] "Specificity"
[[3]]
[1] "SRE" "CTA" "RF" "MARS" "FDA"
[[4]]
[1] "RUN1" "Full"
[[5]]
Myocastor_PA1 Myocastor_PA2
                 "PA2"
       "PA1"
# let's print the TSS scores of Random Forest
myBiomodModelEval["TSS","Testing.data","RF",,]
                  _____ R output _____
      PA1 PA2
RUN1 0.775 0.867
Full 1.000 1.000
# let's print the ROC scores of all selected models
myBiomodModelEval["ROC", "Testing.data",,,]
                _____ R output __
, , PA1
     RUN1 Full
SRE 0.733 0.816
CTA 0.821 0.988
RF 0.926 1.000
MARS 0.948 0.980
FDA 0.948 0.960
, , PA2
     RUN1 Full
SRE 0.867 0.796
CTA 0.859 0.994
RF 0.982 1.000
MARS 1.000 0.979
FDA 0.992 0.975
```

```
____ R input __
• Relative importance of the explanatory variables
                              _{\scriptscriptstyle -} R input _{\scriptscriptstyle -}
   # print variable importances
   getModelsVarImport(myBiomodModelOut)
                     _____ R output __
  , , RUN1, PA1
         SRE CTA RF MARS FDA
  Var1 0.039 0.874 0.440 0.714 0.851
  Var2 0.039 0.263 0.122 0.460 0.000
  Var3 0.089 0.005 0.138 0.121 0.052
  Var4 0.106 0.209 0.348 0.504 0.667
  Var5 0.106 0.136 0.161 0.159 0.148
  , , Full, PA1
         SRE CTA RF MARS FDA
  Var1 0.010 0.855 0.381 0.567 0.823
  Var2 0.010 0.370 0.142 0.528 0.000
  Var3 0.058 0.038 0.085 0.325 0.018
  Var4 0.097 0.239 0.342 0.382 0.575
  Var5 0.043 0.404 0.179 0.119 0.142
  , , RUN1, PA2
                    RF MARS FDA
         SRE CTA
  Var1 0.045 0.901 0.500 0.517 0.902
  Var2 0.036 0.004 0.199 0.551 0.762
  Var3 0.103 0.222 0.105 0.344 0.576
  Var4 0.148 0.534 0.439 0.422 0.544
  Var5 0.062 0.106 0.071 0.294 0.199
  , , Full, PA2
         SRE CTA
                    RF MARS FDA
  Var1 0.018 0.904 0.445 0.492 0.910
  Var2 0.009 0.432 0.200 0.045 0.697
  Var3 0.092 0.096 0.119 0.578 0.873
  Var4 0.122 0.315 0.329 0.643 0.563
```

Var5 0.032 0.181 0.093 0.303 0.150

3.2 Ensemble modeling

Here comes one of the most interesting features of biomod2. BIOMOD_EnsembleModeling combines individual models to build some kind of meta-model. In the following example, we decide to exclude all models having a TSS score lower than 0.85.

NOTE 8:

Models are now combined by repetition, other way to combine them (e.g. by Models, all together...) will be available soon

```
R output ______ R output ______ ! all models available will be included in ensemble.modeling
> Evaluation & Weighting methods summary :
   TSS over 0.85
> PA1
> TSS
> models kept : Myocastor_PA1_Full_CTA, Myocastor_PA1_Full_RF, Myocastor_PA1_Full_MARS
> Mean of probabilities...
> Coef of variation of probabilities...
> Median of ptobabilities...
> Confidence Interval...
   > 2.5 %
   > 97.5 %
> Comittee averaging...
> Prababilities wegthing mean...
> PA2
> TSS
```

You can easily access to the data and outputs of BIOMOD_Modeling using some specific functions to make your life easier.

Let's see the meta-models evaluation scores.

NOTE 9:

We decide to evaluate all meta-models produced even the CV (Coefficient of Variation) one which is quite hard to interpret. You may consider it as: higher my score is, more the variation is localised where my species is forecasted as present.

```
\_ R input \_
 # print summary
myBiomodEM
                           R output -
-=-=- 'BIOMOD.EnsembleModeling.out'
sp.name : Myocastor
expl.var.names : bio_3 bio_4 bio_7 bio_11 bio_12
models computed: Myocastor_PA1_AllRun_EM.TSS, Myocastor_PA2_AllRun_EM.TSS
-------
                          \_ R input \_
# get evaluation scores
getEMeval(myBiomodEM)
                          _{-} R output _{-}
$Myocastor_PA1_AllRun_EM.TSS
, , em.mean
   Testing.data Cutoff Sensitivity Specificity
    0.985 483.0 100.00 98.5
TSS
         0.999 522.8
ROC
                           98.31
                                       98.5
```

```
, , em.cv
  Testing.data Cutoff Sensitivity Specificity
       0.000 1.000 0.000 6.0
        0.002 0.878
ROC
                      1.695
                                 1.5
, , em.ci.inf
   Testing.data Cutoff Sensitivity Specificity
TSS
       0.881 49.17 88.14 100
ROC
       0.941 100.91
                      88.14
                                  100
, , em.ci.sup
   Testing.data Cutoff Sensitivity Specificity
TSS 0.96 965 100 96
         0.98 1000 100
                                   96
ROC
, , em.median
   Testing.data Cutoff Sensitivity Specificity
TSS 0.975 289.1 100.00 97.5
        0.998 483.0
                      96.61
ROC
                                 97.5
, , em.ca
   Testing.data Cutoff Sensitivity Specificity
TSS 0.975 494.9 100 97.5
ROC 0.999 667.0 100 97.5
, , em.pmw
  Testing.data Cutoff Sensitivity Specificity
TSS 0.995 527 100 99.5
ROC
       1.000 533
                        100
                                99.5
$Myocastor_PA2_AllRun_EM.TSS
, , em.mean
   Testing.data Cutoff Sensitivity Specificity
TSS
   0.995 433.0 100 99.5
        1.000 444.8
ROC
                        100
                                 99.5
, , em.cv
   Testing.data Cutoff Sensitivity Specificity
TSS 0.000 0.000 100.000 0
ROC
       0.001 0.939
                      1.695
```

```
, em.ci.inf
    Testing.data Cutoff Sensitivity Specificity
TSS
           0.99 24.00
                           100.00
                                            99
ROC
           1.00 68.04
                             98.31
                                             99
, , em.ci.sup
    Testing.data Cutoff Sensitivity Specificity
TSS
          0.995 846.0
                             100
                                          99.5
           1.000 852.5
                               100
                                           99.5
ROC
, , em.median
    Testing.data Cutoff Sensitivity Specificity
          0.985 293.2 100.00
TSS
ROC
           1.000 389.2
                             98.31
                                           98.5
, , em.ca
    Testing.data Cutoff Sensitivity Specificity
TSS
          0.980 580.8
                               100
          0.999 667.0
                               100
                                             98
ROC
, , em.pmw
    Testing.data Cutoff Sensitivity Specificity
TSS
          0.995 442.5
                               100
                                           99.5
           1.000 457.6
ROC
                                100
                                           99.5
```

4 Projection

Once the models are calibrated and evaluated, we might want to project the potential distribution of the species over space and time. This is made using BIOMOD_Projection

NOTE 10:

All projections are stored directly on your hard drive

First let's project the individual models on our current conditions (the globe) to visualize them.

```
# projection over the globe under current conditions

myBiomomodProj <- BIOMOD_Projection(

modeling.output = myBiomodModelOut,
```

*** Myocastor_PA1_Full_MARS

*** Myocastor_PA1_Full_FDA

*** Myocastor_PA2_RUN1_SRE

*** Myocastor_PA2_RUN1_CTA

*** Myocastor_PA2_RUN1_RF

*** Myocastor_PA2_RUN1_FDA

*** Myocastor_PA2_RUN1_FDA

*** Myocastor_PA2_Full_SRE

*** Myocastor_PA2_Full_CTA

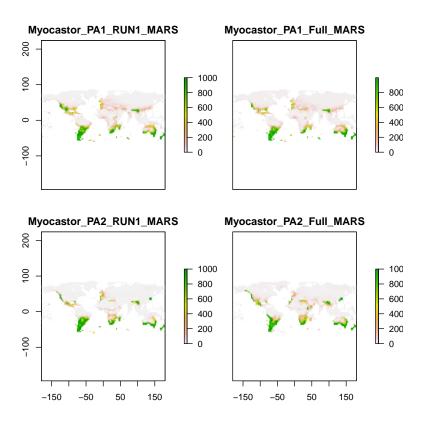
*** Myocastor_PA2_Full_RF

```
proj.name = 'current',
                      selected.models = 'all',
                      binary.meth = 'ROC',
                      compress = 'xz',
                      clamping.mask = F)
Loading required library...
Doing Models Projections...
*** Myocastor_PA1_RUN1_SRE
*** Myocastor_PA1_RUN1_CTA
*** Myocastor_PA1_RUN1_RF
*** Myocastor_PA1_RUN1_MARS
*** Myocastor_PA1_RUN1_FDA
*** Myocastor_PA1_Full_SRE
*** Myocastor_PA1_Full_CTA
*** Myocastor_PA1_Full_RF
```

new.env = myExpl,

R input _____

make some plots sub-selected by str.grep argument
plot(myBiomomodProj, str.grep = 'MARS')

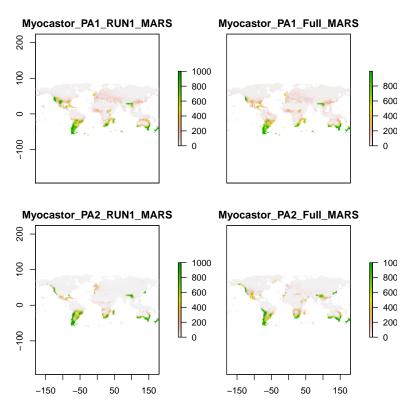


if you want to make custom plots, you can also get the projected map
myCurrentProj <- getProjection(myBiomomodProj)
myCurrentProj

Then we can project the potential distribution of the species over time, i.e. into the future.

```
package="biomod2"),
                     system.file( "external/climat/future/bio11.grd",
                                  package="biomod2"),
                     system.file( "external/climat/future/bio12.grd",
                                  package="biomod2"))
myBiomomodProj2050 <- BIOMOD_Projection(</pre>
                               modeling.output = myBiomodModelOut,
                               new.env = stack(myExp12050),
                               proj.name = 't2050',
                               selected.models = 'all',
                               binary.meth = 'ROC',
                               compress = 'xz',
                               clamping.mask = T)
R output _______ R output ______
  > defining clamping mask
Loading required library...
Doing Models Projections...
*** Myocastor_PA1_RUN1_SRE
*** Myocastor_PA1_RUN1_CTA
*** Myocastor_PA1_RUN1_RF
*** Myocastor_PA1_RUN1_MARS
*** Myocastor_PA1_RUN1_FDA
*** Myocastor_PA1_Full_SRE
*** Myocastor_PA1_Full_CTA
*** Myocastor_PA1_Full_RF
*** Myocastor_PA1_Full_MARS
*** Myocastor_PA1_Full_FDA
*** Myocastor_PA2_RUN1_SRE
*** Myocastor_PA2_RUN1_CTA
*** Myocastor_PA2_RUN1_RF
*** Myocastor_PA2_RUN1_MARS
*** Myocastor_PA2_RUN1_FDA
*** Myocastor_PA2_Full_SRE
*** Myocastor_PA2_Full_CTA
*** Myocastor_PA2_Full_RF
*** Myocastor_PA2_Full_MARS
*** Myocastor_PA2_Full_FDA
Binary transformations...
                       _____ R input ___
```

make some plots, sub-selected by str.grep argument
plot(myBiomomodProj2050, str.grep = 'MARS')



The last step of this vignette is to make Ensemble Forcasting, that means to project the meta-models you have created with BIOMOD_EnsembleModeling. BIOMOD_EnsembleForecasting required the output of BIOMOD_EnsembleModeling and BIOMOD_Projection. It will combine the projections made according to models ensemble rules defined at the ensemble modelling step.

```
"*** Myocastor_PA1_AllRun_EM.TSS ...

> em.mean
> em.cv
> em.ci.inf
> em.ci.sup
> em.median
> em.ca
```

```
> em.pmw

*** Myocastor_PA2_AllRun_EM.TSS ...
> em.mean
> em.cv
> em.ci.inf
> em.ci.sup
> em.median
> em.ca
> em.pmw
```

Nothing is returned but some additional files have been created in your projection folder (RasterStack or array depending on your projection type). This file contains your meta-models projections.

```
plot(Myocastor_PA1_AllRun_EM.TSS)
```

