NRE5585
Final Captone Project Report
Statistical Analysis of Tree Height Estimation
Student Name: Hei Yee Lau

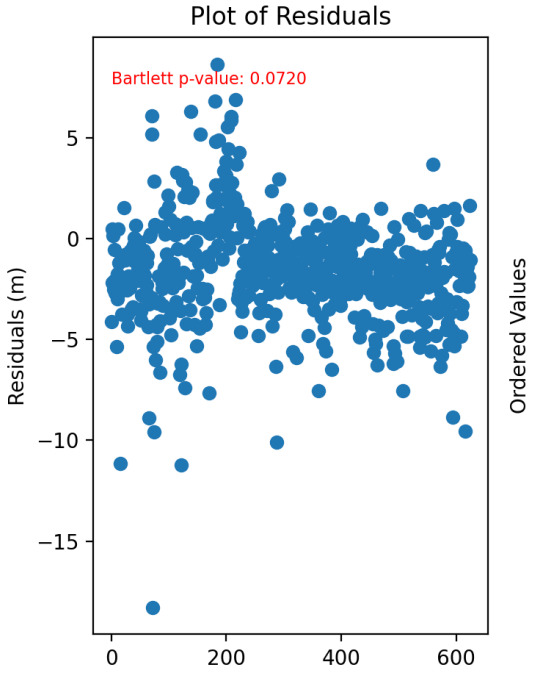
Introduction

The comprehensive analysis aims to evaluate the accuracy of a tree height model based on remote sensing data by comparing it to field measurements. The "Height_validation_data.xls" file contains essential attributes, including LiDAR-derived heights ('LiDAR_Ht'), field measurements('Field_ht'), tree speies('Spp'), and speies types('SppType'). The analysis consists of four main parts, each addressing specific aspects of the evaluation process.

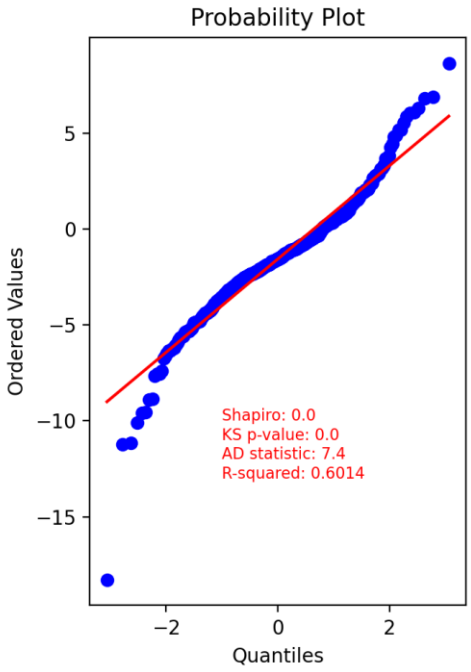
Tests of Equal Variance and Normality

- Residual Analysis
    - **Scatterplot of Residuals**

      In the script, a scatterplot of residual ('LiDAR_Ht – Field_Ht') is presented in Figure 1. The x-axis represents the position of data points in the dataset. The scatter plot visualizes the spread of residuals between 0 and -5. This range indicates the variability or differences between the LiDAR-derived heights and the corresponding field measurements. Additionally, the Bartlett test for equal variance is conducted for each tree species with at least 24 members. We can see that the p-value of Bartlett test is 0.0720, which is larger than the significant level (0.05), it suggests that there is not enough evidence to reject the null hypothesis and conclude that there is equal variance among the groups. The resulting p-value is displayed on the scatterplot in red, indicating statistical significance.



- **Quantile-Quantile Plot with Normality Tests**
  In figure 2, it displays the Quantile-Quantile plot of residual, assessing their normality. Normality tests, including Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling, are conducted. The p-values and statistics are annotated on the plot in red for interpretation.

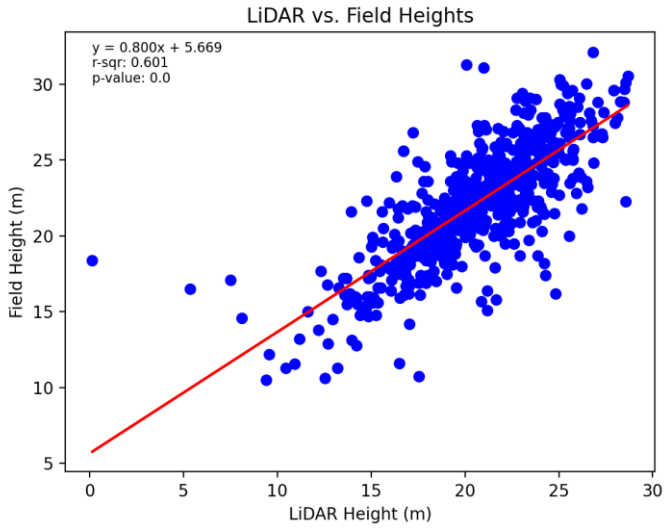


In the above graph, we can see most of the data points do not lie on the diagonal line. The Shapiro-Wilk Test has the p-value 0.0. For the null hypothesis is the data follows a normal distribution, with a p-value of 0.0, we would reject the null hypothesis. This suggests that the data does not follow a normal distribution.
For Kolmogorov-Smirnov Test, similar to the Shapiro-Wilk test, a p-value of 0.0 leads to the rejection of the null hypothesis, The data deviates significantly from a normal distribution.
For Anderson-Darling Statistic is equal to 7.4. An R-squared value of 0.6014 suggests that approximately 60.14% of the variability in the data is explained by the normal distribution.
In summary, the low p-values from Shapiro-Wilk and Kolmogorov-Smirnov tests, along with a potentially high Anderson-Darling statistic, indicate that the data does not follow a normal distribution. The R-squared value suggests a moderate fit of the data to the normal distribution in the Q-Q plot.

Linear Regression Analysis

- **Regression Analysis**

  A linear regression model is fitted to the data, and the results are visualized in Figure 3. The scatterplot include the actual data points (blue) and the regression line (red). The regression equation, R-squared value, and p-value are annotated on the plot.



The equation representing the regression model is y = 0.800x + 5.669. In this case it suggests that on average, for every unit increase in LiDAR_Ht ('x'), the prediction Field_Ht('y') will increase by 0.800 units, and the model intercepts the y-axis at 5.669.

R-squared measures the proportion of the variance in the dependent variable (Field_Ht) that is explained by the independent variable (LiDAR_Ht). An R-squared of 0.601 means that approximately 60.1% of the variability in Field_Ht can be explained by the linear regression model using LiDAR_Ht.

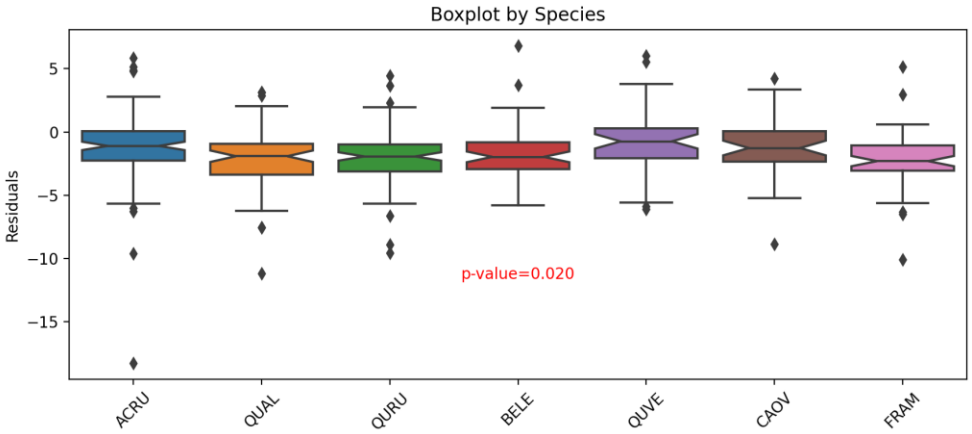
The p-value associated with the regression model tests the null hypothesis that the slope of the regression line is zero. A p-value of 0.0 indicates that we would reject the null hypothesis, suggesting that there is a significant relationship between LiDAR_Ht and Field_Ht.

In the graph, we can see that most of the data points lie on the upper of the diagonal line, which suggests that a positive correlation between LiDAR_Ht and Field_Ht. In other word, as LiDAR_Ht increases, Field_Ht tends to increase as well.

The positive slope in the regression equation indicates a positive linear relationship between LiDAR_Ht and Field_Ht. The R-squared alue of 0.601 suggests that the model explains a moderate proportion of the variability in Field_Ht. The p-value of 0.0 reinforces the statistical significance of the relationship.

Boxplots

- **Residuals by Species and Species Type**
  Boxplots are generated to visualize the distribution of residuals by tree species and species type:
  - **Boxplots by Species**
    For species with a sample size larger than 30, side-by-side boxplots are created. The appropriate test (ANOVA) is selected based on the normality assessment from Part 1. The test results are displayed on the boxplot, indicating significant differences between species.



In the side-by-side box plot, we can see that all the boxes have similar box size and median, indicate that the central tendency and the spread of residuals are relatively consistent across the different tree species.

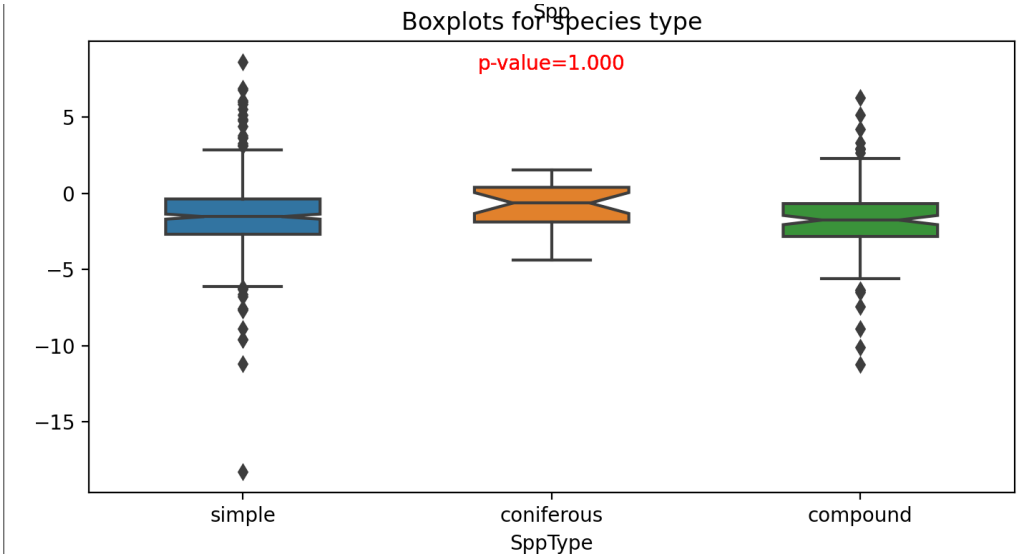
The p-value associated with the boxplot typically results from a statistical test comparing the central tendencies of the groups. Given the p-value of 0.020:

Null Hypothesis: There is no significant difference in the central tendencies of residuals between different tree species.

Alternative Hypothesis: There is significant difference in the central tendencies of residuals between at least two tree species.

With a p-value of 0.020, we have enough evidence to reject the null hypothesis at a conventional significant level (0.05). This suggests that there is evidence to support the claim that the central tendencies of residuals differ among tree species. The similar sizes and medians among the boxes imply that, although there are statistical differences, these differences may not be practically significant or may not manifest in a visually noticeable way in the box plot.

- **Boxplot by Species Type**
  Boxplot for compound, simple, and coniferous species types are presented. Similar to the species boxplots,



In the above box plot, we can see that three of them have the similar box sizes indicate that the interquartile ranges (IQR) of residuals are comparable. And we can see that there are a lot of outlier for simple species types and compound species types suggesting variability in the residuals, potentially driven by outlier.

The p-value associated with the boxplot likely results from a statistical test comparing the central tendencies of the groups (species types) for residuals. A p-value of 1.00 suggests that there is no evidence to reject the null hypothesis. The null hypothesis might stats that there is no significant difference in the central tendencies of residuals between the three species types.

The p-value of 1.00 indicates that there is no statistically significant difference in the central tendencies of residuals among the three species types. In other words, based on this analysis, there is no evidence to suggest that the mean or median residuals significantly differ between compound, simple, and conifer species types. The similar box implies consistent spread in residuals among the species types. However, the presence of numerous outliers, particularly for simple and compound types, may indicate variability or extreme value that are not well captured by the central tendency measures.

Test for Differences Among Residuals of Each Species

- **Pairwise tests of residuals**
  It is conducted for each species, utilizing either the 2-sample t-test or Wilcoxon rank-sum test based on normality. The resulting p-values are organized into a table, providing a clear overview of the significant differences between species.

Pairwise 2-sample tests by species

	ACRU	QUAL	QURU	BELE	QUVE	CAOV	FRAM
ACRU	1	0.008	0.006	0.054	0.205	0.905	0.001
QUAL	0.008	1	0.918	0.624	0.002	0.031	0.548
QURU	0.006	0.918	1	0.664	0.001	0.026	0.452
BELE	0.054	0.624	0.664	1	0.008	0.093	0.256
QUVE	0.205	0.002	0.001	0.008	1	0.303	0.000
CAOV	0.905	0.031	0.026	0.093	0.303	1	0.008
FRAM	0.001	0.548	0.452	0.256	0.000	0.008	1

From the above result, the diagonal elements have a p-value of 1. This is expected since we are comparing a species to itself, and the result should be 1. The off-diagonal elements contain p-values resulting from the comparison of different species. The values seem to vary, suggesting differences in the residual between these species.

The significance of the comparisons is often determined by the p-vale. A p-value less than the significant level (0.05) is considered statistically significant between the two species being compared.

A smaller p-value suggests stronger evidence against the null hypothesis, indicating a more significant difference between the compared species. In the table, QUVE and FRAM have the lowest p-value (0.00) in the pairwise 2 sample t-test, which shows that QUVE and FRAM have the most significant difference between them. ACRU and FRAM, QURU and QUVE also have the smaller p-value (0.01) which show the significant differences between them.

Overall, this type of analysis can help us to identify which species exhibit significantly different residuals, indicating potential differences in the variable being measured.

Conclusion

Our thorough analysis of the tree height estimation model based on remote sensing data revealed several key findings. Residual analysis exposed challenges of unequal variance and non-normality, as indicated by Barlett and normality tests. The linear regression model demonstrated a significant positive correlation between LiDAR-derived heights and field measurements, emphasizing its explanatory power. While boxplots indicated statistical difference in residual among tree species, the practical significance of these differences remained uncertain due to the visual similarity in box sizes and medians Additionally, boxplots by species types showed no significant differences in central tendencies, though the presence of outliers raised considerations. Pairwise test provided valuable insights into specific species exhibiting distinctive residuals. This comprehensive assessment offers a nuanced understanding of the model's strengths and limitation, suggesting avenues for further refinement and exploration of factors influencing tree estimation accuracy.