Machine Learning and Data Mining project: Leaf identification.

Sirianne MADON KENGNE

2022-12-16

Topic:

The goal is to propose a method for leaf identification based on the provided leaf attributes and using a proper unsupervised or supervised learning tool.

```
library(ggplot2)
library(cowplot)
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

Let's read the data and see the first 6 rows.

```
data <- read.csv("leaf.csv", header=FALSE)
head(data)
```

```
##   V1 V2      V3     V4      V5      V6      V7      V8        V9       V10
## 1  1  1 0.72694 1.4742 0.32396 0.98535 1.00000 0.83592 0.0046566 0.0039465
## 2  1  2 0.74173 1.5257 0.36116 0.98152 0.99825 0.79867 0.0052423 0.0050016
## 3  1  3 0.76722 1.5725 0.38998 0.97755 1.00000 0.80812 0.0074573 0.0101210
## 4  1  4 0.73797 1.4597 0.35376 0.97566 1.00000 0.81697 0.0068768 0.0086068
## 5  1  5 0.82301 1.7707 0.44462 0.97698 1.00000 0.75493 0.0074280 0.0100420
## 6  1  6 0.72997 1.4892 0.34284 0.98755 1.00000 0.84482 0.0049451 0.0044506
##         V11      V12       V13        V14        V15     V16
## 1 0.0477900 0.127950 0.0161080 0.00523230 2.7477e-04 1.17560
## 2 0.0241600 0.090476 0.0081195 0.00270800 7.4846e-05 0.69659
## 3 0.0118970 0.057445 0.0032891 0.00092068 3.7886e-05 0.44348
## 4 0.0159500 0.065491 0.0042707 0.00115440 6.6272e-05 0.58785
## 5 0.0079379 0.045339 0.0020514 0.00055986 2.3504e-05 0.34214
## 6 0.0104870 0.058528 0.0034138 0.00112480 2.4798e-05 0.34068
```

Let's give proper columns names to our data

```
colnames(data) <- c("Species","Specimen Number","Eccentricity","AspectRatio","Elongation","Solidity","S
names(data)
```

```
##  [1] "Species"                   "Specimen Number"
##  [3] "Eccentricity"              "AspectRatio"
##  [5] "Elongation"                "Solidity"
##  [7] "Stochastic_Convexity"      "Isoperimetric_Factor"
##  [9] "Maximal_Indentation_Depth" "Lobedness"
## [11] "Average_Intensity"         "Average_Contrast"
## [13] "Smoothness"                "Third_moment"
## [15] "Uniformity"                "Entropy"
```

Let's give the proper name of species

```
last_species_names<-c(1:15,22:36)
new_species_names<-c("Quercus suber","Salix atrocinera","Populus nigra","Alnus sp.","Quercus robur",
     "Crataegus monogyna","Ilex aquifolium","Nerium oleander","Betula pubescens",
     "Tilia tomentosa","Acer palmatum","Celtis sp.","Corylus avellana","Castanea sativa","Populus alba"
     "Primula vulgaris","Erodium sp.","Bougainvillea sp.","Arisarum vulgare","Euonymus japonicus","Ilex
     "Magnolia soulangeana","Buxus sempervirens","Urtica dioica","Podocarpus sp.","Acca sellowiana","Hy
     "Magnolia grandiflora","Geranium sp.")
for(i in last_species_names){
  data[data$Species == i,]$Species <-new_species_names[i]
}
unique(data$Species)
```

```
##  [1] "Quercus suber"            "Salix atrocinera"
##  [3] "Populus nigra"            "Alnus sp."
##  [5] "Quercus robur"            "Crataegus monogyna"
##  [7] "Ilex aquifolium"          "Nerium oleander"
##  [9] "Betula pubescens"         "Tilia tomentosa"
## [11] "Acer palmatum"            "Celtis sp."
## [13] "Corylus avellana"         "Castanea sativa"
## [15] "Populus alba"             "Primula vulgaris"
## [17] "Erodium sp."              "Bougainvillea sp."
## [19] "Arisarum vulgare"         "Euonymus japonicus"
## [21] "Ilex perado ssp. azorica" "Magnolia soulangeana"
## [23] "Buxus sempervirens"       "Urtica dioica"
## [25] "Podocarpus sp."           "Acca sellowiana"
## [27] "Hydrangea sp."            "Pseudosasa japonica"
## [29] "Magnolia grandiflora"     "Geranium sp."
```

let's delete the columns 2 because it is useless for what we want to do

```
dat<-data[,-2]
head(dat)
```

```
##         Species Eccentricity AspectRatio Elongation Solidity
## 1 Quercus suber      0.72694      1.4742    0.32396  0.98535
## 2 Quercus suber      0.74173      1.5257    0.36116  0.98152
## 3 Quercus suber      0.76722      1.5725    0.38998  0.97755
## 4 Quercus suber      0.73797      1.4597    0.35376  0.97566
## 5 Quercus suber      0.82301      1.7707    0.44462  0.97698
## 6 Quercus suber      0.72997      1.4892    0.34284  0.98755
##   Stochastic_Convexity Isoperimetric_Factor Maximal_Indentation_Depth Lobedness
## 1              1.00000              0.83592                 0.0046566 0.0039465
## 2              0.99825              0.79867                 0.0052423 0.0050016
## 3              1.00000              0.80812                 0.0074573 0.0101210
## 4              1.00000              0.81697                 0.0068768 0.0086068
## 5              1.00000              0.75493                 0.0074280 0.0100420
## 6              1.00000              0.84482                 0.0049451 0.0044506
##   Average_Intensity Average_Contrast Smoothness Third_moment Uniformity Entropy
## 1         0.0477900         0.127950  0.0161080   0.00523230 2.7477e-04 1.17560
## 2         0.0241600         0.090476  0.0081195   0.00270800 7.4846e-05 0.69659
## 3         0.0118970         0.057445  0.0032891   0.00092068 3.7886e-05 0.44348
## 4         0.0159500         0.065491  0.0042707   0.00115440 6.6272e-05 0.58785
## 5         0.0079379         0.045339  0.0020514   0.00055986 2.3504e-05 0.34214
## 6         0.0104870         0.058528  0.0034138   0.00112480 2.4798e-05 0.34068
```

```
dat$Species <- as.factor(dat$Species)
str(dat)
```

```
## 'data.frame':    340 obs. of  15 variables:
##  $ Species                 : Factor w/ 30 levels "Acca sellowiana",..: 27 27 27 27 27 27 27 27 27 27
##  $ Eccentricity            : num  0.727 0.742 0.767 0.738 0.823 ...
##  $ AspectRatio             : num  1.47 1.53 1.57 1.46 1.77 ...
##  $ Elongation              : num  0.324 0.361 0.39 0.354 0.445 ...
##  $ Solidity                : num  0.985 0.982 0.978 0.976 0.977 ...
##  $ Stochastic_Convexity    : num  1 0.998 1 1 1 ...
##  $ Isoperimetric_Factor    : num  0.836 0.799 0.808 0.817 0.755 ...
##  $ Maximal_Indentation_Depth: num  0.00466 0.00524 0.00746 0.00688 0.00743 ...
##  $ Lobedness               : num  0.00395 0.005 0.01012 0.00861 0.01004 ...
##  $ Average_Intensity       : num  0.04779 0.02416 0.0119 0.01595 0.00794 ...
##  $ Average_Contrast        : num  0.128 0.0905 0.0574 0.0655 0.0453 ...
##  $ Smoothness              : num  0.01611 0.00812 0.00329 0.00427 0.00205 ...
##  $ Third_moment            : num  0.005232 0.002708 0.000921 0.001154 0.00056 ...
##  $ Uniformity              : num  2.75e-04 7.48e-05 3.79e-05 6.63e-05 2.35e-05 ...
##  $ Entropy                 : num  1.176 0.697 0.443 0.588 0.342 ...
```

Let's run the random forest

```
#write.csv(dat, "mydata.csv")
set.seed(42)
model <- randomForest(Species ~ ., data=dat, proximity=TRUE)
str(model)
```

```
## List of 19
##  $ call           : language randomForest(formula = Species ~ ., data = dat, proximity = TRUE)
##  $ type           : chr "classification"
##  $ predicted      : Factor w/ 30 levels "Acca sellowiana",..: 10 27 27 27 27 27 27 27 27 7 ...
##   ..- attr(*, "names")= chr [1:340] "1" "2" "3" "4" ...
##  $ err.rate       : num [1:500, 1:31] 0.492 0.498 0.476 0.464 0.454 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:31] "OOB" "Acca sellowiana" "Acer palmatum" "Alnus sp." ...
##  $ confusion      : num [1:30, 1:31] 3 0 0 0 0 0 0 1 0 0 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:30] "Acca sellowiana" "Acer palmatum" "Alnus sp." "Arisarum vulgare" ...
##   .. ..$ : chr [1:31] "Acca sellowiana" "Acer palmatum" "Alnus sp." "Arisarum vulgare" ...
##  $ votes          : 'matrix' num [1:340, 1:30] 0.0117 0.0649 0.0581 0.0435 0.0719 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:340] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:30] "Acca sellowiana" "Acer palmatum" "Alnus sp." "Arisarum vulgare" ...
##  $ oob.times      : num [1:340] 171 154 172 184 167 191 190 178 171 182 ...
##  $ classes        : chr [1:30] "Acca sellowiana" "Acer palmatum" "Alnus sp." "Arisarum vulgare" ...
##  $ importance     : num [1:14, 1] 27.4 29.3 29.9 38.9 16.6 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:14] "Eccentricity" "AspectRatio" "Elongation" "Solidity" ...
##   .. ..$ : chr "MeanDecreaseGini"
##  $ importanceSD   : NULL
##  $ localImportance: NULL
##  $ proximity      : num [1:340, 1:340] 1 0 0 0.02 0 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:340] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:340] "1" "2" "3" "4" ...
##  $ ntree          : num 500
##  $ mtry           : num 3
##  $ forest         :List of 14
##   ..$ ndbigtree : int [1:500] 165 167 159 161 157 165 135 159 163 163 ...
##   ..$ nodestatus: int [1:199, 1:500] 1 -1 1 1 1 1 1 1 1 1 ...
##   ..$ bestvar   : int [1:199, 1:500] 4 0 10 1 2 6 14 14 7 3 ...
##   ..$ treemap   : int [1:199, 1:2, 1:500] 2 0 4 6 8 10 12 14 16 18 ...
##   ..$ nodepred  : int [1:199, 1:500] 0 2 0 0 0 0 0 0 0 0 ...
##   ..$ xbestsplit: num [1:199, 1:500] 0.6182 0 0.0707 0.9507 1.6593 ...
##   ..$ pid       : num [1:30] 1 1 1 1 1 1 1 1 1 1 ...
##   ..$ cutoff    : num [1:30] 0.0333 0.0333 0.0333 0.0333 0.0333 ...
##   ..$ ncat      : Named int [1:14] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..- attr(*, "names")= chr [1:14] "Eccentricity" "AspectRatio" "Elongation" "Solidity" ...
##   ..$ maxcat    : int 1
##   ..$ nrnodes   : int 199
##   ..$ ntree     : num 500
##   ..$ nclass    : int 30
##   ..$ xlevels   :List of 14
##   .. ..$ Eccentricity             : num 0
##   .. ..$ AspectRatio              : num 0
##   .. ..$ Elongation               : num 0
##   .. ..$ Solidity                 : num 0
##   .. ..$ Stochastic_Convexity     : num 0
##   .. ..$ Isoperimetric_Factor     : num 0
##   .. ..$ Maximal_Indentation_Depth: num 0
##   .. ..$ Lobedness                : num 0
##   .. ..$ Average_Intensity        : num 0
##   .. ..$ Average_Contrast         : num 0
##   .. ..$ Smoothness               : num 0
##   .. ..$ Third_moment             : num 0
##   .. ..$ Uniformity               : num 0
##   .. ..$ Entropy                  : num 0
##  $ y              : Factor w/ 30 levels "Acca sellowiana",..: 27 27 27 27 27 27 27 27 27 27 ...
##   ..- attr(*, "names")= chr [1:340] "1" "2" "3" "4" ...
##  $ test           : NULL
##  $ inbag          : NULL
##  $ terms          :Classes 'terms', 'formula'  language Species ~ Eccentricity + AspectRatio + Elonga
##   .. ..- attr(*, "variables")= language list(Species, Eccentricity, AspectRatio, Elongation, Solidity
##   .. ..- attr(*, "factors")= int [1:15, 1:14] 0 1 0 0 0 0 0 0 0 0 ...
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:15] "Species" "Eccentricity" "AspectRatio" "Elongation" ...
##   .. .. .. ..$ : chr [1:14] "Eccentricity" "AspectRatio" "Elongation" "Solidity" ...
##   .. ..- attr(*, "term.labels")= chr [1:14] "Eccentricity" "AspectRatio" "Elongation" "Solidity" ...
##   .. ..- attr(*, "order")= int [1:14] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..- attr(*, "intercept")= num 0
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(Species, Eccentricity, AspectRatio, Elongation, Solidity
##   .. ..- attr(*, "dataClasses")= Named chr [1:15] "factor" "numeric" "numeric" "numeric" ...
##   .. .. ..- attr(*, "names")= chr [1:15] "Species" "Eccentricity" "AspectRatio" "Elongation" ...
##  - attr(*, "class")= chr [1:2] "randomForest.formula" "randomForest"
```

We know want to plot the out_of_bag rate and the error rate foreach species in function of the numbers of tree in our random forest. We first put our model data in form of 2 columns(tree, error).

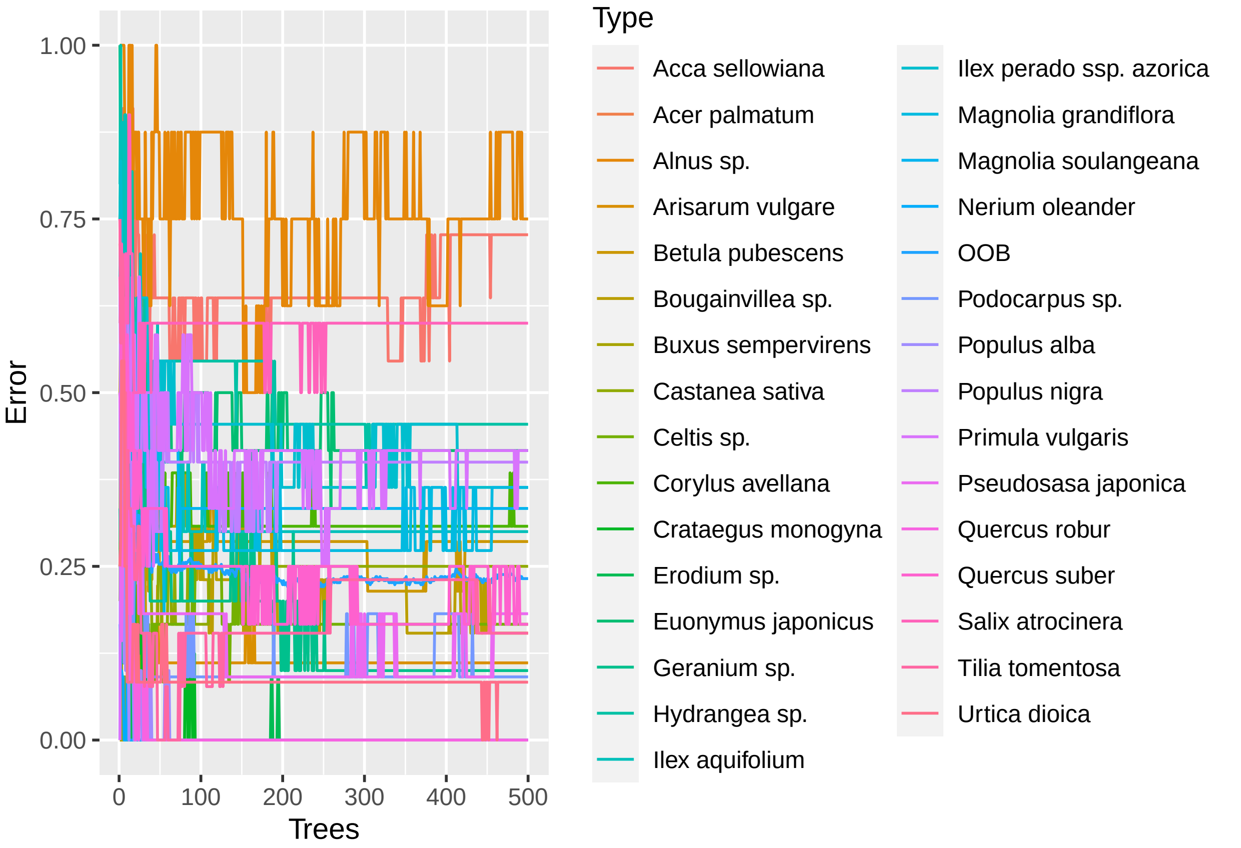
```
clas<-c("Quercus suber" ,"Salix atrocinera","Populus nigra","Alnus sp.","Quercus robur","Crataegus monog

err<-c(model$err.rate[,"OOB"])
for (i in clas) err<-c(err,model$err.rate[,i])

oob.error.data <- data.frame(Trees=rep(1:nrow(model$err.rate), times=31),Type=rep(c("OOB",clas), each=n
str(oob.error.data)
```

```
## 'data.frame':    15500 obs. of  3 variables:
##  $ Trees: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type : chr  "OOB" "OOB" "OOB" "OOB" ...
##  $ Error: num  0.492 0.498 0.476 0.464 0.454 ...
```

```
#head(oob.error.data)
```

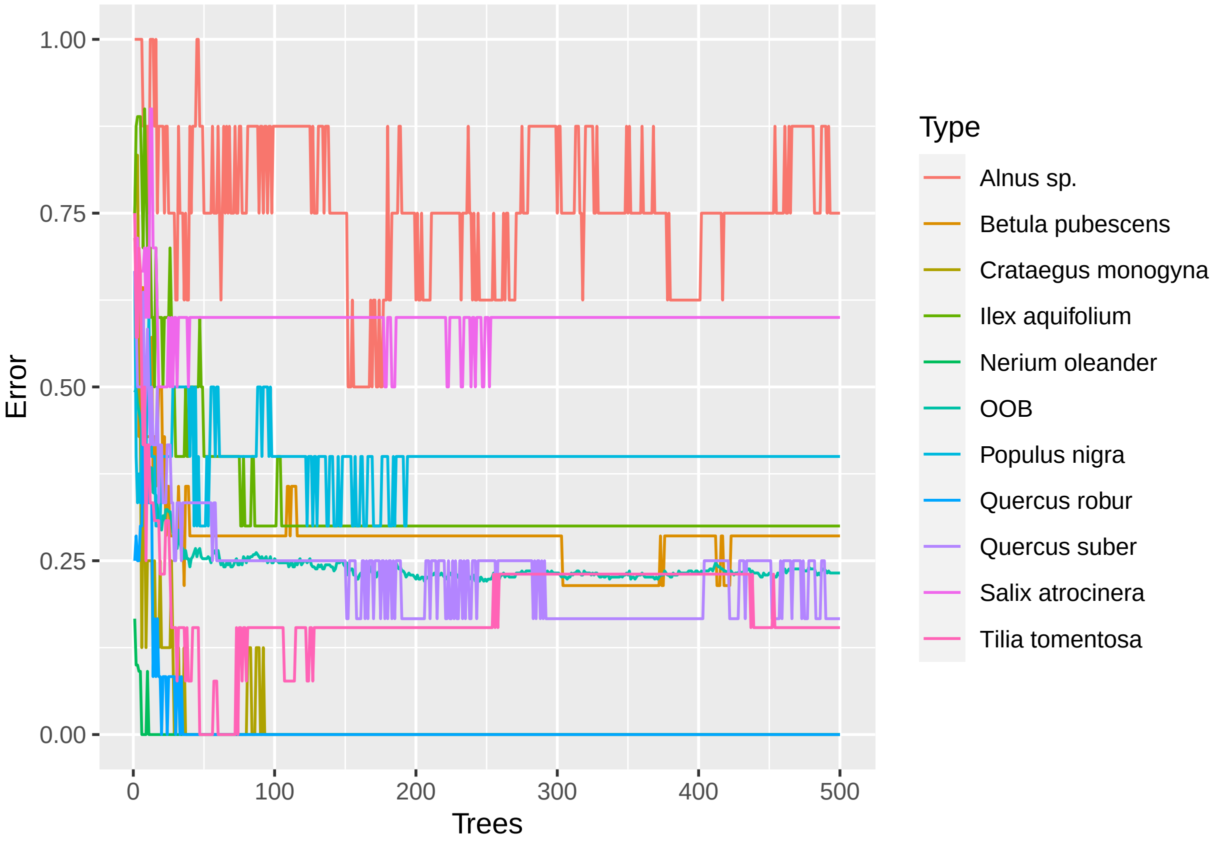
Know we do the plot.

```
ggplot(data=oob.error.data, aes(x=Trees, y=Error)) + geom_line(aes(color=Type))
```

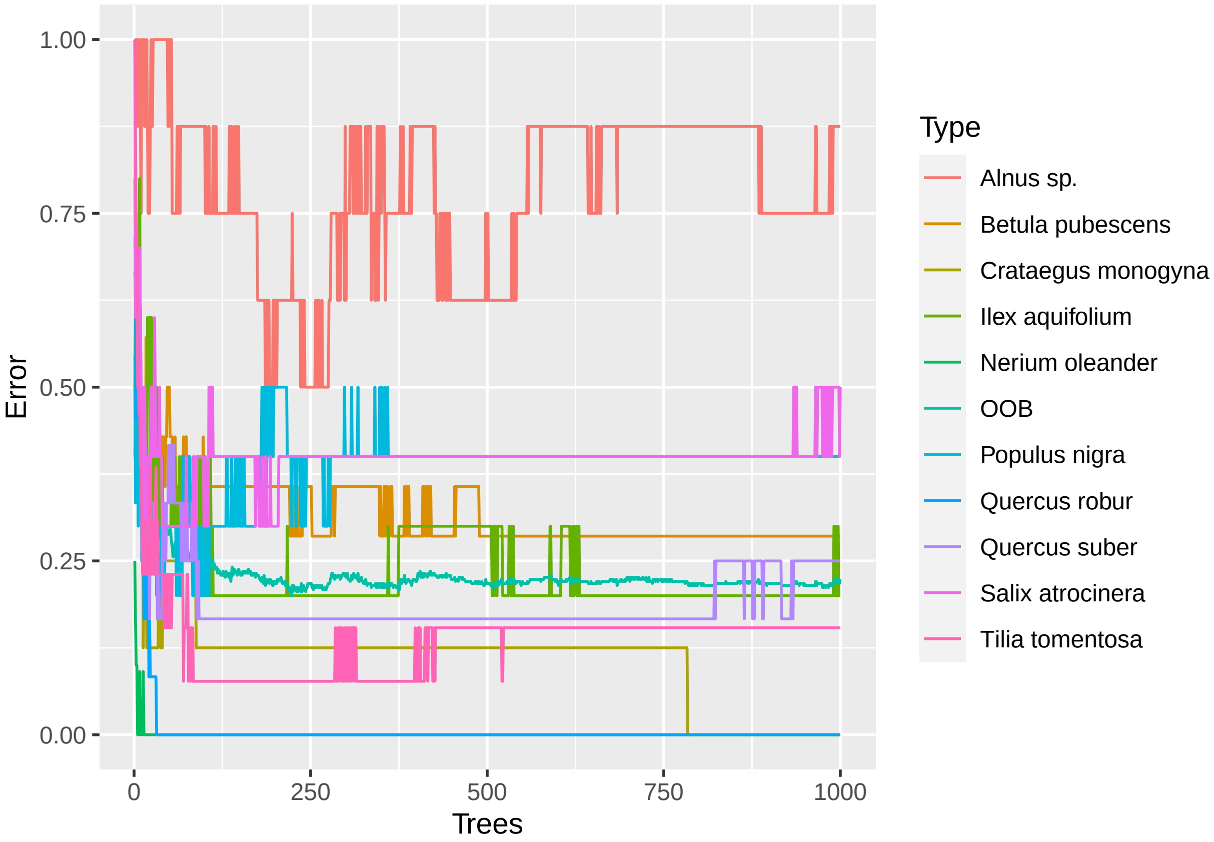


```
# ggsave("oob_error_rate_1000_trees.pdf")
```

```
clas<-c("Quercus suber" ,"Salix atrocinera","Populus nigra","Alnus sp.","Quercus robur","Crataegus mono

err<-c(model$err.rate[,"OOB"])
for (i in clas) err<-c(err,model$err.rate[,i])

oob.error.data <- data.frame(Trees=rep(1:nrow(model$err.rate), times=11),Type=rep(c("OOB",clas), each=n
ggplot(data=oob.error.data, aes(x=Trees, y=Error)) + geom_line(aes(color=Type))
```



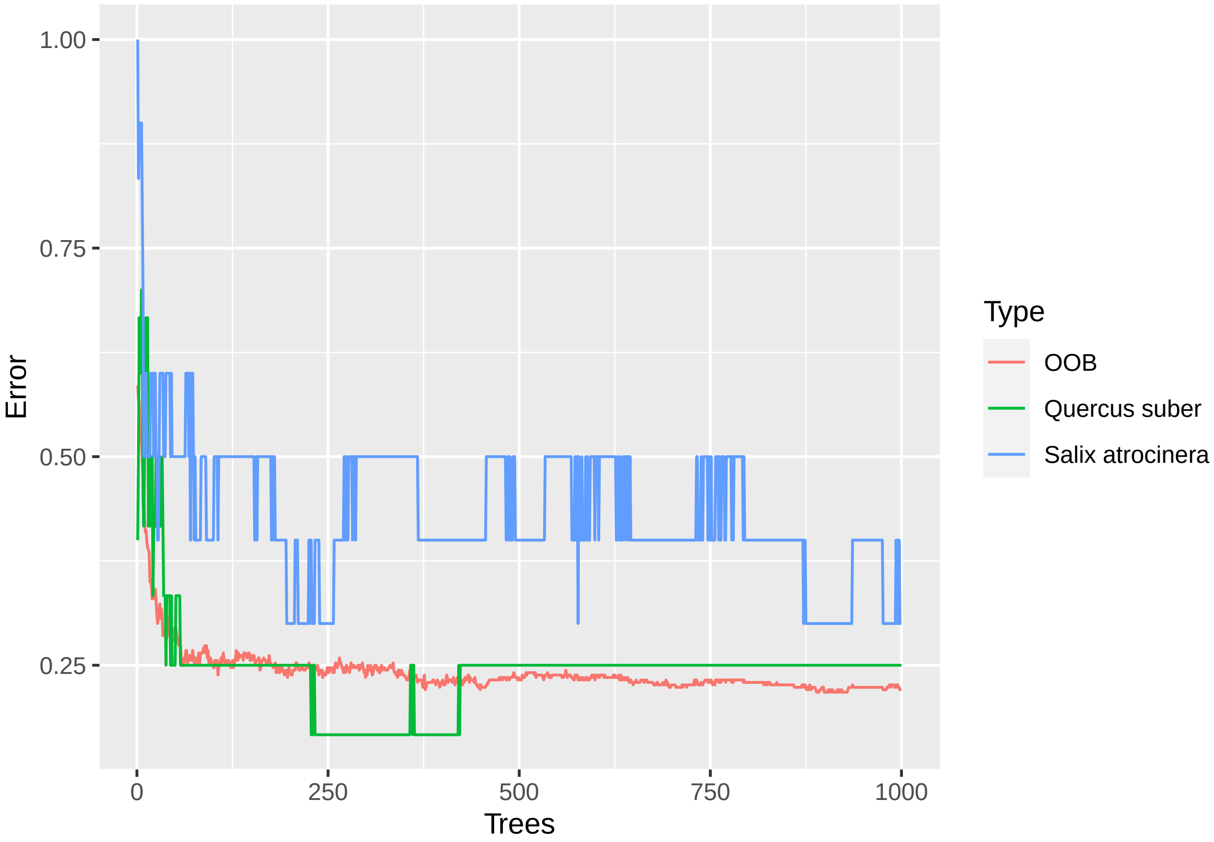
```
model <- randomForest(Species ~ ., data=dat, ntree=1000, proximity=TRUE)

clas<-c("Quercus suber" ,"Salix atrocinera","Populus nigra","Alnus sp.","Quercus robur","Crataegus mono

err<-c(model$err.rate[,"OOB"])
for (i in clas) err<-c(err,model$err.rate[,i])

oob.error.data <- data.frame(Trees=rep(1:nrow(model$err.rate), times=11),Type=rep(c("OOB",clas), each=n
ggplot(data=oob.error.data, aes(x=Trees, y=Error)) + geom_line(aes(color=Type))
```



```
## If we want to compare this random forest to others with different values for
## mtry (to control how many variables are considered at each step)...
oob.values <- vector(length=10)
for(i in 1:10) {
  temp.model <- randomForest(Species ~ ., data=dat, mtry=i, ntree=1000)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
oob.values
```

```
##  [1] 0.2500000 0.2382353 0.2235294 0.2117647 0.2117647 0.2205882 0.2264706
##  [8] 0.2382353 0.2205882 0.2264706
```

```
model <- randomForest(Species ~ ., data=dat,mtry=2, ntree=1000, proximity=TRUE)

clas<-c("Quercus suber" ,"Salix atrocinera")

err<-c(model$err.rate[,"OOB"])
for (i in clas) err<-c(err,model$err.rate[,i])

oob.error.data <- data.frame(Trees=rep(1:nrow(model$err.rate), times=3),Type=rep(c("OOB",clas), each=nr
ggplot(data=oob.error.data, aes(x=Trees, y=Error)) + geom_line(aes(color=Type))
```


1.00
0.75
0.50
0.25
Error
0
250
500
750
1000
Trees
Type
OOB
Quercus suber
Salix atrocinera