R ASSIGNMENT: RESEARCH METHODS IN FOREST SCIENCES

Group Members	**Student No.**
Amaitum Joshua Elukut	2216274
Uche Ndudi Omeoga	330344

School of Forest Sciences

University of Eastern Finland

INTRODUCTION

This report is about a task that followed lectures on use of R software during the Research Methods in Forest Sciences Course. R is a use tool for analyzing data because of its statistical computing power and graphics development potential which supports data visualization, interpretation, and consequently it's utility.

For this exercise we were given data (Group4) for a particular stand with variables like diameter, height, age, and basal area. We were then tasked to present some descriptive statistics and visually represent some relationships. Furthermore, we performed inferential statistics (regression analysis) which entailed developing a model for estimating the total stem volume in the stand. This model was a function of basic stand characteristics. Thereafter, we would test the suitability of the model, firstly by comparing its performance against data from which it was developed (modelling dataset), and secondly against data from an entirely new forest stand (validation dataset) with the aim of judging the precision, accuracy, and overall performance.

Objectives

- To determine the descriptive characteristics of the stand variables e.g. mean, minimum, maximum, and standard deviation.
- To make visual representations of some attributes/variables of the data such as the total volume.
- To create linear regression models for estimating total stem volume in the stand (m3 ha-1) as a function of basic stand characteristics (e.g. diameter, height, possibly other variables).
- To select the most ideal model and determine its suitability for predicting volume through comparisons with volume values from two datasets.
- To calculate Bias and RMSE values.

Note: Group4 dataset was used as the modelling data set

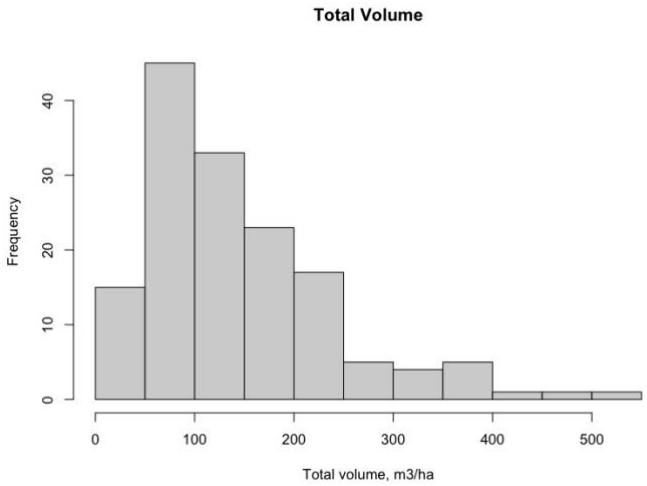
MATERIALS

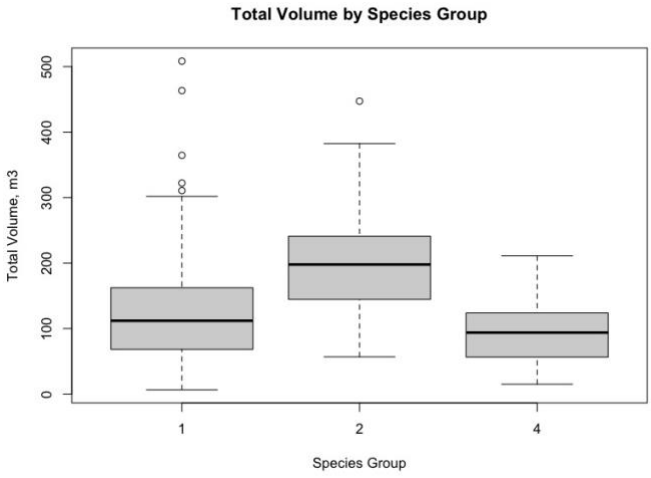
Modelling dataset

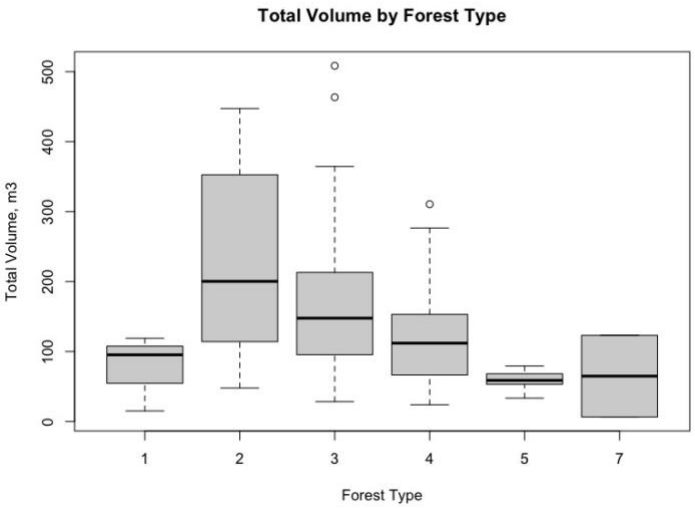
Summary statistics of modelling dataset

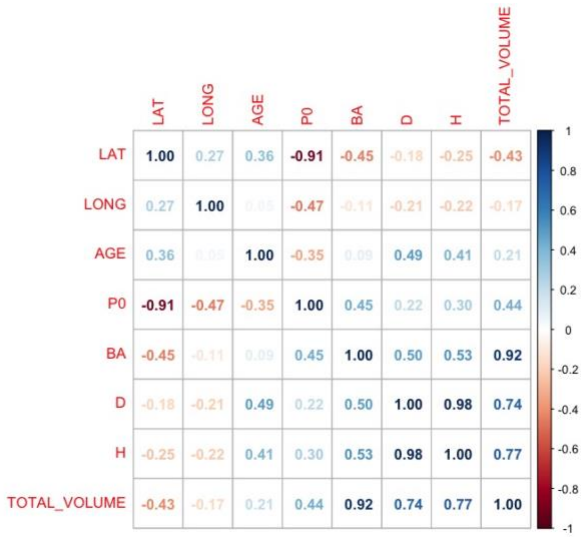
	AGE	P0	BA	D	H	TOTAL_VOLUME
Min	21.00	707.9	1.465	5.808	5.329	6.463
Max	256.00	1360.3	48.993	33.236	25.049	508.432
Mean	69.95	1046.8	21.641	17.063	12.403	144.928
Sd	43.42	135.2	9.962	5.986	3.990	94.776

Number of plots = 150

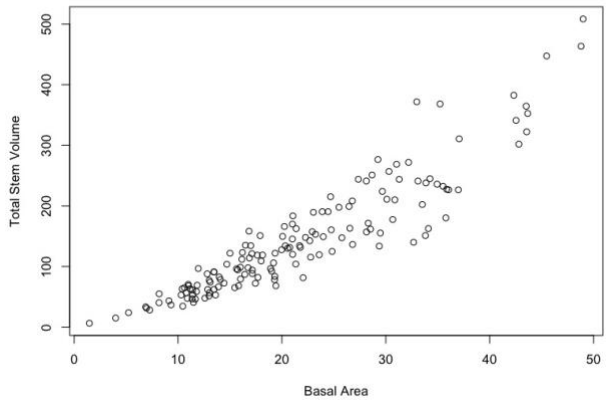


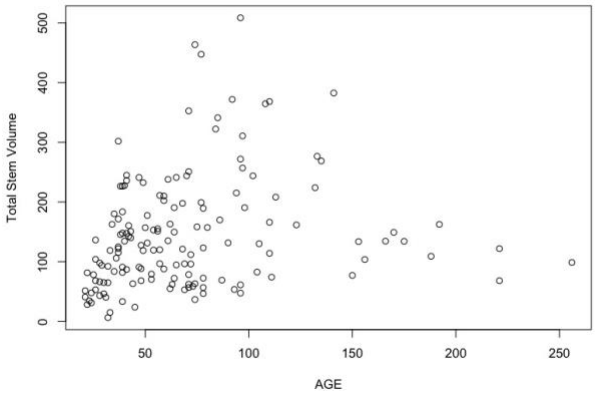


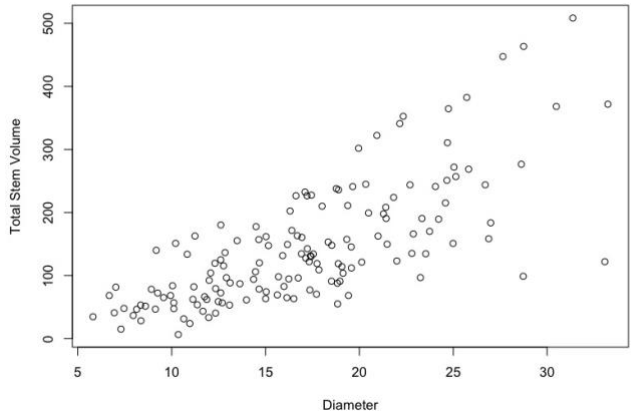


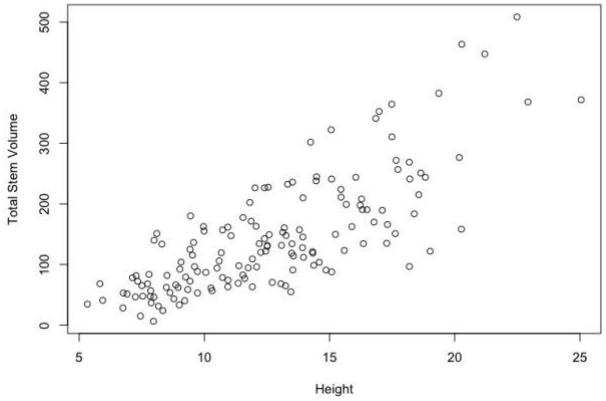


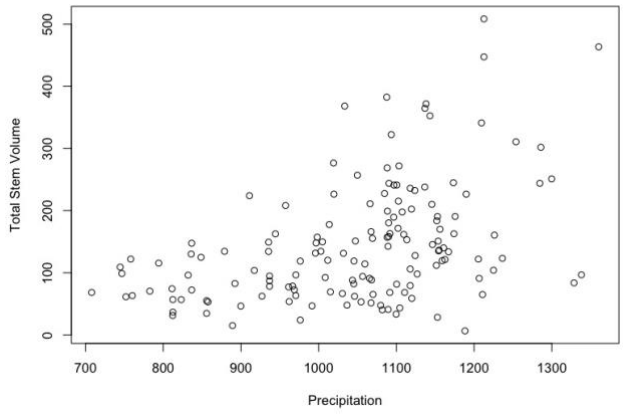
Scatter plots of total volume and other variables in modelling data set









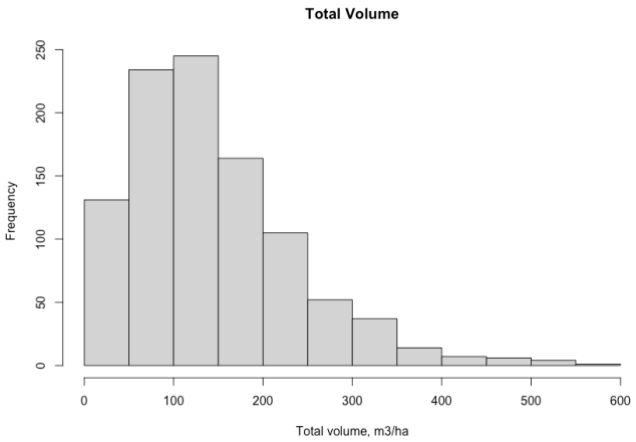


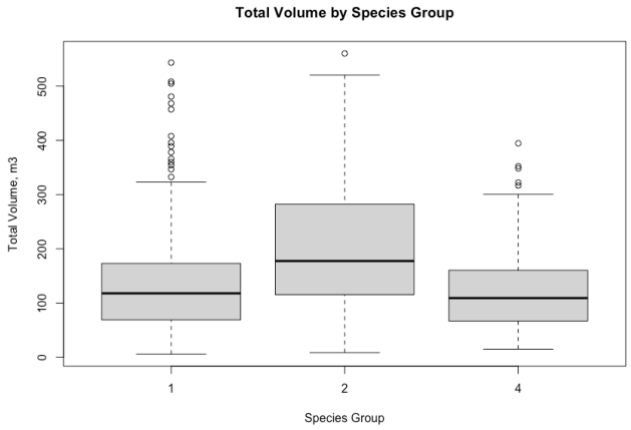
Validation dataset

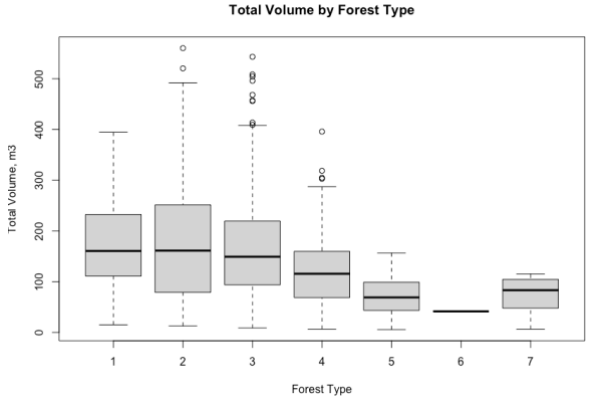
Summary statistics of validation dataset

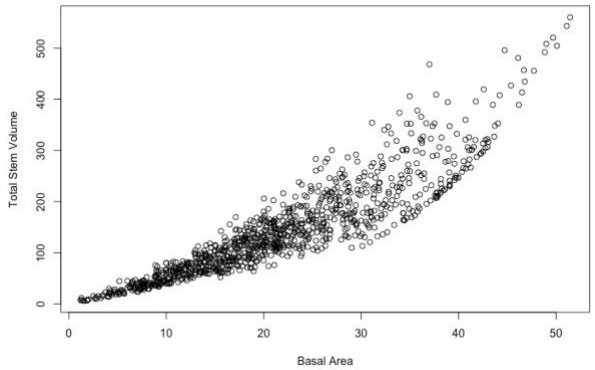
	AGE	P0	BA	D	H	VOLUME
Min	11.00	708.5	1.191	5.813	5.314	5.691
Max	337.00	1405.4	51.442	39.994	28.509	560.238
Mean	67.57	1045.4	21.567	17.191	12.638	143.360
Sd	43.94	133.77	10.217	6.028	3.939	92.111

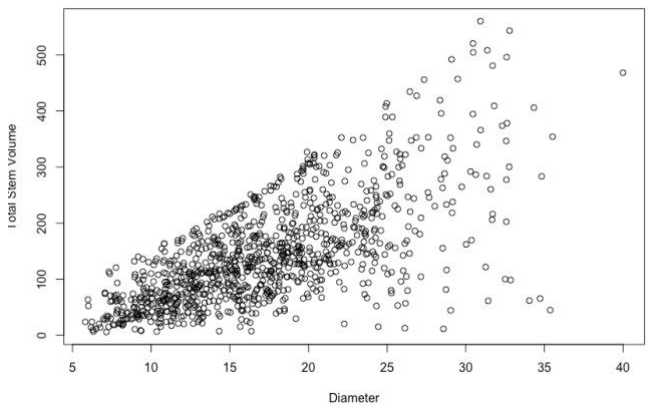
Number of plots = 1000

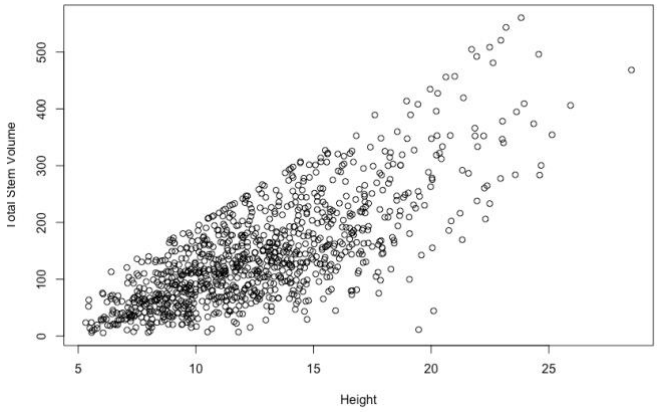












METHODS

The steps below were followed

- First step was to start with a model that included all the ratio scale variables since they are the suitable ones.
- The summary of that model revealed that only 3 independent variables (BA, H, D, & SP_GROUP) were significant.
- We dropped the non-significant variables and saved only 3 variables in a new object.
- We run a correlation analysis between the dependent variable and the 3 independent variables to determine its extent, to guide the selection of variables for the next models.
- The results indicated that TOTAL_VOLUME had the specified correlations with the three different variables: BA= 0.920752, H= 0.77428475, D= 0.73508882.
- Thereafter we began to trial different models with varying independent variables, their combinations, and transformations.
- In total about 20 different models were tried before arriving at the final 3. In each case, for each model, the R^2, p-value, plot of residuals and fitted values, together with the histogram of residuals were examined to arrive at the final shortlist. The suitable models had good attributes for the previously highlighted parameters.
- The chosen model was used to predict new volumes in the modelling dataset and these were compared to the reference volumes to assess how good the model was.
- Similarly, the selected model was then used to predict volumes for the validation dataset which was from a new area and contained even more plots (in this case 1000 compared to 150) in the first case.
- The newly modeled volume was then also compared with the reference volume in the validation dataset.
- Furthermore, the RMSE and Bias were determined to find out the accuracy and precision of the best choice model.
- **RMSE Equation:** sqrt(mean((obs.pred1$Modeled - obs.pred1$Original)^2))
- **BIAS Equation:** pbias(obs.pred1$Modeled, obs.pred1$Original)
- **BIAS Significance:** t.test(obs.pred1$Original,obs.pred1$Modeled,paired=TRUE)

RESULTS

Modelling results

The best 3 models

First model: Stand volume = $10^{0.351256 + 0.938934*\log(BA*H)}$

Second model: Stand volume = $10^{-0.344490 + 0.891245*\log(BA*D)}$

Third model: Stand volume = $(1.079893 + 0.551283*\sqrt{(BA*D)})^2$

Model Summary

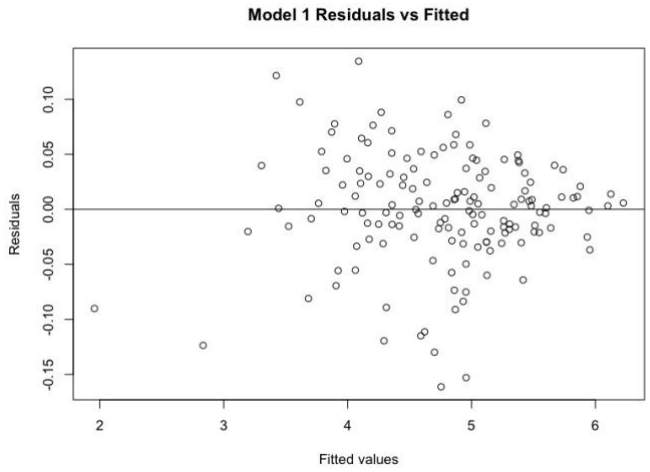
Model	R	R-Square	Adjusted R Square	Std. Error of the estimate
1	0.9973	0.9947	0.9947	0.05096
2	0.9914	0.9829	0.9827	0.09165
3	0.9923	0.9847	0.9846	0.4661

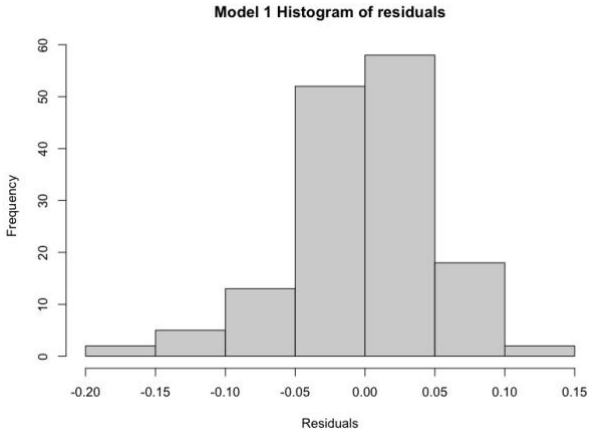
Coefficients

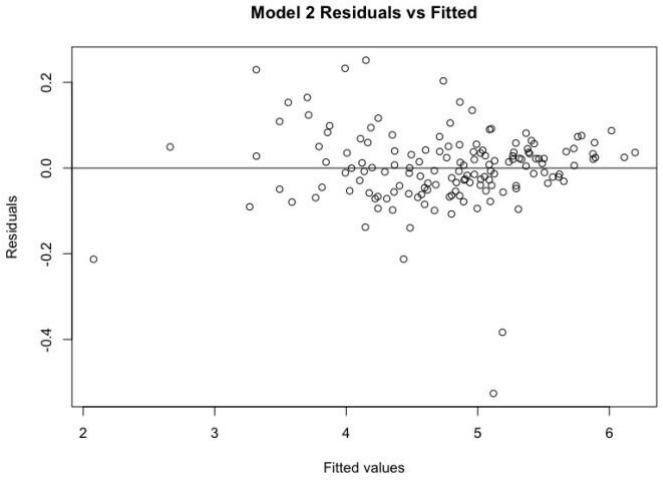
Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	-0.351256	0.030953	-11.35	.000
	log(BA*H)	0.938934	0.005633	166.68	.000
2	(Constant)	-0.344490	0.055927	-6.16	.000
	log(BA*D)	0.891245	0.009675	92.12	.000
3	(Constant)	1.079893	0.112773	9.576	.000
	√ (BA*D)	0.551283	0.005648	97.609	.000

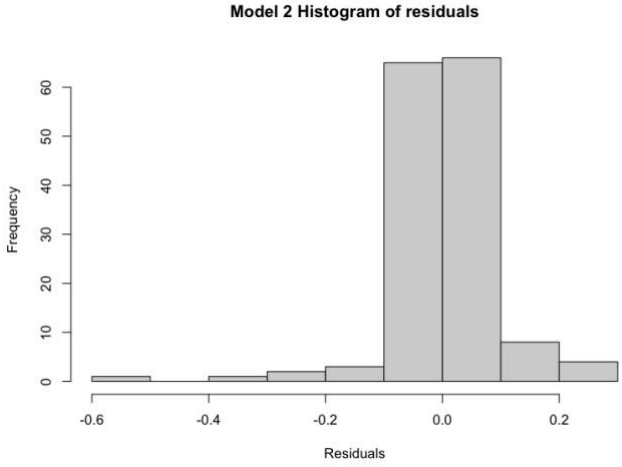
Note: Figures, where your independent variables are plotted against the dependent variable have been shared in the MATERIALS section of the report.

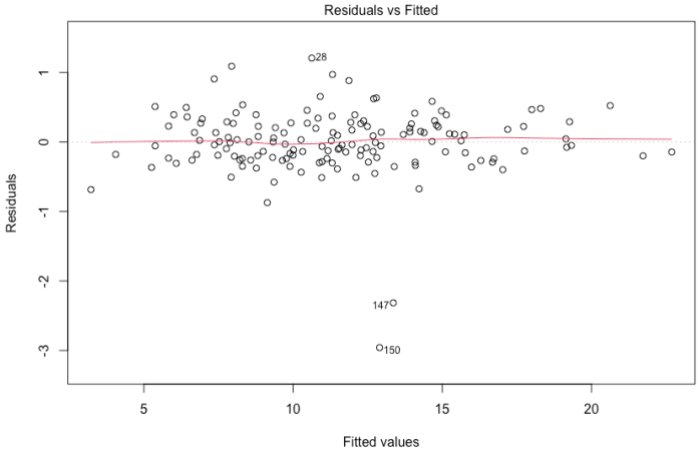
Figures showing the model residuals

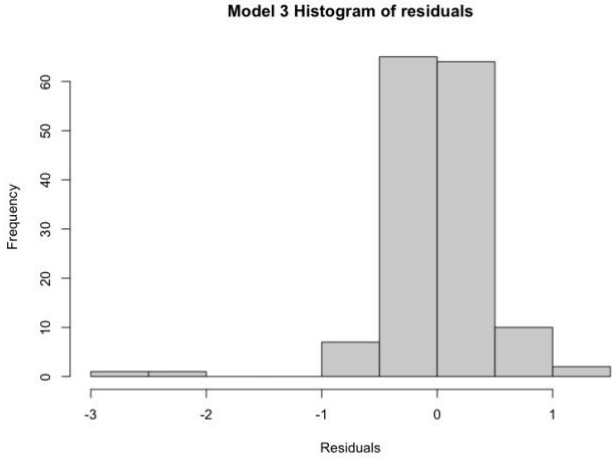






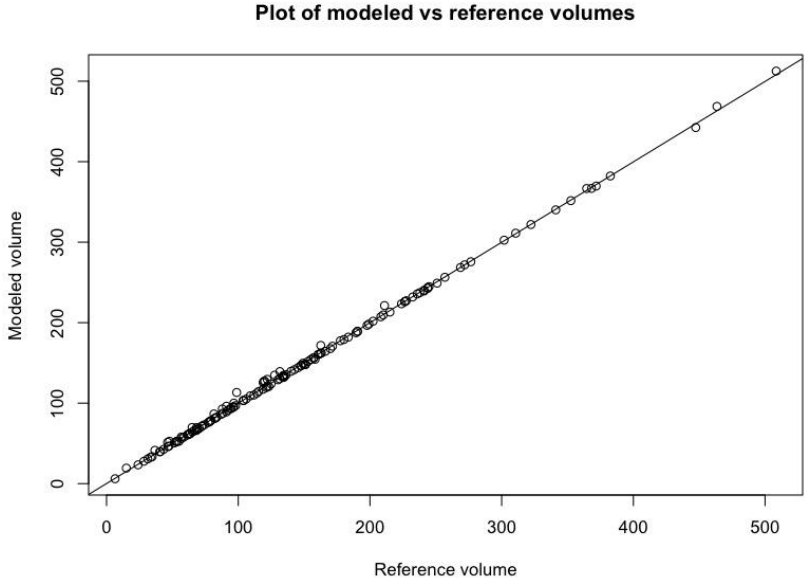






Model 3 is selected as the best model and used in the next stages of the report

Stand volume = $(1.079893 + 0.551283 * \sqrt{(BA*D)})^2$

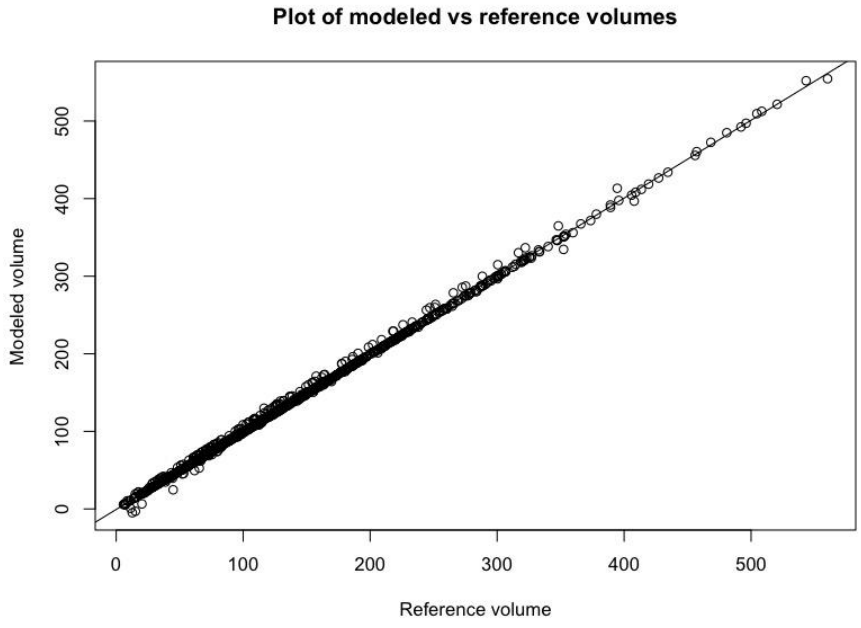


Validation results

RMSE: 3.371825

BIAS: 0.1

With t-test: t = 1.312, df = 999, p-value = 0.1898, Since p-value is not less than 0.05, we conclude the bias is not significant



DISCUSSION

Initially, it was difficult to decide on which variables to include in the dataset because all those at the ratio scale would seem justifiable at face value, for example, one could consider that amount of precipitation or age would be a suitable predictor. However, this was addressed after trying out a first model with all the ratio scale variables and only the significant ones were viable for further consideration in the analysis.

The model development process was rigorous, and it was difficult to arrive a t a desirable one based off only a single independent variable. This was partly solved through combinations with other variables following guidance from aspects like correlation analysis, R^2, values, and significance. The final model was confirmed after comparing various exponential and quadratic transformations.

There were a few outliers in the datasets, which could have resulted from either wrong measurements, or mistakes during data input. This affected the plots of residuals versus fitted values and made some histogram residual plots seem unsuitable. We could argue that for Model 2 and Model 3, the histogram of residuals suggested that the residuals (and hence the error terms) are normally distributed but have a few extreme outliers. The solution could be to create a new data set removing the rows with outliers or proposing verification of the observation to confirm whether it was an anomaly.

Several residuals versus fitted values plots were heteroscedastic and did not meet the criteria for a linear relationship. This was also addressed through trying out several complex transformations before arriving at constant variance.

While using the R software, there were errors while performing some operations such as the RMSE calculations and the graphical correlation plot. This was solved by installing new packages that enable such analysis to be conducted.