Practical - Classification and regression tree

Biodiversity modelling

F.G. Blanchet – August 19–23, 2019

Introduction

This practical document presents R code for the section of the course regarding classification and regression trees.

In the present document, we will focus on the distribution of *Acanthis flammea*.

Load R packages

```
library(raster)
```

```
## Loading required package: sp
```

Load the data

```
bird <- readRDS("birdAll.RDS")
climatePresent <- readRDS("climate_Present.RDS")
road <- readRDS("road_Distance.RDS")
```

Extract and organize the data

Let's organize the data based on what we learned from logistic regression.

```
# Species data
sp <- bird$Acanthis.flammea
spDat <- values(sp)

# Climate data
climateAll <- values(climatePresent)
climateDat <- scale(climateAll)
climateDat <- as.data.frame(climateDat)

# Pixels within 50 km of roads
roadDat <- values(road)
road50 <- which(roadDat < 50000)

# Build the raster for the subset region
```

```r
roadSub <- raster(road)
values(roadSub)[road50] <- values(road)[road50]

# Make sure only land pixels are considered
roadSub <- mask(roadSub, climatePresent[[1]])

# Find pixels with values
locPixRoad <- which(!is.na(values(roadSub)))

# For the species
spSub <- spDat[locPixRoad]

# For the climate
climateSub <- climateDat[locPixRoad,]
colnames(climateSub) <- colnames(climateDat)
climateSub <- as.data.frame(climateSub)

# Extract coordinates
xyAll <- coordinates(sp)
xySub <- xyAll[locPixRoad,]

# Combine all the data in one data.frame
explanaAll <- as.data.frame(cbind(climateDat, xyAll))
explanaSub <- explanaAll[locPixRoad,]
```

Build the model

To estimate classification and regression tree models you need to load the R packages `rpart` and `randomForest`

```r
library(rpart)
library(randomForest)
```

Classification and regression tree models can be very useful to account for the non-linearity in the response variable, but in a different ways than for generalized additive models.

When we build our tree model, we will include climate, distance to major roads and spatial coordinates in the data.

Using `rpart`

Again for our species *Acanthis flammea*, lets build a tree model

```r
# Formula
formu <- as.formula(paste0("spSub ~ ",
                    paste0("bio",1:19,collapse = "+"),
                    "+ x + y"))
```

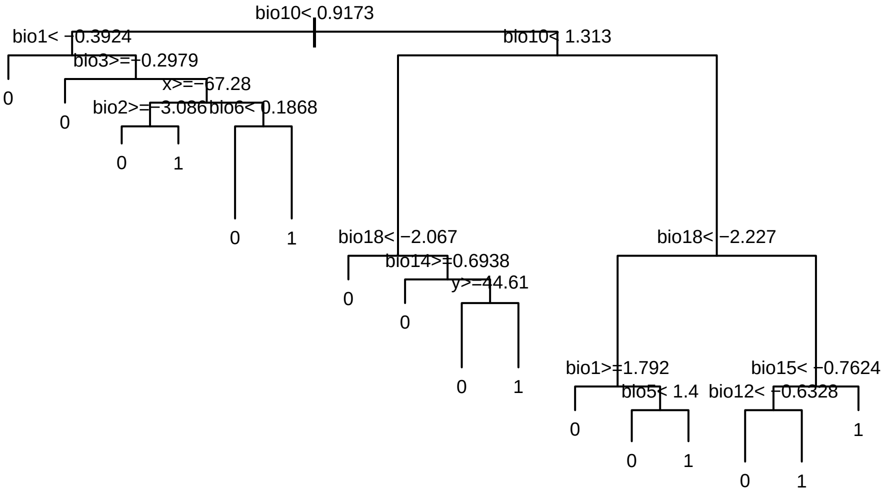
```r
# Model
spRpart <- rpart(formu,
                     data = explanaSub,
                     method = "class",
                     control = list(minsplit = 10,
                                    minbucket = 3,
                                    cp = 0.005))
```

Results

```r
spRpart
```

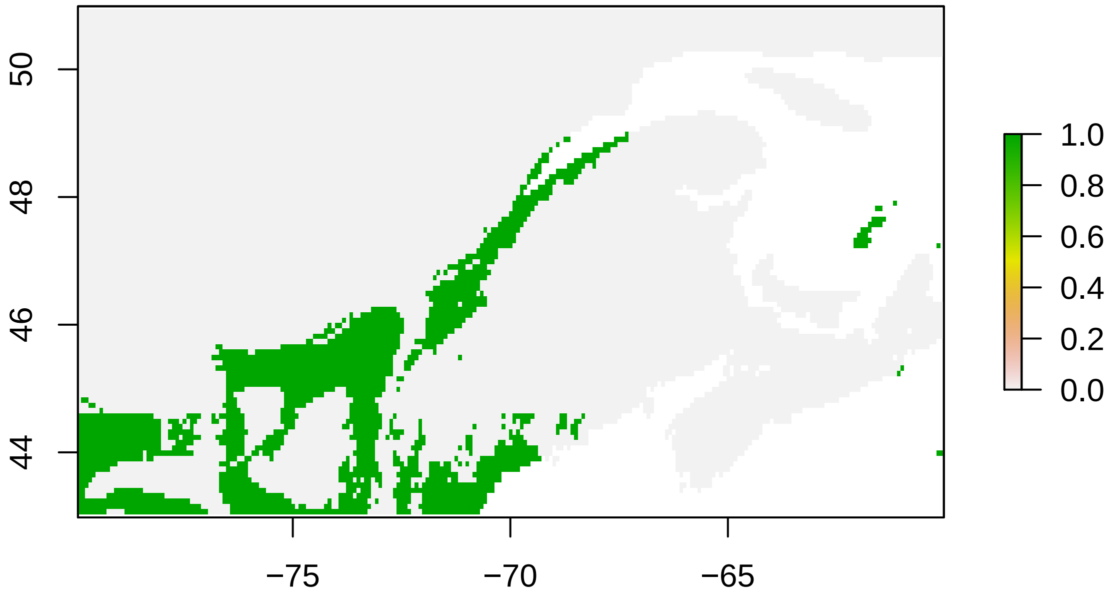
```
## n=13816 (489 observations deleted due to missingness)
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 13816 3133 0 (0.77323393 0.22676607)
##    2) bio10< 0.9172968 10818 1639 0 (0.84849325 0.15150675)
##      4) bio1< -0.3923933 4262  211 0 (0.95049273 0.04950727) *
##      5) bio1>=-0.3923933 6556 1428 0 (0.78218426 0.21781574)
##       10) bio3>=-0.2978698 5094  895 0 (0.82430310 0.17569690) *
##       11) bio3< -0.2978698 1462  533 0 (0.63543092 0.36456908)
##         22) x>=-67.28071 431   83 0 (0.80742459 0.19257541)
##           44) bio2>=-3.08584 415   67 0 (0.83855422 0.16144578) *
##           45) bio2< -3.08584 16    0 1 (0.00000000 1.00000000) *
##         23) x< -67.28071 1031  450 0 (0.56353055 0.43646945)
##           46) bio6< 0.1868088 701  242 0 (0.65477889 0.34522111) *
##           47) bio6>=0.1868088 330  122 1 (0.36969697 0.63030303) *
##    3) bio10>=0.9172968 2998 1494 0 (0.50166778 0.49833222)
##      6) bio10< 1.312863 1073  438 0 (0.59179870 0.40820130)
##       12) bio18< -2.066597 63    7 0 (0.88888889 0.11111111) *
##       13) bio18>=-2.066597 1010  431 0 (0.57326733 0.42673267)
##         26) bio14>=0.6938034 227   70 0 (0.69162996 0.30837004) *
##         27) bio14< 0.6938034 783  361 0 (0.53895275 0.46104725)
##           54) y>=44.61093 547  213 0 (0.61060329 0.38939671) *
##           55) y< 44.61093 236   88 1 (0.37288136 0.62711864) *
##      7) bio10>=1.312863 1925  869 1 (0.45142857 0.54857143)
##       14) bio18< -2.226872 420  149 0 (0.64523810 0.35476190)
##         28) bio1>=1.792219 173   23 0 (0.86705202 0.13294798) *
##         29) bio1< 1.792219 247  121 1 (0.48987854 0.51012146)
##           58) bio5< 1.400469 117   44 0 (0.62393162 0.37606838) *
##           59) bio5>=1.400469 130   48 1 (0.36923077 0.63076923) *
##       15) bio18>=-2.226872 1505  598 1 (0.39734219 0.60265781)
##         30) bio15< -0.7623913 491  240 0 (0.51120163 0.48879837)
##           60) bio12< -0.6327784 169   55 0 (0.67455621 0.32544379) *
##          61) bio12>=-0.6327784 322  137 1 (0.42546584 0.57453416) *
##        31) bio15>=-0.7623913 1014  347 1 (0.34220907 0.65779093) *
```

```
plot(spRpart, margin = 0.05)
text(spRpart, cex = 0.6)
```



Project the estimation on a map

```
# Focus only on the land pixels
locPixLand <- !is.na(climateDat[,1])
explanaLand <- explanaAll[locPixLand,]

# Calculate estimated values
spRpartPred <- predict(spRpart,
                       newdata = explanaLand,
                       type = "class")

# Build raster
spRpartRaster <- raster(climatePresent)

# Convert result into numeric for raster
spRpartPredNum <- as.numeric(as.character(spRpartPred))

# Find where to place the values in the raster
values(spRpartRaster)[locPixLand] <- spRpartPredNum

# Map
plot(spRpartRaster, zlim = c(0,1))
```



Using `randomForest`

Again for our species *Acanthis flammea*, lets build a randomForest model.

For the random forest to be carried out on class data, the response variable needs to be defined as a factor.

```
spFac <- as.factor(spSub)

spRF <- randomForest(spFac ~ .,
                         data = explanaSub,
                         ntree = 50,
                         na.action = na.omit,
                         importance = TRUE)
```

Results

```
spRF
```

```
##
## Call:
##  randomForest(formula = spFac ~ ., data = explanaSub, ntree = 50,      importance = TRUE, na
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 17.31%
## Confusion matrix:
##      0    1 class.error
## 0 9766  917  0.08583731
## 1 1474 1659  0.47047558
```

Extract some useful values

`importance`

This is the importance of each explanatory variable with regards to the levels of the class as well as in the context of a information criterion

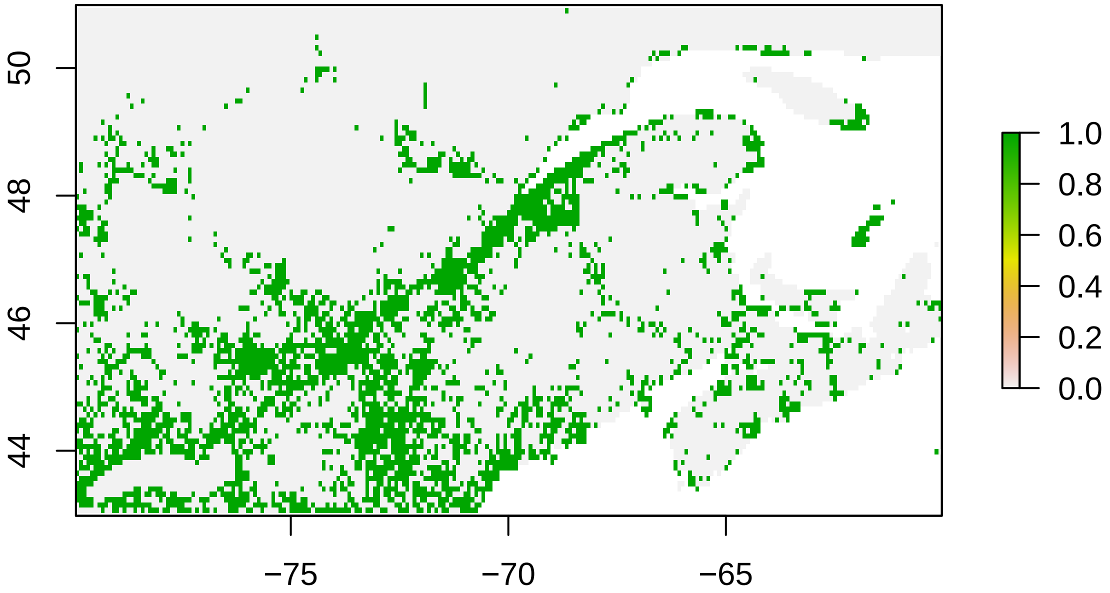
```
importance(spRF)
```

```
##               0        1 MeanDecreaseAccuracy MeanDecreaseGini
## bio1  10.670389 2.139118            11.700282         315.2768
## bio2  15.586344 3.921784            18.244636         221.2158
## bio3  16.754327 3.942050            17.978401         258.3153
## bio4   8.953240 2.218066             9.623560         210.4148
## bio5  11.637268 5.151114            13.274206         407.0826
## bio6   8.525704 2.157426             8.681344         209.6116
## bio7   9.993753 4.090502            11.298268         194.2992
## bio8  14.421621 3.737660            15.379653         224.3772
## bio9   9.255493 1.583365             9.702283         232.9379
## bio10 14.863271 3.573442            15.263553         402.2346
## bio11  8.588179 1.540778             9.493868         206.2739
## bio12  8.972310 2.326418            10.699409         172.9879
## bio13 14.781573 3.577625            16.432217         186.2600
## bio14 13.193019 2.381943            15.428812         203.3266
## bio15 13.932032 3.874415            15.312562         216.3068
## bio16 13.523474 2.699972            14.020872         183.6759
## bio17 10.503015 2.023702            11.584168         190.9608
## bio18 12.743575 3.060072            13.459958         219.0006
## bio19 11.131091 1.398199            12.619758         188.5559
## x     14.774508 5.232327            16.354807         241.1186
## y      8.731528 2.975578             9.899900         153.0197
```

Project the estimation on a map

```
# Focus only on the land pixels
locPixLand <- !is.na(climateDat[,1])
explanaLand <- explanaAll[locPixLand,]

# Calculate estimated values
spRFPred <- predict(spRF,
                        newdata = explanaLand,
                        type = "class")

# Build raster
```

```
spRFRaster <- raster(climatePresent)

# Convert result into numeric for raster
spRFPredNum <- as.numeric(as.character(spRFPred))

# Find where to place the values in the raster
values(spRFRaster)[locPixLand] <- spRFPredNum

# Map
plot(spRFRaster, zlim = c(0,1))
```



A few things to try

- For the CART model tweak the parameters to get a more relevent model (the ones chosen were chosen partly for illustration purposes)
- Model the distribution for another bird species
- Model the distribution for all bird species
- Model the distribution for a proposed climatic scenario
- Compare the result of a random forest to a CART model, a GAM and/or a GLM for the same species

To keep in mind

- This is not only model fitting, have an ecological perspective
- The data has particularities, be aware of it
- Remember the assumptions that you make when building your model
- Remember that the reason why you are modelling a species may skew the way you build and interpret the model.