Report

Raphael Idewele

22/02/2022

ANALYSIS AND MODELLING OF FOREST DATA

Introduction

At the core of this task is the need to analyze the situation within the datasets and develop models that can be used in predicting the stem volume in the stand in response to a combination of the independent variables and making predictions on the validation data. This work under consideration is made up of two datasets. These are the modelling datasets and validation datasets[1,2]. The features in both datasets are: "LAT – latitude, LONG - longitude, SP_GROUP – dominant species (1: Scots pine, 2: Norway spruce, 4: Birch and other deciduous), AGE – stand age (years), BA – stand basal area (m2/ha), D – average diameter (cm), H – average height (m), FOREST_TYPE – site fertility class (1: very fertile, 2: fertile, 3: semi-fertile, .., 7: very poor), VOLUME – stand volume (m3/ha), P0 - Annual precipitation (mm)"[3]. However, there is a mismatch between the other fields in the datasets. Thus, the need for data cleaning. Although stand ID was reportedly included, however, it is ostensibly missing from the datasets.

Import Datasets

-Model dataset

```
data_trees=read.csv2(file = "C://Users//Raph//Documents//Group1.csv", header
= T, sep = ",")
```

-Validation dataset

```
data_trees_validation=read.csv2(file =
"C://Users//Raph//Documents//validating_data.csv", header = T, sep = ",")
```

Overview of the modelling data revealed it is made up of 150 samples trees with each instances having thirteen fields while validation data shows 1000 trees with twelve features in the dataset. This leaves the modelling data at about 13% of the composite datasets. If it was not that the datasets were predefined and the mode of generation unknown from the data source, I would have opted for a split in the upward direction of 50% in favour of modelling dataset.

```
colnames(data_trees)
```

```
##  [1] "X.1"          "X"            "LAT"          "LONG"
"SP_GROUP"
##  [6] "AGE"          "P0"           "BA"           "D"            "H"
## [11] "YEAR"         "FOREST_TYPE"  "TOTAL_VOLUME"
```

```
colnames(data_trees_validation)
```

```
##  [1] "LAT"         "LONG"        "SP_GROUP"    "AGE"         "P0"
##  [6] "BA"          "D"           "H"           "N"           "YEAR"
## [11] "FOREST_TYPE" "VOLUME"
```

The column N is a distinct feature that is not mentioned in the modelling dataset. Meanwhile, variables X and X1 should not be used because there is no way to validate a model built on them. Also, N should be omitted as it is absent in the model.

Data Cleaning

The dplyr's select function is used to drop the extra and irrelevant fields in the datasets. Volume column was homogenized tin both datasets too.

```
trees<-select(trees,-c(X.1, X))
treesv<-select(treesv,-c(N))
colnames(trees)[which(colnames(trees) %in% c("TOTAL_VOLUME"))] <- c("VOLUME")
```

Summary statistics

The inbuilt summary function would have been used but was forgone for describe function of the psch package because of the ease with which it handles character without conversion to factor. Moreover, describe gives elegant summary table.

```
desct<-describe(trees)
desctv<-describe(treesv)
```

```
subset(desct, select = -c(vars, trimmed,mad,range))
```

```
##                n    mean    sd median  min  max  skew kurtosis   se
## LAT*         150   27.71 15.01   25.5    1   57  0.26    -0.98 1.23
## LONG*        150   38.47 18.48   39.0    1   71 -0.24    -0.85 1.51
## SP_GROUP     150    1.54  0.95    1.0    1    4  1.78     1.97 0.08
## AGE          150   72.01 42.86   62.0   18  269  1.91     4.96 3.50
## P0*          150   70.65 40.87   70.5    1  142  0.03    -1.23 3.34
## BA*          150   74.15 42.69   73.5    1  148  0.02    -1.22 3.49
## D*           150   75.50 43.45   75.5    1  150  0.00    -1.22 3.55
## H*           150   75.50 43.45   75.5    1  150  0.00    -1.22 3.55
## YEAR         150 2010.00  0.00 2010.0 2010 2010   NaN      NaN 0.00
## FOREST_TYPE  150    3.22  0.95    3.0    1    7  0.64     2.32 0.08
## VOLUME*      150   75.50 43.45   75.5    1  150  0.00    -1.22 3.55
```

```
#Summary of validation data

subset(desctv, select = -c(vars, trimmed,mad,range))
```

```
##                n    mean     sd median  min  max  skew kurtosis   se
## LAT*        1000   36.90  18.66   34.0    1   77  0.38    -0.72 0.59
## LONG*       1000   50.97  23.45   51.0    1   96 -0.11    -0.88 0.74
## SP_GROUP    1000    1.49   0.94    1.0    1    4  1.93     2.44 0.03
## AGE         1000   67.57  43.94   58.0   11  337  2.29     7.71 1.39
## P0*         1000  328.90 203.40  322.5    1  683  0.06    -1.27 6.43
## BA*         1000  478.34 283.56  474.5    1  972  0.03    -1.22 8.97
## D*          1000  500.50 288.82  500.5    1 1000  0.00    -1.20 9.13
## H*          1000  500.50 288.82  500.5    1 1000  0.00    -1.20 9.13
## YEAR        1000 2010.00   0.00 2010.0 2010 2010   NaN      NaN 0.00
## FOREST_TYPE 1000    3.33   0.94    3.0    1    7  0.12     1.13 0.03
## VOLUME*     1000  500.50 288.82  500.5    1 1000  0.00    -1.20 9.13
```

From tables above, the age of the youngest and oldest tree is 18 and 269 years, whereas the average age for the trees is 66 years. Also, the shortest and tallest trees are 1m and 150m. Again, the smallest and largest diameters are 1cm and 150cm. The average annual rainfall is 328.90mm.

Data type conversion

Some key variable types required for the model are characters and not numeric, and thus should be adjusted. Meanwhile, BA, H, D, PO, and TOTAL_VOLUME are cast as double for our data, a replace parameter is used to effect the change. For this conversion, within function could be used too.

Data Visualization

It is possible to get information on the datasets intuitively and know how to go about modelling them by simply visualizing them. The ggplot2 extension package, GGally, was used to get scatter plots for the numeric fields to see patterns over the datasets. This is favoured in this report because it is easier to get a bird's-eye view of the datasets without the need to create several individual plots and calculations. The desired features were selected. The modelling and validation datasets showed quite similar pattern (e.g., the most correlated are the same in both datasets). Moreover, the interaction among the independent variables and that with the response variable (stand volume) is conspicuous looking at the chart. Some important observations are:

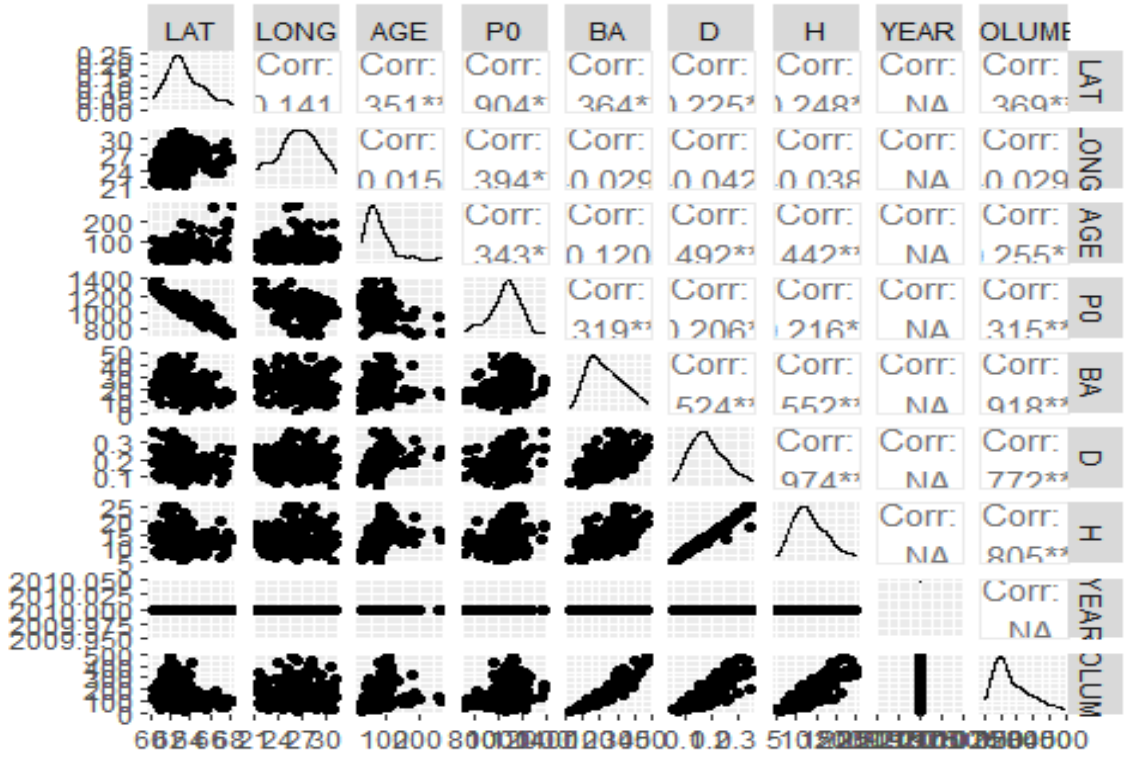
1. The sample datasets are for a single year (2010). Thus, the flat nature of year plots. Using unique function on the year field of the datasets confirms this.

2. Based on the relationship of the response variable to the independent predictors in decreasing order of correlation: BA, H, D are the best whereas others do not show meaningful correlation. Interestingly, longitude is negatively correlated for the modelling dataset. Again, this is probably not significant enough to influence the data.

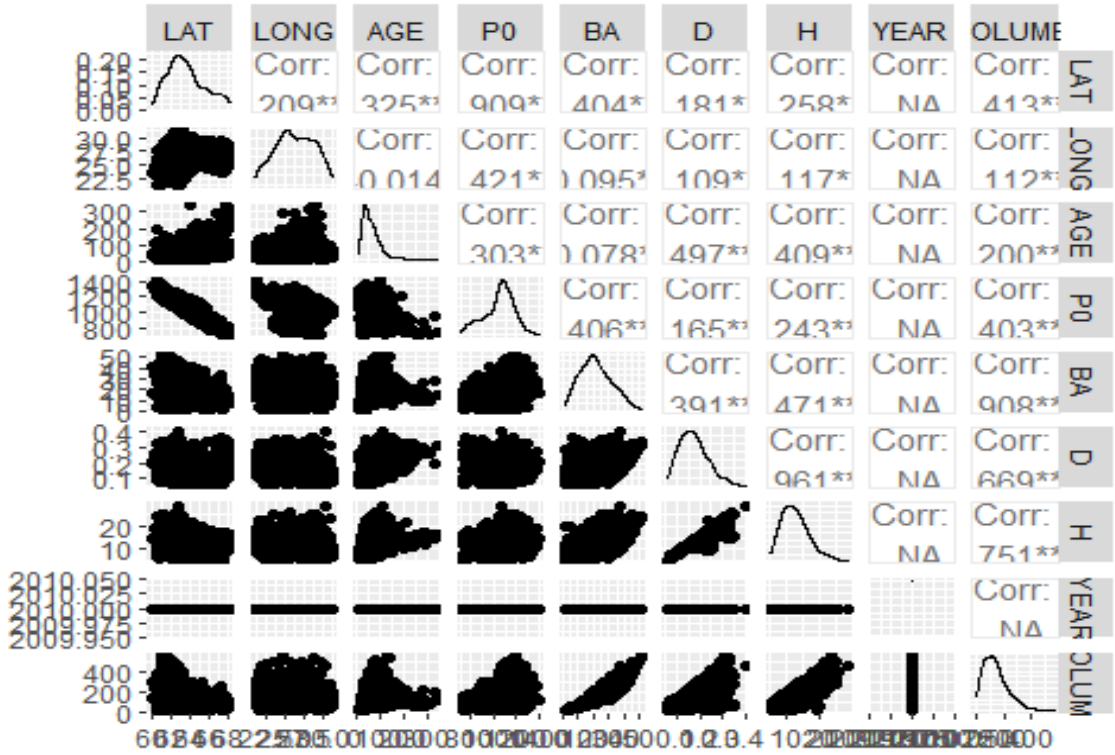
3. There is a strong correlation between height and diameter which may be an issue in model development as the assumption on which linear regression models are built

is that significant interaction is not expected to exist between the independent variables.

```
library(GGally)

trees_sub<-subset(trees, select = -c(SP_GROUP, FOREST_TYPE) )
treesv_sub<-subset(treesv, select = -c(SP_GROUP, FOREST_TYPE) )

 #Modelling data
ggpairs(trees_sub)
```




```
#Validation data
ggpairs(treesv_sub)
```



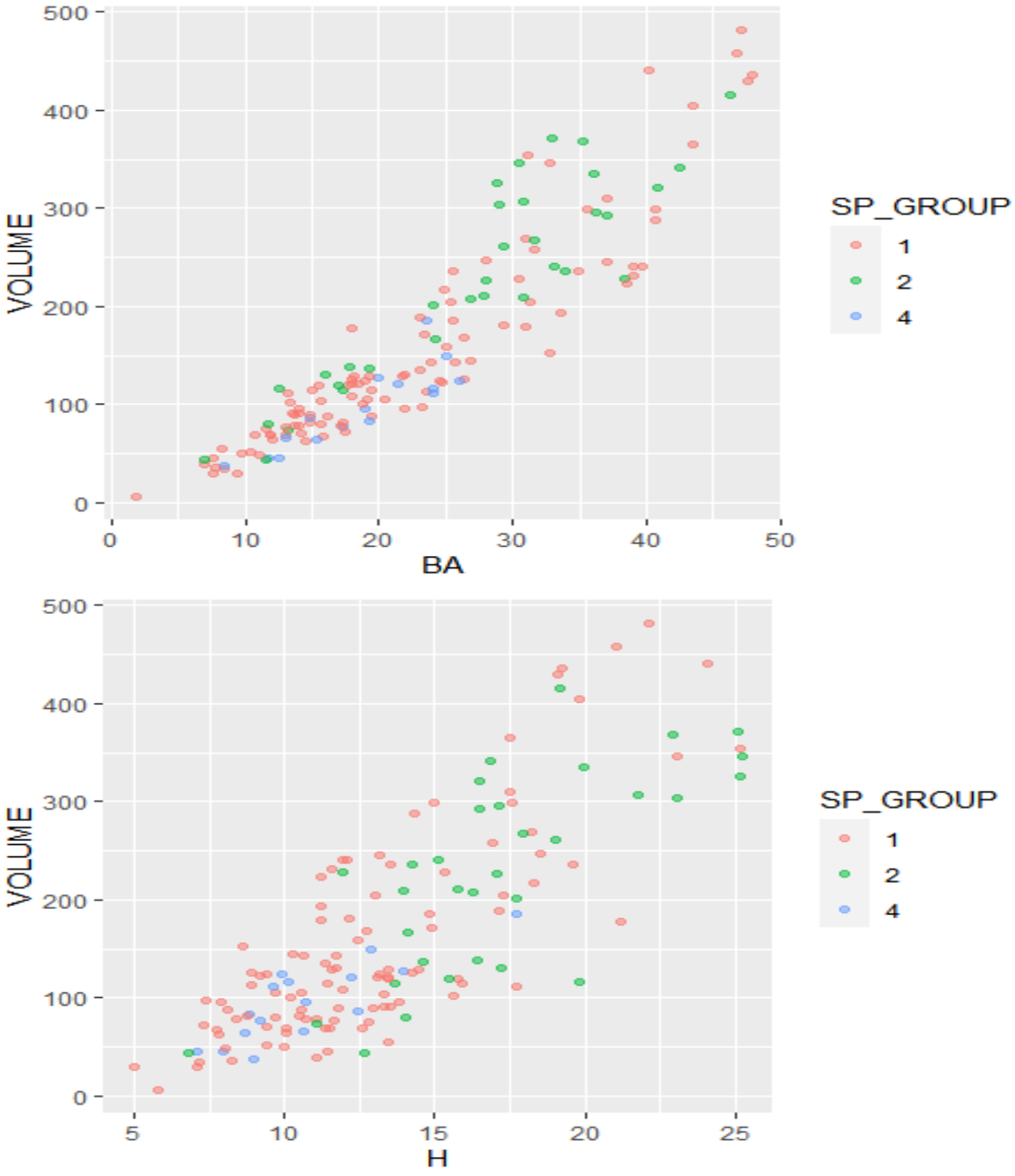
Volume by species

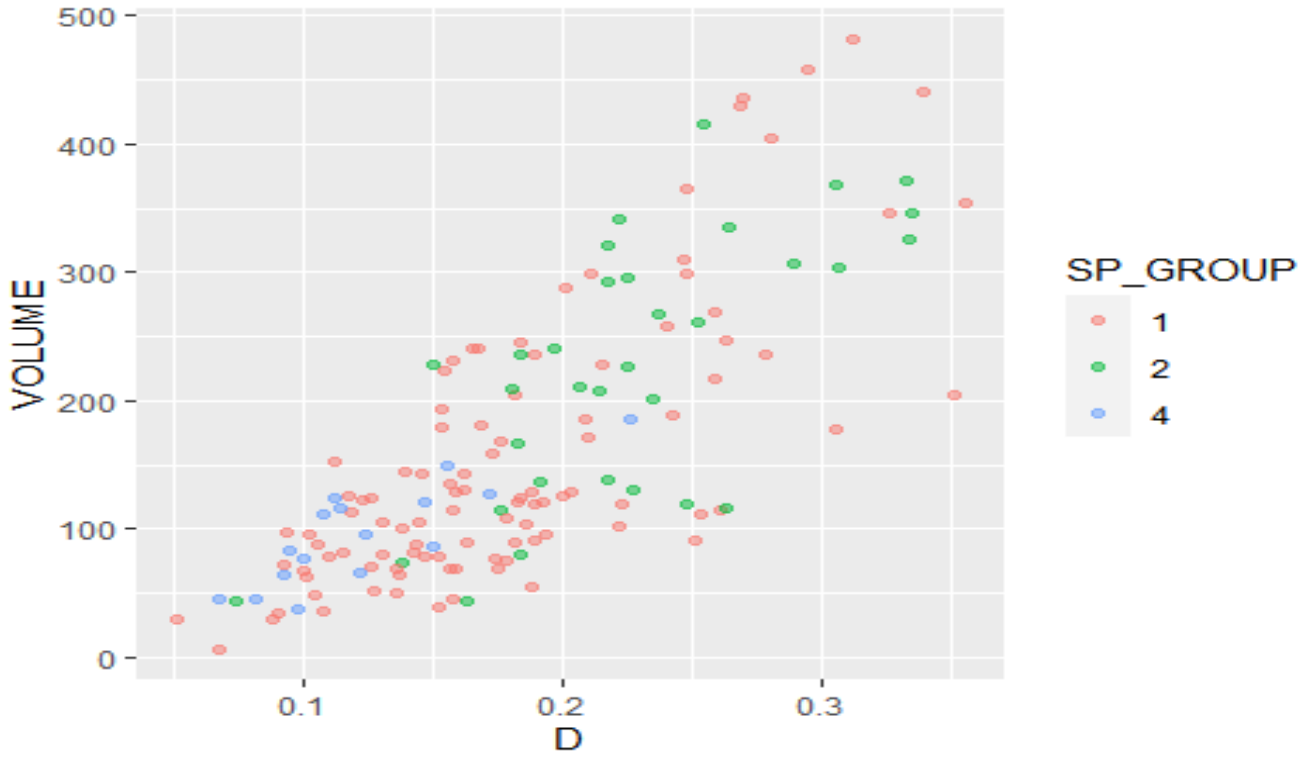
The way features that are strongly correlated with stand volume are by dominant species are presented below. Considering trees on average by largest size (diameter and height), it is Scots pine followed by Norway spruce. Consequently, by area and volume, it is Scots pine, next is Norway spruce, then Birch and other deciduous are the smallest. Thus, the bigger trees are from the first two, whereas Birch and other deciduous trees are largely smaller. This view is further supported by the boxplots. This can be useful in making decisions on forest operations as it is easy to know from which tree subset a certain volume could be expected. Also, there are few trees that look like outliers with values at the extremes relative to others in the data, but they fit the bid as they are well within the size range for the named trees[4,5]. At the upper end of the spectrum, they are potentially high performing trees.

Again, a look at the histogram shows the plots are approximately bell-shaped and symmetrical around the mean. Therefore, it is right to assume that the considered independent fields fairly approximate a pattern of normal distribution. Comparing these plots with tables of summary, the kurtosis is less than zero, so this is a platykurtic distribution and has a thinner tail (light tailed). However, because the skewness of these features is zero except for BA that is slightly right-skewed (skewness= 0.02), the assumption of normal distribution is maintained. Similar symmetry is observed in the test data too but BA in this case is more skewed to the right.

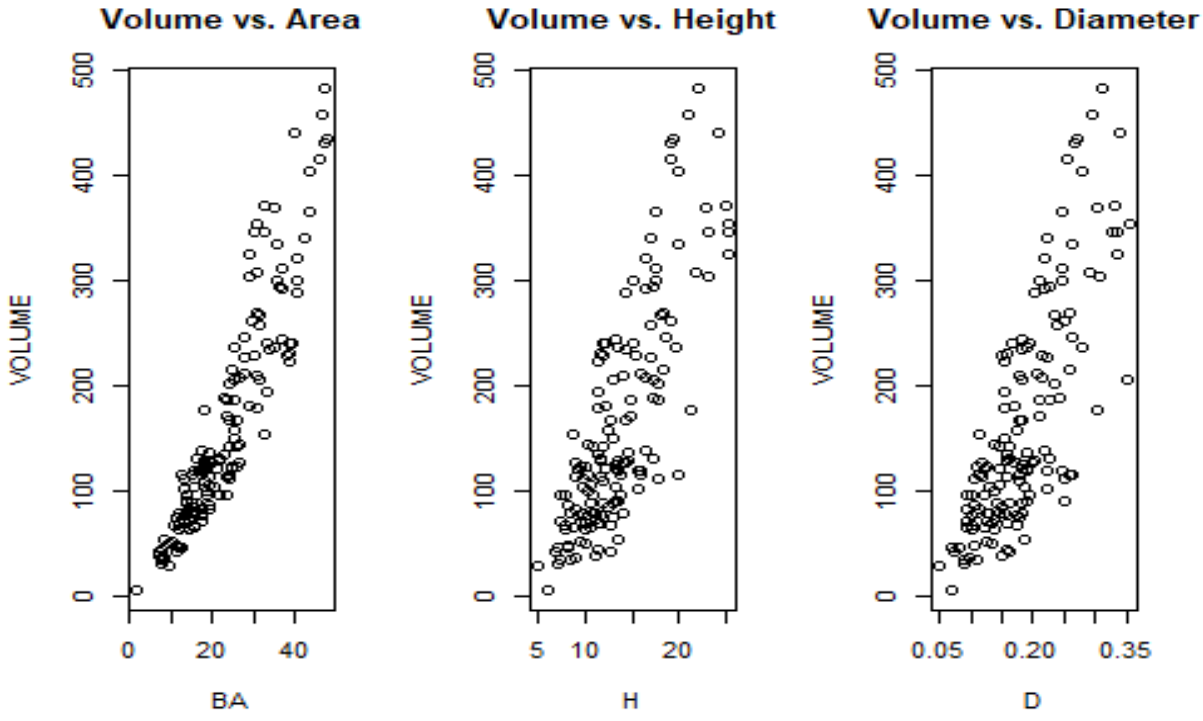
Lastly, further analysis using the sophisticated tidyverse package's pipe operator with dplyr's slice_max and slice_min functions gave some specific interesting findings. First and

foremost, the various forests under consideration are in Finland. Additionally, the lowest and the highest points on the map where the forest are using the geographic coordinates are Western Finland (Uusikaupunki and Närpes) and North Karelia region (Ilomantsi) respectively. Also, the oldest tree, and the trees with the highest height, largest diameter, area, and volume are found in the fertile forest, and not the very fertile one. Besides, the shortest, smallest and largest diameter trees, and those with the largest area and volume are Scots pine but the tallest tree is a Norway spruce. Next, the oldest and youngest trees are also Scots pine, they are however found in not so fertile forest type 3 and 4 respectively. It would have been interesting to know more about the forests because additional information such as the ages of the forests, difference in ages etc. could give further insight and explain observations within the forest types. Finally, the tallest tree is in Kitukankaantie, Jyväskylä while the oldest is in Sodankylä, Lapland.

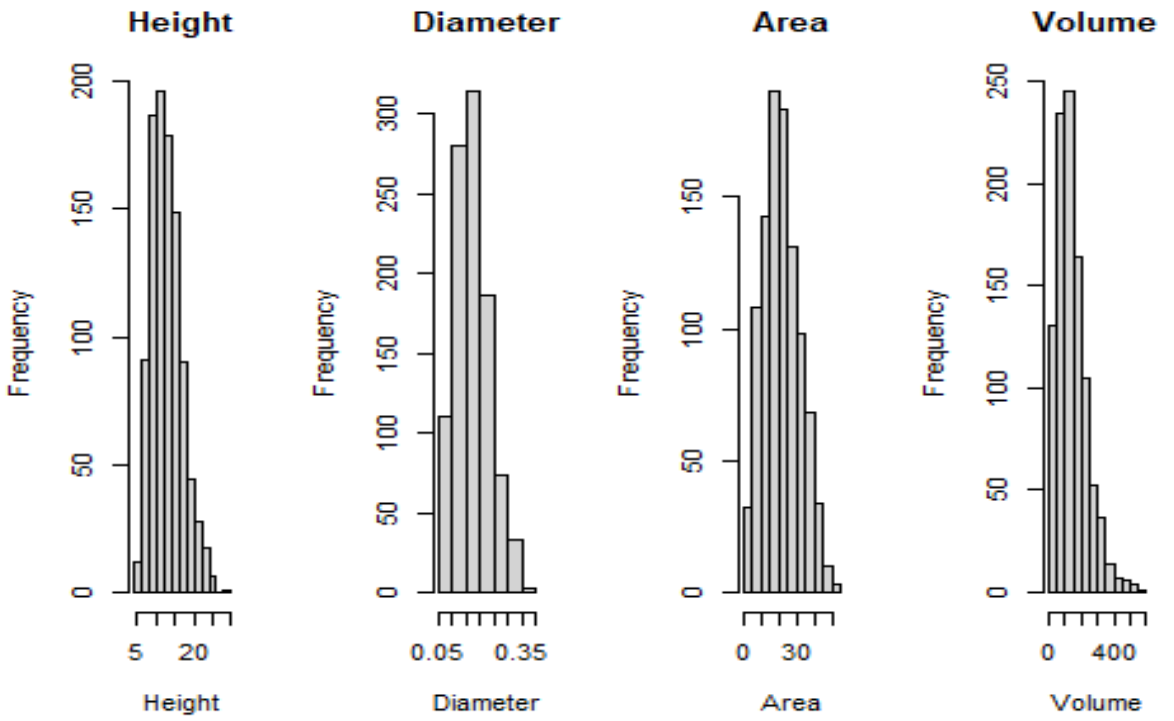

VOLUME
500
400
300
200
100
0
0
10
20
30
40
50
BA
SP_GROUP
1
2
4
VOLUME
500
400
300
200
100
0
5
10
15
20
25
H
SP_GROUP
1
2
4



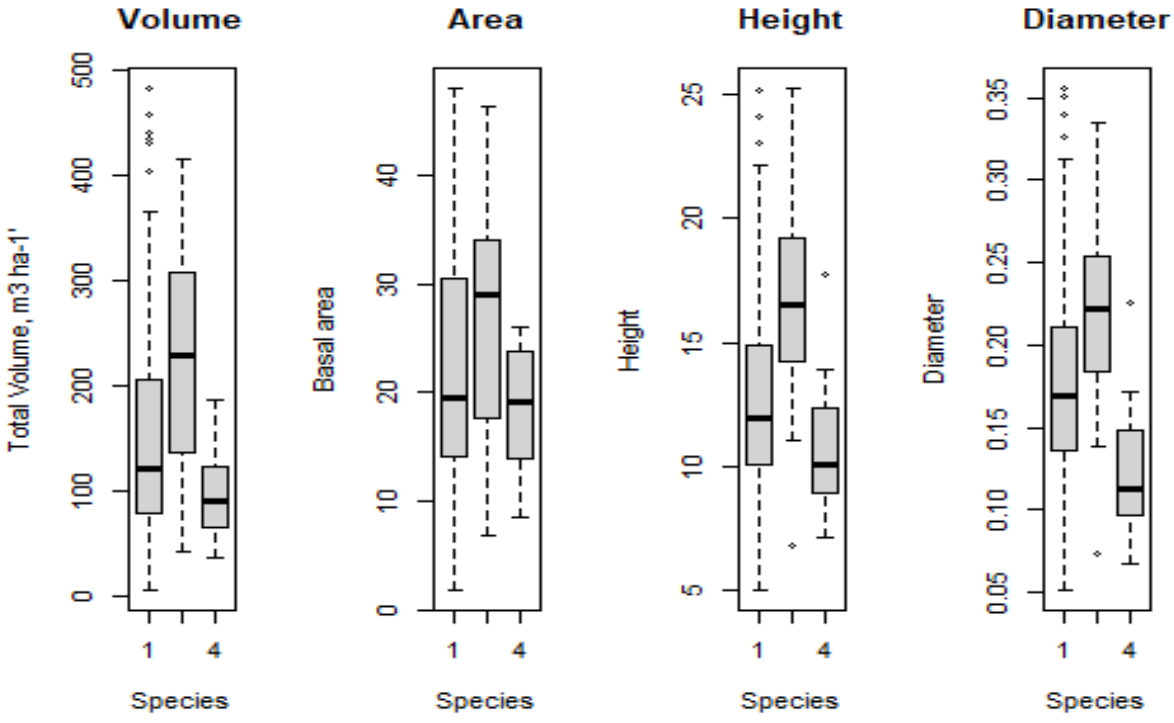
Scatter: plots against volume



Histogram: distribution



Boxplot: features by species group



Further analysis

To mine for specific information within datasets to better understand the dataset

```
library(tidyverse)

# Record of tallest tree

trees%>% slice_max(H)

#trees[which.max(trees$H),]
#trees[trees$H==max(trees$H,]

##    LAT LONG SP_GROUP AGE       P0       BA         D        H YEAR
FOREST_TYPE
## 1 62.2 25.6        2  84 1102.735 30.49293 0.3343574 25.21582 2010
2
##     VOLUME
## 1 345.7836

#Shortest tree
trees%>% slice_min(H)

##    LAT LONG SP_GROUP AGE       P0      BA          D       H YEAR
FOREST_TYPE
## 1 63.2 30.8        1  72 1003.327 9.39354 0.05130065 5.00662 2010
5
##     VOLUME
## 1 29.92449
```

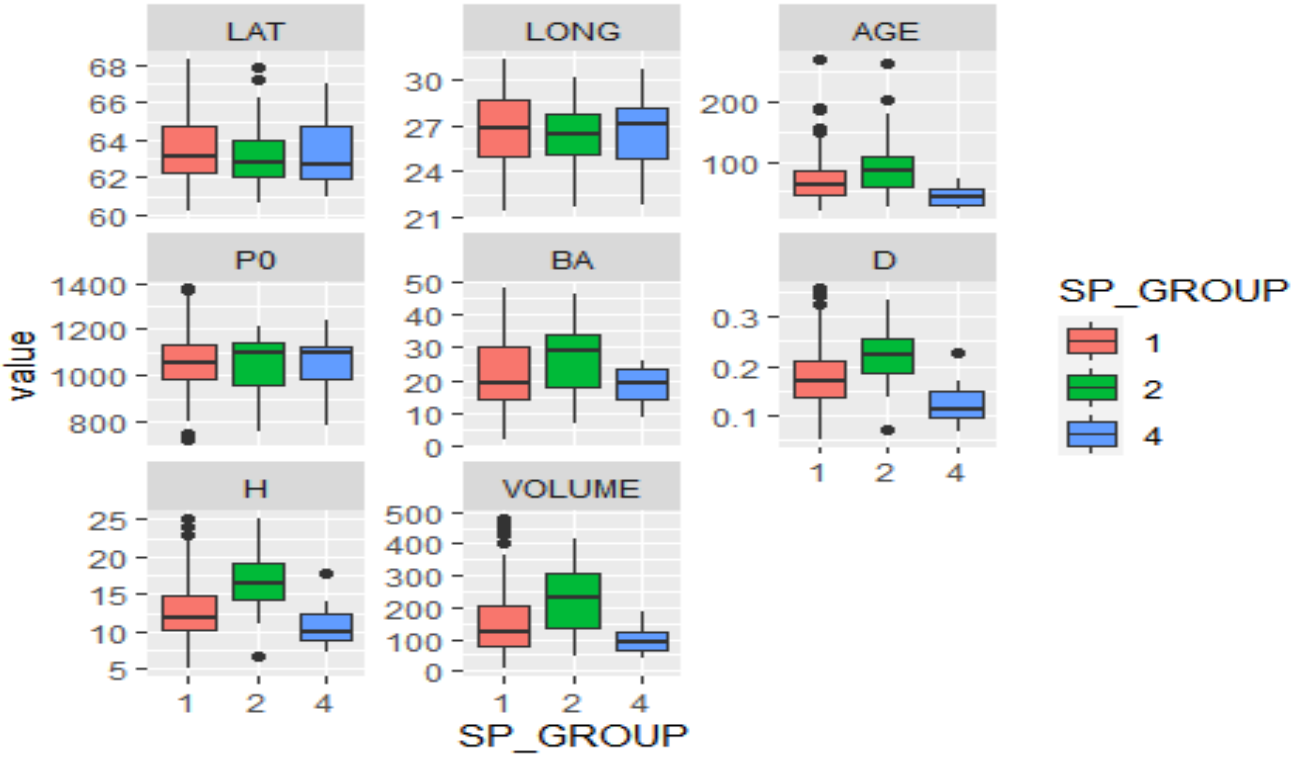
Additional pattern

Given it is more difficult to observe change of character/factor across the levels in continuous variables, I decided to turn around and look at the spread of the different continuous variables in every level of the factors. These operations were performed using the melt function of the reshape package in R together with the ubiquitous ggplot2 visualization package. In this way, the distribution of these variables over the three dominant tree species and the forest types is presented below. Clearly, the pattern in distribution is consistent for the variables that are strongly correlated with volume for dominant species, but the opposite is true across forest types. Therefore, if decision is to be taken from the preceding analysis and visualization on what feature to use in building a predictive model, the strongly correlated fields, BA, H, D will stand out as top picks. Contrariwise, longitude can be dropped since it is negatively correlated, by extension, latitude will be unnecessary as they are both pairs denoting geographical location. Furthermore, among discrete variables, forest type can be readily removed too.

```
require(reshape2)

treesz<-melt(subset(trees, select = -c(YEAR)), id=c("SP_GROUP",
"FOREST_TYPE") )
```

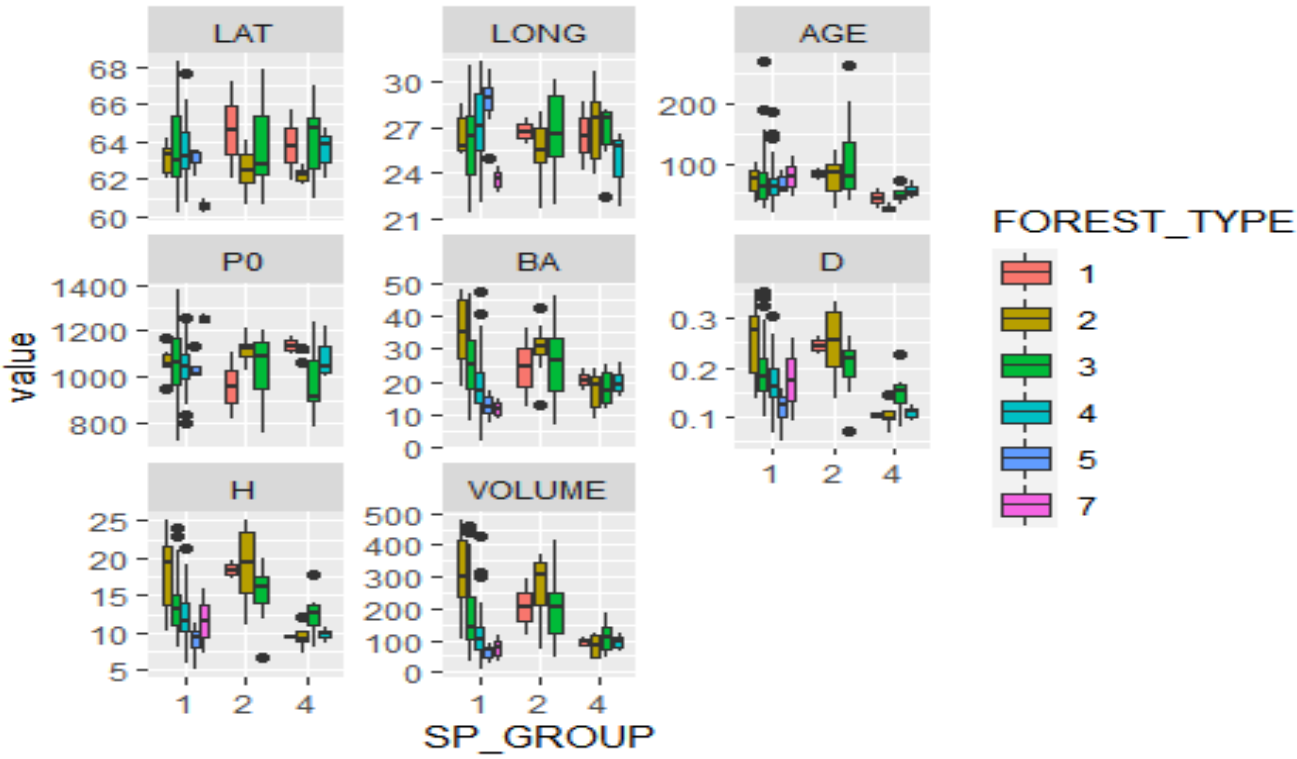
Boxplot by species

```
ggplot(treesz, aes(SP_GROUP, y = value, fill=SP_GROUP)) +
  geom_boxplot() +
  facet_wrap(~variable, scales="free_y")
```



Boxplot by Forest types

```
ggplot(treesz, aes(SP_GROUP, y = value, fill=FOREST_TYPE)) +
  geom_boxplot() +
  facet_wrap(~variable, scales="free_y")
```



Predictive models

The linear models and algorithms used here are from literature[6]. The approach is to first build multiple linear regression models upon all the parameters and use search algorithm stepAIC to select variables that are good enough for these models. For StepAIC to automate the model selection process, and by extension features, using Stepwise Algorithm a certain criterion is used. Akaike Information Criterion (AIC) penalizes models that utilize excessive parameters while tracking that which gives best variation in the dataset. Thus, it gives penalty for making the model complicated than necessary. The stepAIC function is from packages MASS and the companion library car. The variable selections are compared with what we have from the previous analysis and visualization of the data as well as domain knowledge. Subsequently, different linear models are built on these results in order to find the best.

Model lmForest0

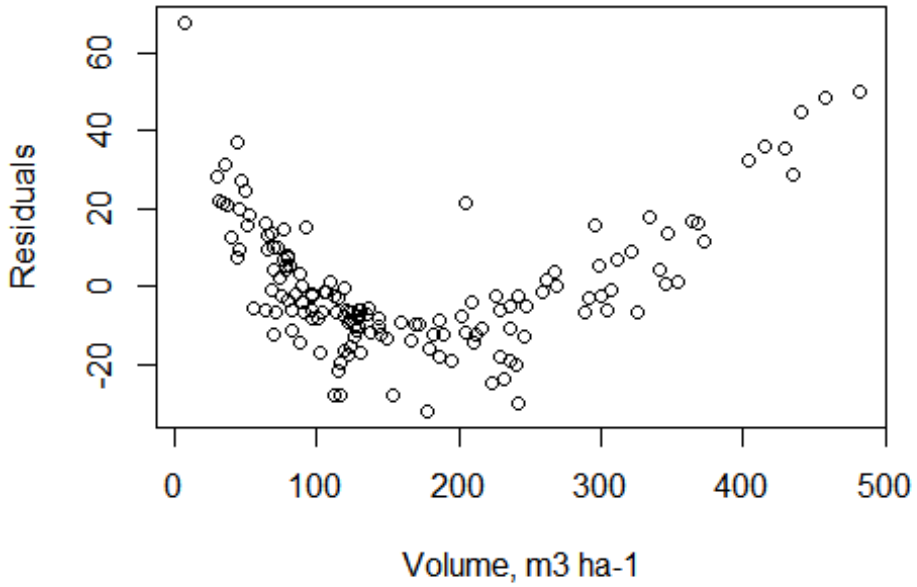
```
library(MASS)

library(car)

lmForest0<-lm(VOLUME~BA+D+H+P0+ SP_GROUP+FOREST_TYPE+LAT+LONG+AGE, data =
trees)

summary(lmForest0)
```

```
##
## Call:
## lm(formula = VOLUME ~ BA + D + H + P0 + SP_GROUP + FOREST_TYPE +
##     LAT + LONG + AGE, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.582  -9.964  -3.525   8.440  67.566
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -99.65659  191.89522  -0.519   0.6044
## BA              6.94259    0.18777  36.973  < 2e-16 ***
## D            -315.64877  139.07028  -2.270   0.0248 *
## H              15.34453    2.07214   7.405 1.27e-11 ***
## P0             -0.01350    0.03524  -0.383   0.7024
## SP_GROUP2     -11.30211    5.07091  -2.229   0.0275 *
## SP_GROUP4     -16.00420    6.29960  -2.541   0.0122 *
## FOREST_TYPE2   16.18546    9.77789   1.655   0.1002
## FOREST_TYPE3    7.94591    9.52040   0.835   0.4054
## FOREST_TYPE4    8.43666   10.12829   0.833   0.4063
## FOREST_TYPE5   26.56334   12.62564   2.104   0.0372 *
## FOREST_TYPE7   24.04939   16.91420   1.422   0.1574
## LAT            -0.49263    2.25923  -0.218   0.8277
## LONG           -0.17500    0.76918  -0.228   0.8204
## AGE            -0.02838    0.05157  -0.550   0.5830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.44 on 135 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9734
## F-statistic: 390.9 on 14 and 135 DF,  p-value: < 2.2e-16

steplmForest0<-stepAIC(lmForest0, direction = "both")

## Start:  AIC=871.76
## VOLUME ~ BA + D + H + P0 + SP_GROUP + FOREST_TYPE + LAT + LONG +
##     AGE
##
##               Df Sum of Sq    RSS     AIC
## - LAT          1        14  41056  869.81
## - LONG         1        16  41058  869.81
## - P0           1        45  41087  869.92
## - AGE          1        92  41134  870.09
## <none>                      41042  871.76
## - FOREST_TYPE  5      3506  44548  874.05
## - D            1      1566  42608  875.37
## - SP_GROUP     2      2396  43438  876.27
## - H            1     16671  57713  920.89
## - BA           1    415593 456635 1231.15
```

```
##
## Step:  AIC=869.81
## VOLUME ~ BA + D + H + P0 + SP_GROUP + FOREST_TYPE + LONG + AGE
##
##               Df Sum of Sq    RSS     AIC
## - LONG         1         6  41062  867.83
## - P0           1        49  41106  867.99
## - AGE          1       103  41159  868.18
## <none>                      41056  869.81
## + LAT          1        14  41042  871.76
## - FOREST_TYPE  5      3791  44848  873.06
## - D            1      1615  42672  873.60
## - SP_GROUP     2      2390  43447  874.30
## - H            1     17127  58183  920.11
## - BA           1    428283 469340 1233.27
##
## Step:  AIC=867.83
## VOLUME ~ BA + D + H + P0 + SP_GROUP + FOREST_TYPE + AGE
##
##               Df Sum of Sq    RSS     AIC
## - P0           1        45  41107  866.00
## - AGE          1        97  41159  866.18
## <none>                      41062  867.83
## + LONG         1         6  41056  869.81
## + LAT          1         5  41058  869.81
## - FOREST_TYPE  5      3833  44895  871.22
## - D            1      1640  42702  871.70
## - SP_GROUP     2      2431  43493  872.46
## - H            1     17163  58225  918.22
## - BA           1    441964 483026 1235.58
##
## Step:  AIC=866
## VOLUME ~ BA + D + H + SP_GROUP + FOREST_TYPE + AGE
##
##               Df Sum of Sq    RSS     AIC
## - AGE          1        55  41162  864.20
## <none>                      41107  866.00
## + P0           1        45  41062  867.83
## + LAT          1        21  41087  867.92
## + LONG         1         1  41106  867.99
## - FOREST_TYPE  5      3809  44916  869.29
## - D            1      1840  42947  870.56
## - SP_GROUP     2      2650  43757  871.37
## - H            1     17721  58828  917.76
## - BA           1    478955 520062 1244.66
##
## Step:  AIC=864.2
## VOLUME ~ BA + D + H + SP_GROUP + FOREST_TYPE
##
##               Df Sum of Sq    RSS     AIC
```

```
## <none>                      41162  864.20
## + AGE          1        55  41107  866.00
## + P0           1         3  41159  866.18
## + LONG         1         1  41161  866.19
## + LAT          1         0  41162  866.19
## - FOREST_TYPE  5      3862  45024  867.65
## - D            1      2293  43455  870.33
## - SP_GROUP     2      2981  44143  870.68
## - H            1     19093  60255  919.36
## - BA           1    484477 525639 1244.26

summary(steplmForest0)

##
## Call:
## lm(formula = VOLUME ~ BA + D + H + SP_GROUP + FOREST_TYPE, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.869 -10.395  -3.626   8.005  67.694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -149.9021    11.0808 -13.528  < 2e-16 ***
## BA              6.9319     0.1714  40.448  < 2e-16 ***
## D            -350.0849   125.8157  -2.783  0.00614 **
## H              15.7122     1.9568   8.030 3.72e-13 ***
## SP_GROUP2     -12.3039     4.7067  -2.614  0.00993 **
## SP_GROUP4     -16.7088     6.0754  -2.750  0.00675 **
## FOREST_TYPE2   16.7035     9.4589   1.766  0.07961 .
## FOREST_TYPE3    7.9888     9.2137   0.867  0.38741
## FOREST_TYPE4    8.1572     9.6705   0.844  0.40039
## FOREST_TYPE5   25.7168    11.6658   2.204  0.02914 *
## FOREST_TYPE7   23.1477    15.6063   1.483  0.14028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.21 on 139 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9741
## F-statistic: 561.7 on 10 and 139 DF,  p-value: < 2.2e-16

vif(steplmForest0)

##                  GVIF Df GVIF^(1/(2*Df))
## BA           1.615735  1        1.271116
## D           35.542696  1        5.961770
## H           38.154129  1        6.176903
## SP_GROUP     2.698118  2        1.281638
## FOREST_TYPE  1.938092  5        1.068409
```

```
#Plot of residuals
plot(trees$VOLUME, residuals(steplmForest0),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```

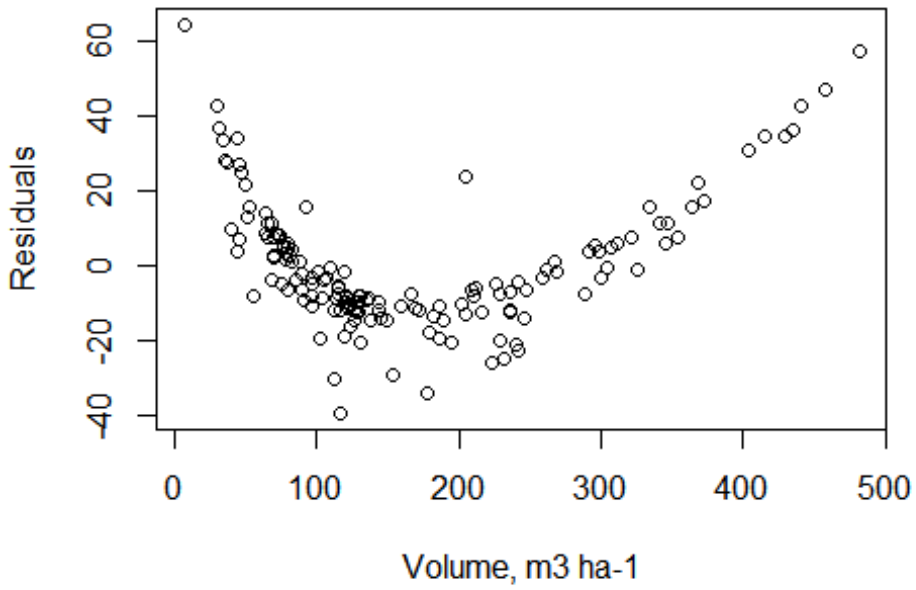



```
#plot(trees$VOLUME, residuals(lmForest0),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```

1. Model lmForest1

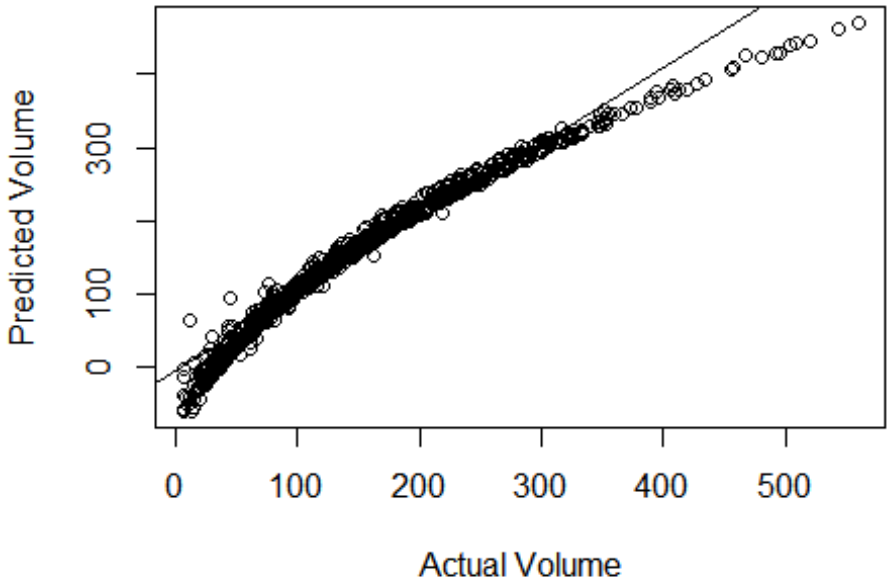
```
lmForest1<-lm(VOLUME~ BA+D+H+SP_GROUP,data = trees) #Drop other variables
that are poor and negatively correlated
summary(lmForest1)

##
## Call:
## lm(formula = VOLUME ~ BA + D + H + SP_GROUP, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.519 -10.776  -3.589   7.470  64.537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -139.4624     5.2683 -26.472  < 2e-16 ***
## BA             6.8775     0.1678  40.978  < 2e-16 ***
## D           -391.8161   125.1276  -3.131  0.00211 **
## H             16.3658     1.9434   8.421 3.43e-14 ***
## SP_GROUP2    -12.3470     4.4647  -2.765  0.00643 **
```

```
## SP_GROUP4    -18.2707     5.7825  -3.160  0.00193 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.68 on 144 degrees of freedom
## Multiple R-squared:  0.9736, Adjusted R-squared:  0.9727
## F-statistic:  1062 on 5 and 144 DF,  p-value: < 2.2e-16

vif(lmForest1)

##                GVIF Df GVIF^(1/(2*Df))
## BA         1.467632  1        1.211458
## D         33.295825  1        5.770253
## H         35.643948  1        5.970255
## SP_GROUP   2.128176  2        1.207819

#Plot of residuals
plot(trees$VOLUME, residuals(lmForest1),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```




```
#Validate with test data: treesv. No validation with this search

VOLUME_Pred<-predict(lmForest1, treesv)
#predicted is VOLUME_Pred


#Add predicted volume as column to validation data(treesv)
```

```
treesv["Predicted"]<-VOLUME_Pred

#Comparism table
Volume_comp<-treesv[,c("VOLUME", "Predicted")]
#Volume_comp
#plot actual volume vs predicted volume

plot(treesv$VOLUME,treesv$Predicted, xlab = "Actual Volume", ylab =
"Predicted Volume")
#plot(Volume_comp$VOLUME, Volume_comp$Predicted, xlab = "Actual Volume", ylab
= "Predicted Volume") #same
#Add trendline
abline(lm(treesv$Predicted~treesv$VOLUME))
```




```
#Comparison
library(Metrics)

## Warning: package 'Metrics' was built under R version 4.1.2

bias(actual =Volume_comp[,1] ,predicted =Volume_comp[,2] )

## [1] -0.8560809

rmse(actual =Volume_comp[,1] ,predicted =Volume_comp[,2])

## [1] 19.11263
```

II Linear mixed effect model 2. Model lmForest2

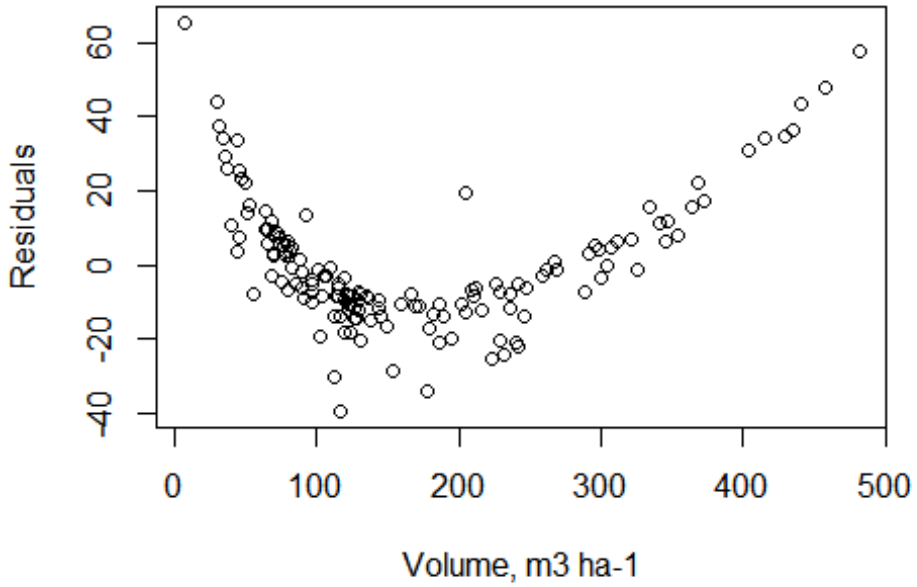
```
library(lme4)

library(sjPlot)

## Warning: package 'sjPlot' was built under R version 4.1.2

lmForest2<-lmer(VOLUME~BA+D+H+ (1|SP_GROUP), data = trees)
summary(lmForest2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: VOLUME ~ BA + D + H + (1 | SP_GROUP)
##    Data: trees
##
## REML criterion at convergence: 1275.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2341 -0.6014 -0.2450  0.4089  3.6996
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SP_GROUP (Intercept)  72.55    8.518
##  Residual             312.65   17.682
## Number of obs: 150, groups:  SP_GROUP, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -148.1462     7.8049 -18.981
## BA             6.8890     0.1676  41.096
## D           -348.0120   121.2174  -2.871
## H             15.7052     1.8846   8.333
##
## Correlation of Fixed Effects:
##    (Intr) BA     D
## BA -0.058
## D   0.380  0.125
## H  -0.477 -0.236 -0.975

tab_model(lmForest2)
```

Marginal R2 / Conditional R2

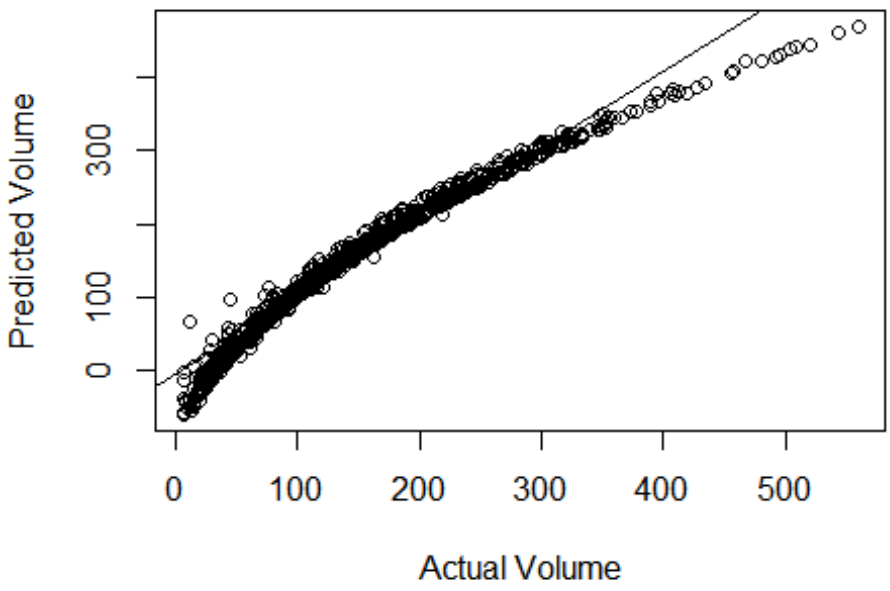
0.967 / 0.973

```
vif(lmForest2)

##        BA         D         H
##  1.393056 26.571414 27.707553
```

```
#Plot of residuals
plot(trees$VOLUME, residuals(lmForest2),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```




```
#Validate with test data: treesv

VOLUME_Pred<-predict(lmForest2, treesv)
#predicted is VOLUME_Pred

#Add predicted volume as column to validation data(treesv)
treesv["Predicted"]<-VOLUME_Pred

#Comparism table
Volume_comp<-treesv[,c("VOLUME", "Predicted")]
#Volume_comp
#plot actual volume vs predicted volume

plot(treesv$VOLUME,treesv$Predicted, xlab = "Actual Volume", ylab =
"Predicted Volume")

#Add trendline
abline(lm(treesv$Predicted~treesv$VOLUME))
```




```
#Comparison
bias(actual =Volume_comp[,1] ,predicted =Volume_comp[,2] )

## [1] -0.8511779

rmse(actual =Volume_comp[,1] ,predicted =Volume_comp[,2])

## [1] 19.16494
```

III Log model 3. Model lmForest3

```
lmForest3<-lm(log(VOLUME)~ log(BA)+log(D)+log(H)+SP_GROUP,data = trees) #Drop
other variables that are poor and negatively correlated
summary(lmForest3)

##
## Call:
## lm(formula = log(VOLUME) ~ log(BA) + log(D) + log(H) + SP_GROUP,
##     data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032203 -0.003503 -0.001748  0.002784  0.040278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.824567   0.058850  -14.01   <2e-16 ***
## log(BA)      0.997823   0.001556  641.11   <2e-16 ***
```

```
## log(D)      -0.129021   0.012641  -10.21   <2e-16 ***
## log(H)       0.971356   0.014830   65.50   <2e-16 ***
## SP_GROUP2   -0.026391   0.002109  -12.51   <2e-16 ***
## SP_GROUP4   -0.117740   0.003154  -37.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007965 on 144 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 2.304e+05 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
#steplmForest3<-stepAIC(lmForest3, direction = "both")
```

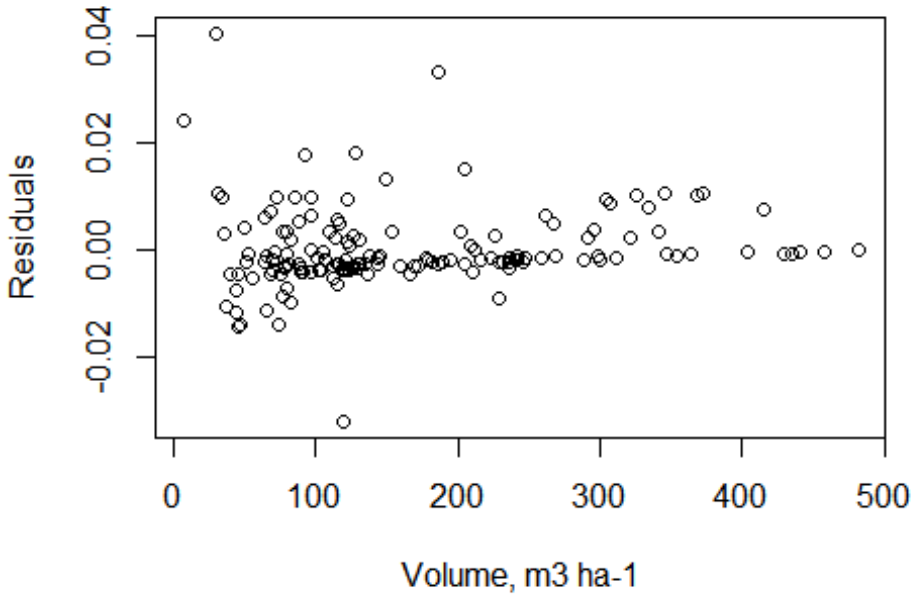
```
summary(lmForest3)
```

```
##
## Call:
## lm(formula = log(VOLUME) ~ log(BA) + log(D) + log(H) + SP_GROUP,
##     data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032203 -0.003503 -0.001748  0.002784  0.040278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.824567   0.058850  -14.01   <2e-16 ***
## log(BA)      0.997823   0.001556  641.11   <2e-16 ***
## log(D)      -0.129021   0.012641  -10.21   <2e-16 ***
## log(H)       0.971356   0.014830   65.50   <2e-16 ***
## SP_GROUP2   -0.026391   0.002109  -12.51   <2e-16 ***
## SP_GROUP4   -0.117740   0.003154  -37.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007965 on 144 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 2.304e+05 on 5 and 144 DF,  p-value: < 2.2e-16
```

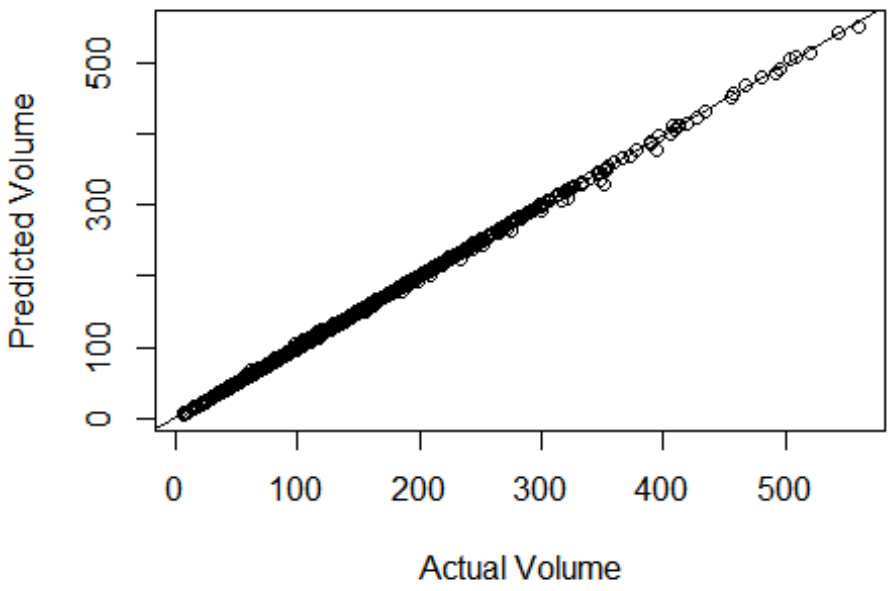
```
vif(lmForest3)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## log(BA)   1.474849  1        1.214434
## log(D)   54.269501  1        7.366784
## log(H)   55.536772  1        7.452300
## SP_GROUP  3.036069  2        1.320012
```

```
#Plot of residuals
plot(trees$VOLUME, residuals(lmForest3),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```




```
#Validate with test data: treesv

VOLUME_Pred<-exp(predict(lmForest3, treesv))
#predicted is VOLUME_Pred


#Add predicted volume as column to validation data(treesv)
treesv["Predicted"]<-VOLUME_Pred

#Comparism table
Volume_comp<-treesv[,c("VOLUME", "Predicted")]
#Volume_comp
#plot actual volume vs predicted volume

plot(treesv$VOLUME,treesv$Predicted, xlab = "Actual Volume", ylab =
"Predicted Volume")
#plot(Volume_comp$VOLUME, Volume_comp$Predicted) #same
#Add trendline
abline(lm(treesv$Predicted~treesv$VOLUME))
```




```
#Volume comparison table for the first five rows
head(Volume_comp)
```

```
##      VOLUME Predicted
## 1 126.96147 127.41878
## 2 301.08814 301.33219
## 3 229.05027 229.58897
## 4 115.16212 114.51892
## 5 319.11954 317.11840
## 6  59.13552  58.56241
```

```
#4. Model lmForest4

# No D: to investigate if the supposed multicolinearity between H&D is an
issue
lmForest4<-lm(log(VOLUME)~ log(BA)+log(H)+SP_GROUP,data = trees) #Drop other
variables that are poor and negatively correlated
summary(lmForest3)
```

```
##
## Call:
## lm(formula = log(VOLUME) ~ log(BA) + log(D) + log(H) + SP_GROUP,
##     data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032203 -0.003503 -0.001748  0.002784  0.040278
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.824567   0.058850  -14.01   <2e-16 ***
## log(BA)      0.997823   0.001556  641.11   <2e-16 ***
## log(D)      -0.129021   0.012641  -10.21   <2e-16 ***
## log(H)       0.971356   0.014830   65.50   <2e-16 ***
## SP_GROUP2   -0.026391   0.002109  -12.51   <2e-16 ***
## SP_GROUP4   -0.117740   0.003154  -37.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007965 on 144 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 2.304e+05 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
#steplmForest4<-stepAIC(lmForest4, direction = "both")
```

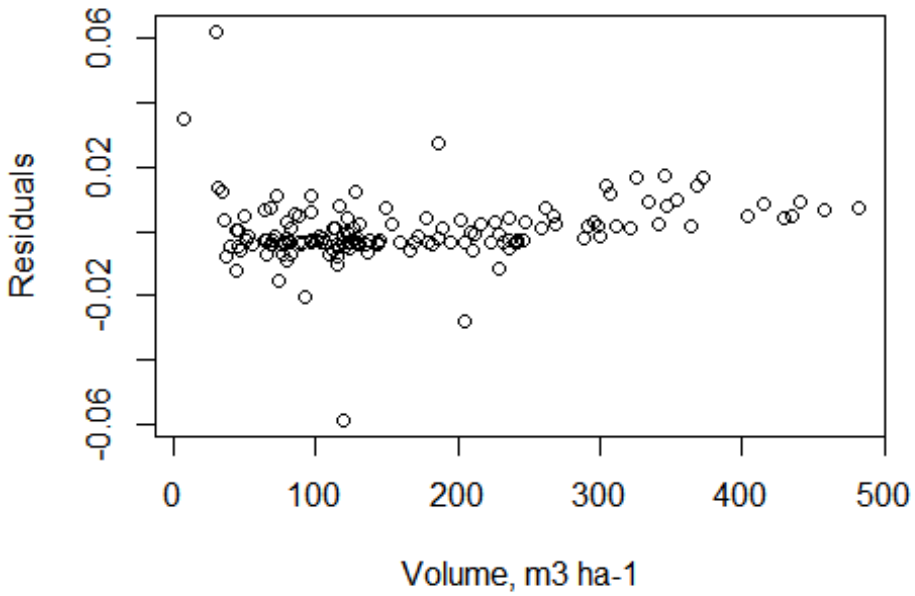
```
summary(lmForest4)
```

```
##
## Call:
## lm(formula = log(VOLUME) ~ log(BA) + log(H) + SP_GROUP, data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.058831 -0.003868 -0.002027  0.003385  0.062121
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.226765   0.007477 -30.328  < 2e-16 ***
## log(BA)      0.999410   0.002026 493.303  < 2e-16 ***
## log(H)       0.822357   0.003415 240.790  < 2e-16 ***
## SP_GROUP2   -0.014085   0.002264  -6.222 4.97e-09 ***
## SP_GROUP4   -0.094414   0.002844 -33.197  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01042 on 145 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 1.683e+05 on 4 and 145 DF,  p-value: < 2.2e-16
```

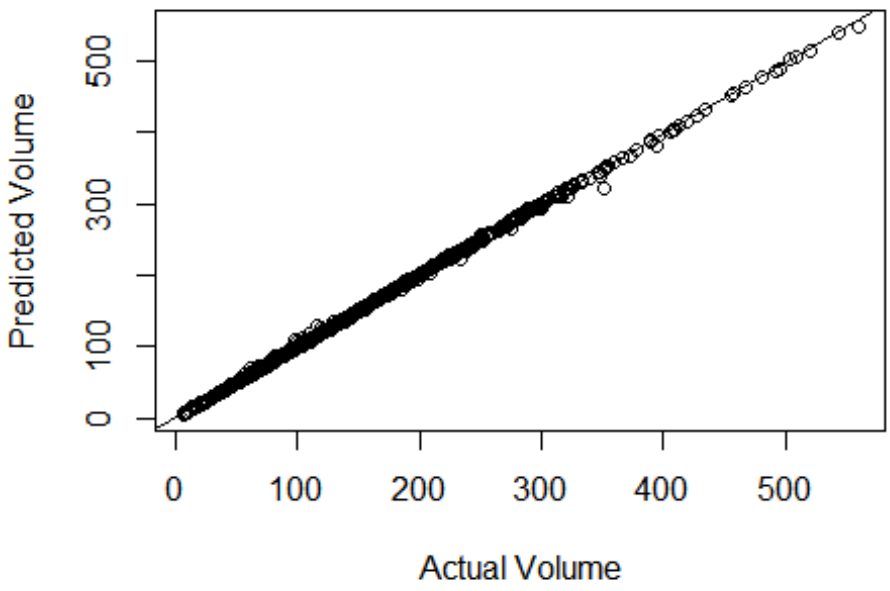
```
vif(lmForest4)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## log(BA)  1.460121  1        1.208355
## log(H)   1.720865  1        1.311817
## SP_GROUP 1.236342  2        1.054471
```

```r
#Plot of residuals
plot(trees$VOLUME, residuals(lmForest4),xlab = 'Volume, m3 ha-1', ylab =
'Residuals')
```




```r
#Validate with test data: treesv

VOLUME_Pred<-exp(predict(lmForest4, treesv))
#predicted is VOLUME_Pred


#Add predicted volume as column to validation data(treesv)
treesv["Predicted"]<-VOLUME_Pred

#Comparism table
Volume_comp<-treesv[,c("VOLUME", "Predicted")]
#Volume_comp
#plot actual volume vs predicted volume

plot(treesv$VOLUME,treesv$Predicted, xlab = "Actual Volume", ylab =
"Predicted Volume")
#plot(Volume_comp$VOLUME, Volume_comp$Predicted) #same
#Add trendline
abline(lm(treesv$Predicted~treesv$VOLUME))
```




```
#Comparison
bias(actual =Volume_comp[,1] ,predicted =Volume_comp[,2] )

## [1] 0.07094697

rmse(actual =Volume_comp[,1] ,predicted =Volume_comp[,2])

## [1] 2.038959

#Comparison
# RMSE Determination based on the best log model


bias(actual =Volume_comp[,1] ,predicted =Volume_comp[,2] )

## [1] 0.07094697

rmse(actual =Volume_comp[,1] ,predicted =Volume_comp[,2])

## [1] 2.038959

#Volume comparison table for the first five rows
head(Volume_comp)

##      VOLUME Predicted
## 1 126.96147 127.31018
## 2 301.08814 301.87056
## 3 229.05027 229.73504
```

```
## 4 115.16212 116.35167
## 5 319.11954 316.85447
## 6  59.13552  58.51299
```

Discussion

For a start, without preempting which variable to use based on earlier analysis that revealed likelihood of different features being usable in model development for forest data, multiple linear regression model lmForest0 was first built using all the features available in both datasets. This model was lumped into the stepAIC search algorithm to carry out feature selection automatically. Interestingly, in its iterative operation, it first eliminated the poorly and negatively correlated longitude. Subsequently, it excluded latitude, then precipitation level before age. This is in sync with earlier observations from our analysis and visualization. Based on the summary, the other fields are favoured, and from the first iteration are proven to be statistically significant using alpha level of 0.05, as they each have $P<0.05$. Upon this, model lmForest1 was developed using the remaining features except forest types. Multiple R-squared decreased to 97.36 from 97.59% in the previous model, which by the way was an insignificant difference. The Variance inflation factor(vif) showed there is a chance that there exists multicollinearity, as a value greater than 5 is a good pointer. Thus, this suggests there is possibly a strong interaction between some predictor variables in the model and raises the suspicion of aliased coefficients existing in the model. Markedly, this matches observations from the visualization plots in this report where there was high correlation between the height and diameter. With a closer look at the plot of residuals against the model, it is obvious that there is nonlinearity in the model. This can be confirmed from the plot of predicted volume against actual volume.The Root Mean Square Error (RMSE) and bias values from the lmforest1 are 19.11 and -0.86 respectively.

Furthermore, the preferred fields from the last model (lmForest1) were used in developing the linear mixed model lmForest2 to study how fixed effects (BA, H, D) and random effects (species group) work for this data. Clearly, there was no improvement over previous models. Also, the same linearity problem persists within this succeeding mixed model. Model lmForest2 had rmse value of 19.16 and bias of -0.85. Perhaps, it would have been better to investigate the effect of Stand ID as a random effect variable for this model, but this field was absent in the datasets.

In addition, linear regression models were developed using logarithmic adjustment to effect correction in the nonlinearity. First, lmForest3 was developed using the same preferred predictor feature selection. This model gave significant improvement over previous models considering the predictor variables explained 99.99% variation in stand volume with a significantly better RMSE and bias of 2.04 and 0.07 respectively. Besides, the model corrected the nonlinearity that burdened earlier models. Moreover, to see if multicollinearity is a real problem, lmForest4 was developed, excluding the diameter, whose inclusion seems to be less significant so far among the preferred fields. Model lmForest4 explained 99.98% variation in volume, and this was a little less than the former. When performing comparison of regression models that use similar dependent variables over the same estimation frame, as the adjusted R-squared increases, the standard error

for the due to the aforementioned reduction in R-squared and the relationship between height and diameter does not appear to be problematic, both features can be left in our models. Therefore, we jettison lmForest4 for lmForest3.

In conclusion, lmForest3 is the best model for predicting the stand volume where BA, H, D and SP_Group explains 99.99% variation in volume. In terms of how close the predicted volumes are to the reference volumes, because of its positive bias, it shows that the model overestimates. This means the predictions are higher than the actual volumes on average. In this case, a bias of 0.07 is reasonable as can be seen from the table above. This is better than the previous models, lmforest1 and lmforest2, which underestimate by larger margins. Besides, the smaller the values of RMSE, the smaller the difference between predicted and actual volumes, which means the better the model fits the data. Undoubtedly, lmForest3 (RMSE 2.04) trumps the other examined models using this parameter. There are some noticeable outliers in the data. The unequal scatter of residuals shows there is a sort of heteroscedasticity resulting from unequal variance for our residuals across the range of volumes. Thus, the variability for the random disturbance over the measurements is different.

Links:

1. https://wiki.uef.fi/download/attachments/44434211/Group1.csv?version=1&modificationDate=1479667748000&api=v2
2. https://wiki.uef.fi/download/attachments/44434211/validating_data.csv?version=1&modificationDate=1479668013000&api=v2
3. https://wiki.uef.fi/display/RESMET/Assignment
4. https://conifersociety.org/conifers
5. https://ferriseeds.com/products/paper-birch-betula-papyrifera
6. Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer