# Project 1: Moisture Content of Tree Branches

Yingxue YU, Guanqun LIU, Yulun Jiang

2022-03-31

## I. Introduction

This dataset is originally collected and used for J. Joseph McDermott's botanical research on the effect of cutting methods have on the moisture content of tree branch samples in 1941. In experiments, researchers use twig segments to confirm the moisture content of woody structure. If the sample is simultaneously cut at both ends, then the release of tension in both direction will not result in instantaneous water removal; whereas cutting at one single end will lead to rapid water loss in the vicinity of cut.

Grounded on this fact, we conduct a statistical analysis on validating how different cutting methods will affect the extent of water removal given different cut samples. We intend to find an estimation model of how each factor given in the dataset and their interaction terms with cutting methods are related to the mass of moisture content. We will carry out an exploratory data analysis in section II to examine and preprocess the data. Then we will process our data modeling and assessment results in section III, IV, and V. Finally we will conclude our analysis and which part in the original analysis should be criticized.

## II. EDA

### 2.1 Data Validation

The dataset includes 120 branch samples, includes 4 variable columns: **species, branch/species, location/branch**, and **transpiration** . The moisture content in the last column is expressed in $10 \times \%$ of its dry sample weight. Under each species, for each possible combination of cutting location (Location/Branch) and transpiration type(Transpiration), we have 5 sample measurements of moisture contents, which also reveals that the dataset is balanced. We will later exclude the **branch/species** column because it works as an indicator of the former combination. A variance summary table is provided below. In "Location/Branch", "central" refers to simultaneous cut at both ends, "distal" and "proximal" refer to single cuts at branch-terminal end and tree-side respectively. We will keep the numerical values of the classes for analysis consistency and transform the response scale to $1 \times \%$ of dry sample weight.

| Var Name | Type | Category | Classes | Detail |
|---|---|---|---|---|
| **Species** | num | categorical | 4 | 1=Loblolly Pine<br>2=Shortleaf Pine<br>3=Yellow Poplar<br>4=Red Gum |
| **Location/Branch** | num | categorical | 3 | 1=Central<br>2=Distal<br>3=Proximal |
| **Transpiration** | num | categorical | 2 | 1=Rapid 2=Slow |
| **MoistureContent** | num | continuous | N/A | **Target Response** |

Table 2: Two-way Frequency Table of Categorical Variations

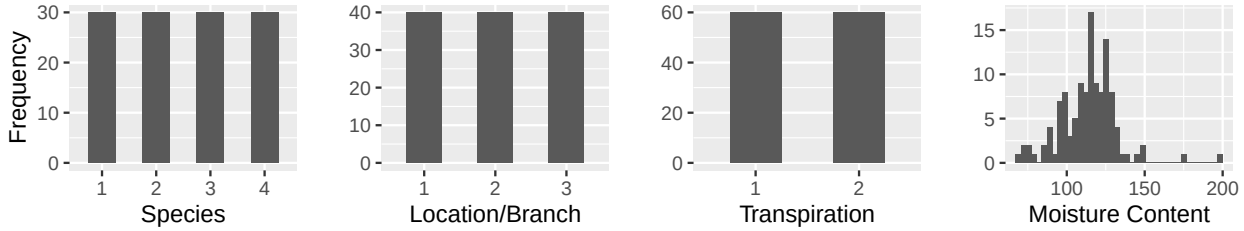
| S/L | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 1 | 10 | 10 | 10 | 30 |
| 2 | 10 | 10 | 10 | 30 |
| 3 | 10 | 10 | 10 | 30 |
| 4 | 10 | 10 | 10 | 30 |
| Total | 40 | 40 | 40 | 120 |

| S/T | 1 | 2 | Total |
|---|---|---|---|
| 1 | 15 | 15 | 30 |
| 2 | 15 | 15 | 30 |
| 3 | 15 | 15 | 30 |
| 4 | 15 | 15 | 30 |
| Total | 60 | 60 | 120 |

| L/T | 1 | 2 | Total |
|---|---|---|---|
| 1 | 20 | 20 | 40 |
| 2 | 20 | 20 | 40 |
| 3 | 20 | 20 | 40 |
| Total | 60 | 60 | 120 |

### 2.2 Univariate Analysis

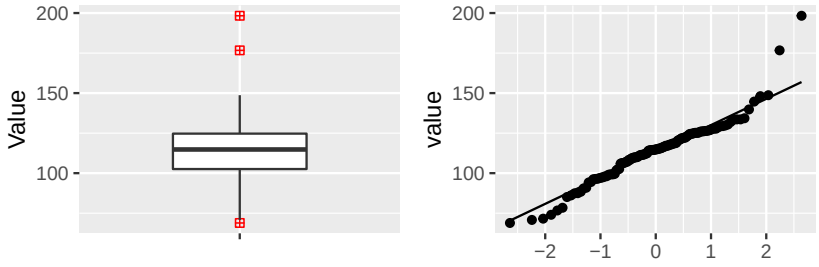
Fig. 1: Frequency Count of Variates




From the histograms above, we can see that for "Species", "Location/Branch" and "Transpiration", the variety in each variate is equally distributed. For the continuous response variable, we expand its value distribution analysis (Fig. 2). The distribution of moisture content is slightly left-skewed, with more suspected extreme outliers at higher percentiles. 50% of samples fall into the interval $[102.5, 124.7]$. From the Q-Q plot, we can corroborate that the distribution is more left skewed. The result of Kolmogorov-Smirnov test also rejects the null hypothesis of its normality.

Fig. 2: Basic Statistics of Response Variate

| | Moisture Content |
|---|---|
| *Min.* | 68.9 |
| *1st Qu.* | 102.575 |
| *Median* | 114.8 |
| *Mean* | 113.7233 |
| *3rd Qu.* | 124.7 |
| *Max.* | 198.3 |
| *IQR* | 22.125 |
| *SD* | 18.5918 |




### 2.3 Bivariate/Multivariate Analysis

Between categorical variables, we use two way tables (Table 1) to capture the frequency under each combined category. Frequencies of all categories under each pair of categorical variables are equal. Based on these tables, we conduct $\chi^2$ tests and the result (Table 2) shows that all any two categorical variables are independent to each other. As for the categorical variables with the continuous target response (moisture content), we draw side-by-side boxplots of each pair and conduct the ANOVA analysis. ANOVA confirms the significance of category mean difference of **species**(tree species), **location/branch** (cutting method), **transpiration**(transpiration), and the interaction term of species and transpiration at 0.001 level. This shows an apparent difference from the original paper that **location/branch** and **species** are at 0.001 level; **transpiration**, the interaction term of cutting method and species, and the interaction term of all three variables are at 0.01 level; and the interaction term of transpiration and species is at 0.05 level.

Table 3: Chi-square Test Result

| X2.test | X.squared | df | p.value |
|---|---|---|---|
| Species/Location | 0 | 6 | 1 |
| Species/Transpiration | 0 | 3 | 1 |
| Location/Transpiration | 0 | 2 | 1 |

Fig. 3: Boxplots

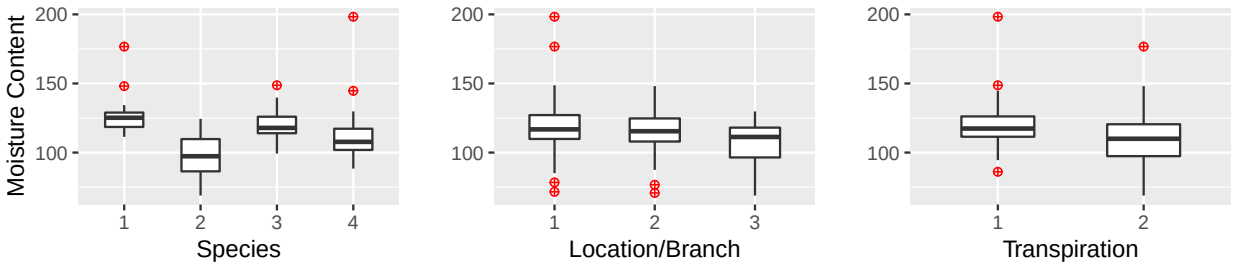



Table 4: One-way/Two-way ANOVA Test Result (Comma - Interaction Term)

| ANOVA | df | meanSq | Fvalue | signif.level |
|---|---|---|---|---|
| Species | 3 | 4775 | 31.729 | 0.001 |
| Loation | 2 | 1272 | 8.458 | 0.001 |
| Transpiration | 1 | 3942 | 26.194 | 0.001 |
| S:L | 6 | 79 | 0.523 | 1.000 |
| S:T | 3 | 1020 | 6.776 | 0.001 |
| L:T | 2 | 352 | 2.338 | 1.000 |
| S:L:T | 6 | 272 | 1.811 | 1.000 |

## III. Model Fitting

## IV. Assessment

## V. Plots

## VI. Conclusion