Bakeoff Pipeline

Anna Norberg

03 July 2018

Bakeoff is a pipeline for fitting various single and joint species distribution models to community data. The main script used is 'bakeoff.pipeline.r', which guides through the importing and formatting the data, fitting the models, producing predictions based on validation data.

All the scripts sourced, the data used and the folder structure are available in Github: AnnaNorb/bakeoff. For the pipeline to run smoothly, the required folder structure is created as a part of the pipeline to the location indicated by the user. Please make sure, that the location of the pipeline (where it is downloaded), is also a location where additional results folders can be created and results saved in the end.

The pipeline works in R environment, except for the method HMSC, for which all models are fitted and predictions made in Matlab (see the document 'bakeoffHMSCvignette'). The workflow is the following: 1) first we construct the folder structure needed, 2) then fit the HMSC models in Matlab, 3) after which we fit all the rest of models in R, 4) produce all predictions and perfomance measures, 5) and plot the results.

1 Preparations

First clean the workspace, define the path for the location of the pipeline folder, and run a series of setting.

```
# preliminaries
#############################################################################################
rm(list = ls(all=TRUE)); gc()   # clear workspace

pth<-"..."  # write here the path to the location of the 'bakeoff' folder
#pth<-"~/OneDrive - University of Helsinki"
#pth<-"D:/HY-data/NORBERG/OneDrive - University of Helsinki"

SETT<-file.path(pth,"bakeoff","pipeline","SCRIPTS","settings.r")    # path to settings
source(SETT)    # run settings
setwd(WD)       # set working directory to the pipeline folder

source(file.path(SD,"pipe","get_os.r"))     # identify your OS
source(file.path(SD,"pipe","pkgs.r"))       # install required packages,
source(file.path(SD,"pipe","dirs.r"))       # create directories (if don't exist),
source(file.path(SD,"pipe","parall.r"))     # and identify your OS and define settings
                                            # for parallel computing
```

At this point, the user should switch to fitting the HMSC models: open the 'bakeoffHMSCvignette' and follow the workflow until you have successfully produced the fits and predictions.

2 Model fitting and predictions

The provided data sets have been divided into training and validation beforehand (see the main text of Norberg et al. for details). The set of 300 sampling units is randomly sampled from the full set and the set of 150 is randomly sampled from this set.

Then the user has to choose whether to fit the Bayesian methods with shorter or longer chains. The default ('FALSE') means shorter chains (max. MCMC 50 000 iterations and/or 24h fitting time), and 'TRUE' mean double the default.

```
MCMC2<-FALSE
sz<-1
#sz<-2
#sz<-3
```

Then we can fit the models and make predictions. Within the model fitting scripts, also computation times are calculated and they are saved in the results directory.

```
for (d in 1:length(Sets)) {

    set_no <- Sets[d]
    source(readdata)
    source(fitmodels)
    source(makepreds)

}
```

3 Measures for evaluating the predictive performance of the models

First we calculate species-specific occurrence probabilities, species richnesses and beta index values. Then we produce the measures for evaluating the predictive performance of the modelling frameworks and their variants, and compile all the results into one table. Finally we'll plot the raw results, also for computation times.

```
ENS<-list(NULL,unlist(mod_names),c("HMSC3", "GLM5", "MISTN1", "GNN1", "MARS1"))
PRV<-list(NA, 0.1)

source(modpreds)
source(pms)
source(pms_comb)
source(pms_tbl)
source(pms_plot)
source(compt_times)
```