Random Forest

Rationale:

To predict the likelihood of a distribution of a characteristic or feature over an area where some sample data is available but not so many individual samples to cover the whole area. It uses a method called “Random Forest” to create a prediction for data. Takes a file of groundtruth data and predicts presence data. If the groundtruth data has just presence data but no absence data the program will make an equal number of randomly placed absence points in the geographic area defined by the first raster image. The groundtruth data will have a mixture of groundtruth points with “Presence” or “Absence” criteria. The user can enter the definition of the presence points and all other points will be counted as absence.

The program will create a prediction raster image at a given resolution of the likelihood of the pixel being definitely defined as Presence (value 1) or Absence (value 0) and many values in between. Accuracy assessment can be determined by statistical values.

Values provided are:

- Errors of commission - sometimes also called "false positives."
- Errors of omission - a mistake that consists of not including something such as an amount or fact that should be included. Errors of omission are likely to be more common than errors of commission.
- Producer's Accuracy - the map accuracy from the point of view of the map maker (the producer). This is how often real features on the ground that are correctly shown on the classified map or the probability that a certain land cover of an area on the ground is classified correctly.
- The User's Accuracy - the accuracy from the point of view of a map user, not the map maker. The User's accuracy essentially tells us how often the class on the map will actually be present on the ground. This is often referred to as reliability.
- Overall Accuracy - the proportion of reference sites were mapped correctly. It is usually expressed as a percent, with 100% accuracy being a perfect classification where all reference sites were classified correctly.
- Cohen's Kappa - a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). It considers random chance (0%) i.e. agreement with a random classifier, to 100% perfect prediction (with the given data), which generally means it is less misleading than simply using accuracy as a metric.

Usage:

There are several parameters required for this analysis:

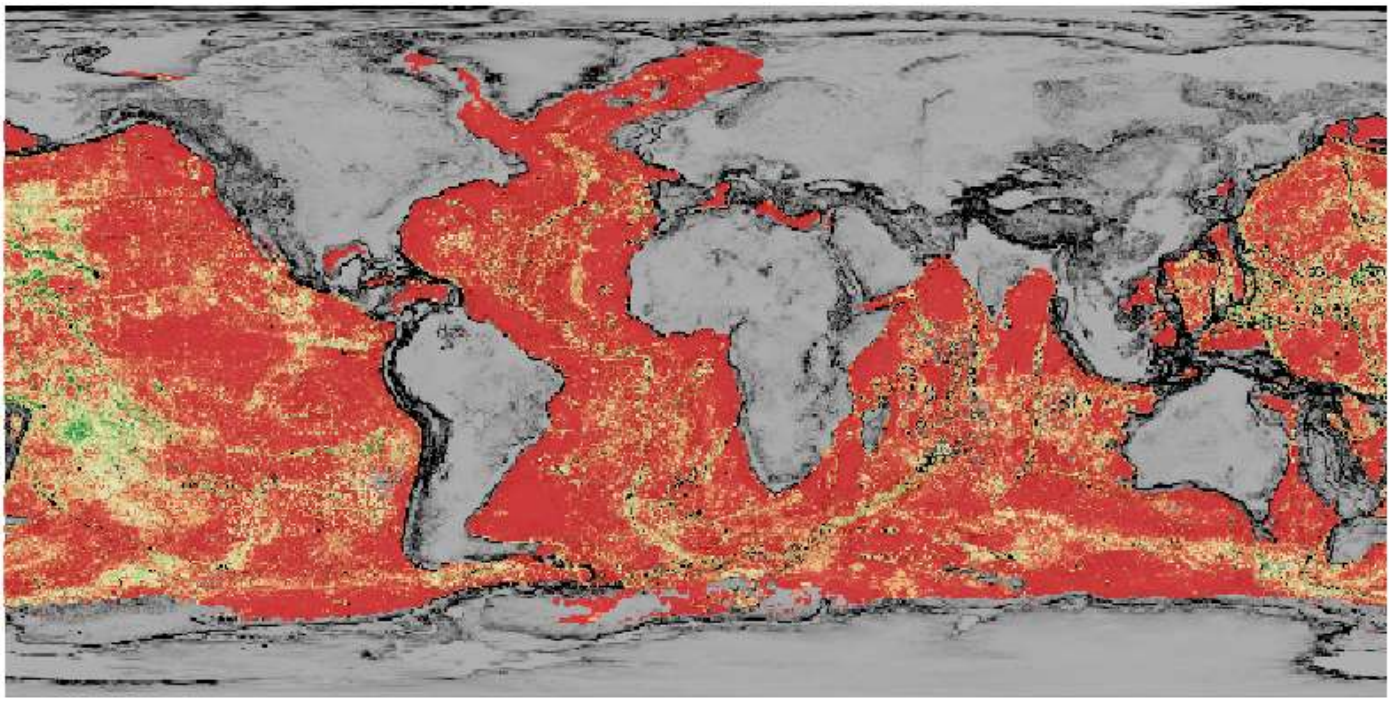
- The groundtruth point sample data giving the feature to be predicted. The point data must have an attribute field that allows determination of presence or absence of the feature. Examples of this are: attribute “Class” has values of “Crust” or “Not Crust”, or attribute “Coral” has values greater or less than 50%. The projection of the input sample data should be the same as the input predictor grids.

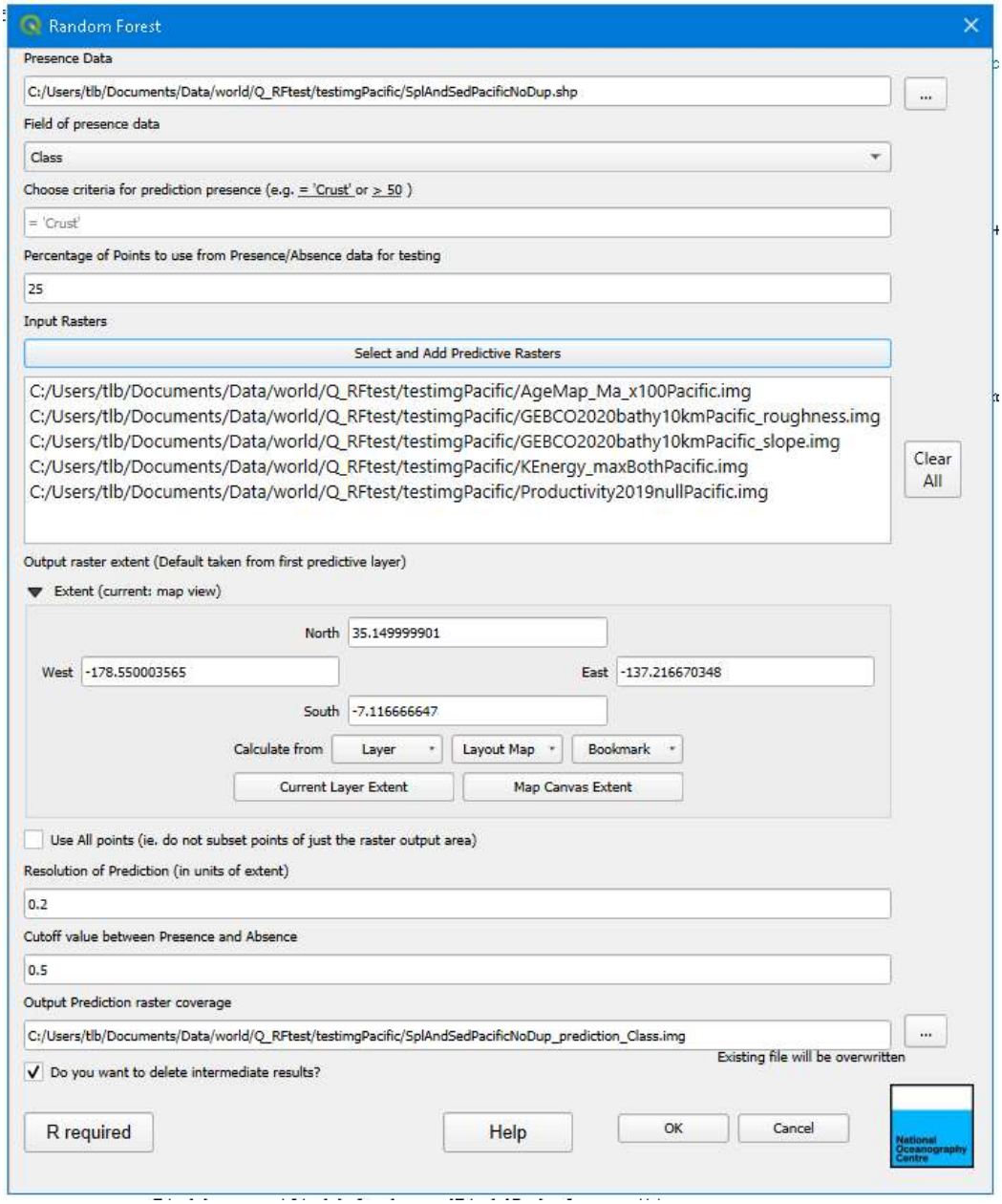
- A pulldown menu of the attributes in the Presence (and Absence) datafile.
- The criteria for prediction presence. This can either be a text criterion (using single quotes such as **= 'Crust'**) or a numeric criterion (such as **> 49.9**)
- A proportion of the input point sample data is held back from the model production and prediction for testing of the final output. A random selection is made of the input sample data being a percentage of the total number of input points.
- Next is the predictor grid files required to do the prediction. Initially, it is suggested that many predictors are entered, but in later predictions, some of those input layers can be omitted due to not be of much use in the final predicted output. All the predictors should be single layered grids in .img or .tif format.
- The extent box can be used to reduce the area of the final prediction grid. This is used to speed up the process or for testing the data. Default extent is the extent of the first predictor grid entered. Output will not be calculated for a pixel if there is a missing value in any of the input predictive grid.
- If all input points are used in the prediction points outside the extent box will used in the production of the prediction model and subsequently affect the output prediction.
- The resolution of the final grid can be in metres or degrees but should be in the same units as the input predictor grids.
- The resulting output grid has values between 1 and 0 for presence and absence. Intermediate values give a level of likelihood. For the statistical calculation of binary presence and absence a cutoff value of 0.5 is often used, but can be nuanced by biasing the resulting division.
- Output is a single polygon vector shapefile and its default filename is the same as the first input sample point filename with "_prediction_<PresenceField>" added to the name. This is default but can be edited by the user.
- The software will create many intermediate files and will put these is a temporary directory /tempMT. Most of these can be deleted by the software but some cannot be removed until exiting QGIS
-

The prediction output grid is accompanied by three other files:

- "RF_Model.Rdata" has the model information used in the prediction
- "RF_Model_importance.png" has a graph of the importance of each of the variable in the model production
- "RF_Model_summary.txt"" has the model statistics. For example:

```
MODEL SUMMARY:
==============

Call:
 randomForest(formula = Presence2 ~ AgeMap_M00 + KEnergy_01 +      Producti02, ntree =
500s, replace = FALSE, importance = TRUE)

               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.326
                    % Var explained: 45.6

Estimated predictor variable importance:

            %IncMSE IncNodePurity
AgeMap_M00    0.312           151
KEnergy_01    0.267           152
Producti02    0.198           113

Presence Absence File used = SplAndSedPacificNoDup.shp
Working in directory       = C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific
Using presence identity      "Class" = 'Crust'
Number of samples used     = 1739
A random selection of      = 1304 samples for model creation
The remainder of samples   = 435 for model testing (25.0 %)
Input files used:
                           = AgeMap_Ma_x100Pacific.img  AgeMap_M00
                           = KEnergy_maxBothPacific.img  KEnergy_01
                           = Productivity2019nullPacific.img  Producti02
Area extent predicted      = -178.550003565,-137.216670348,-7.116666647,35.149999901
Output resolution of grid  = 0.2

Confusion Matrix (based on 50% value presence absence)
                      Presence  Absence
  Predicted Presence      43       17
  Predicted Absence       38      333

                      Errors of    Errors of    Producer     User
                      Commission   Omission     Accuracy     Accuracy
  Predicted Presence    0.28         0.46         0.53         0.71
  Predicted Absence     0.10         0.04         0.95         0.89

  Overall Accuracy       87.238 %
  Cohen's Kappa          53.566 %
```




Random Forest
Presence Data
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/SplAndSedPacificNoDup.shp
...
Field of presence data
Class
Choose criteria for prediction presence (e.g. = 'Crust' or > 50)
= 'Crust'
Percentage of Points to use from Presence/Absence data for testing
25
Input Rasters
Select and Add Predictive Rasters
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/AgeMap_Ma_x100Pacific.img
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/GEBCO2020bathy10kmPacific_roughness.img
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/GEBCO2020bathy10kmPacific_slope.img
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/KEnergy_maxBothPacific.img
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/Productivity2019nullPacific.img
Clear All
Output raster extent (Default taken from first predictive layer)
Extent (current: map view)
North 35.149999901
West -178.550003565
East -137.216670348
South -7.116666647
Calculate from
Layer
Layout Map
Bookmark
Current Layer Extent
Map Canvas Extent
Use All points (ie. do not subset points of just the raster output area)
Resolution of Prediction (in units of extent)
0.2
Cutoff value between Presence and Absence
0.5
Output Prediction raster coverage
C:/Users/tlb/Documents/Data/world/Q_RFtest/testimgPacific/SplAndSedPacificNoDup_prediction_Class.img
...
Existing file will be overwritten
Do you want to delete intermediate results?
R required
Help
OK
Cancel
National Oceanography Centre