Project 1: Moisture Content of Tree Branches

Yingxue Yu, Guanqun Liu, Yulun Jiang

2022-04-08

1. Introduction

This dataset is originally collected and used for J. Joseph McDermott's botanical research on the effect 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 structures. If the sample is simultaneously cut at both ends, then the release of tension in both directions will not result in instantaneous water removal; whereas cutting at one single end will lead to rapid water loss in the vicinity of the 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 is 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 2 to examine and preprocess the data. Then we will process our data modeling and assessment results in section 3. Finally, we will conclude our analysis and which part of the original analysis should be criticized.

2. 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 cuts at both ends, "distal" and "proximal" refer to single cuts at the 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 the dry sample weight.

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

2.2 Univariate Analysis

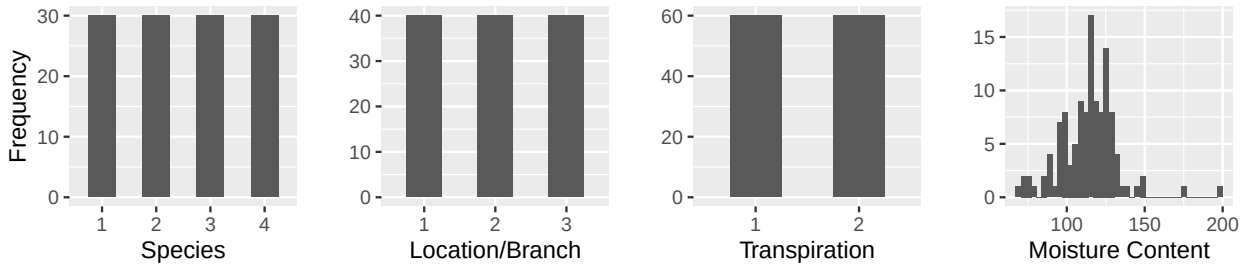


Figure 1: Frequency Count of Variates

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 the Kolmogorov-Smirnov test also rejects the null hypothesis of its normality.

	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

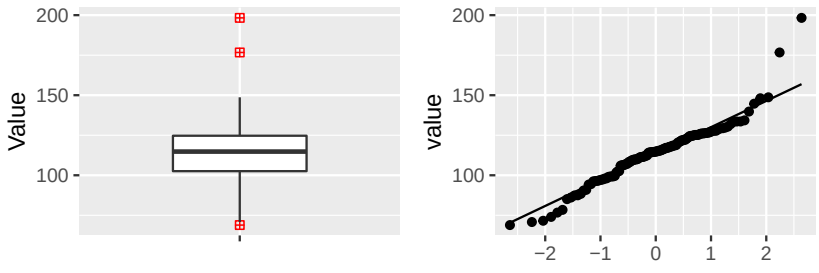


Figure 2: Basic Statistics of Response Variate

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

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

variables are equal. Based on these tables, we conduct χ^2 tests and the result (Table 2) shows that any two categorical variables are independent of each other. As for the categorical variables with the continuous target response (moisture content), we draw side-by-side boxplots (Table 3) of each pair and conduct the ANOVA analysis. ANOVA confirms the significance of the 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

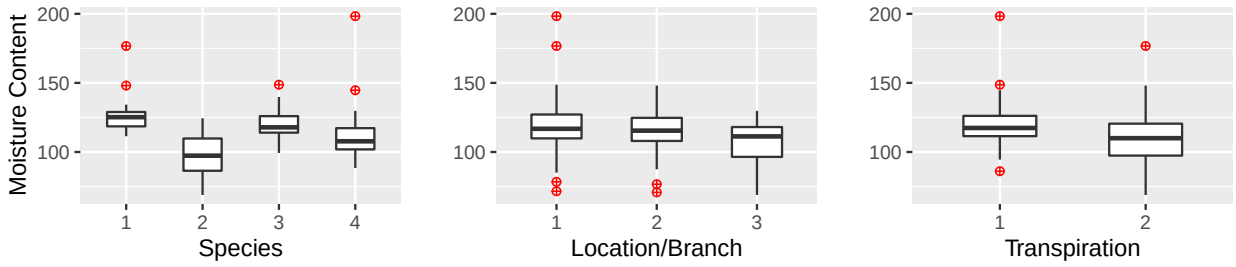


Figure 3: Boxplots by Different Factor Levels

Table 4: One-way/Two-way ANOVA Test Result

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

3. 2-way ANOVA

3.1 Model Fitting

To test the effect of the method of cutting, which is indicated by location variable, on the moisture contents, an analysis of variance is made. Model is chosen by performing stepAIC. Three runs of stepAIC are performed: a forward selection starting from intercept, a backward elimination from all first order factors and second order interactions, and a forward-backward search that starts from all first order factors. The same model is selected by all 3 runs, which is

$$\begin{aligned} moisture = \beta_0 &+ \beta_1 \cdot ShortleafPine + \beta_2 \cdot YellowPoplar + \beta_3 \cdot RedGum + \beta_4 \cdot Distal \\ &+ \beta_5 \cdot Proximal + \beta_6 \cdot Slow + \beta_7 \cdot ShortleafPine \cdot Slow \\ &+ \beta_8 \cdot YellowPoplar \cdot Slow + \beta_9 \cdot RedGum \cdot Slow \\ &+ \beta_{10} \cdot Distal \cdot Slow + \beta_{11} \cdot Proximal \cdot Slow \end{aligned}$$

Table 5: ANOVA Table for Final Model

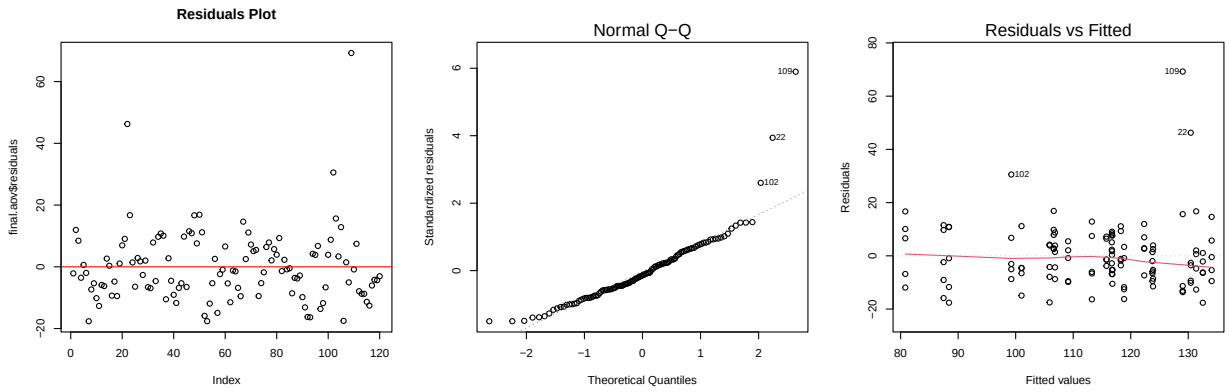
ANOVA	df	meanSq	Fvalue	signif.level
Species	3	4775	31.15	0.001
Loation	2	1273	8.30	0.001
Transpiration	1	3942	25.71	0.001
Species:Transpiration	3	1020	6.65	0.001
Location:Transpiration	2	352	2.30	1.000
Residuals	108	153	NA	NA

The Least Square means of the fitted model is 153. The ANOVA table can be found in Table 5. Values for βs are provided in the Appendix. All other terms in the formula are indicator variables. For example, a sample taken at distal location of a branch of Shortleaf Pine with slow transpiration would have terms $ShortleafPine, Distal, Slow$ equal to 1, and all other terms equal to 0.

3.2 Model Assessment

Our model makes the following assumptions on data:

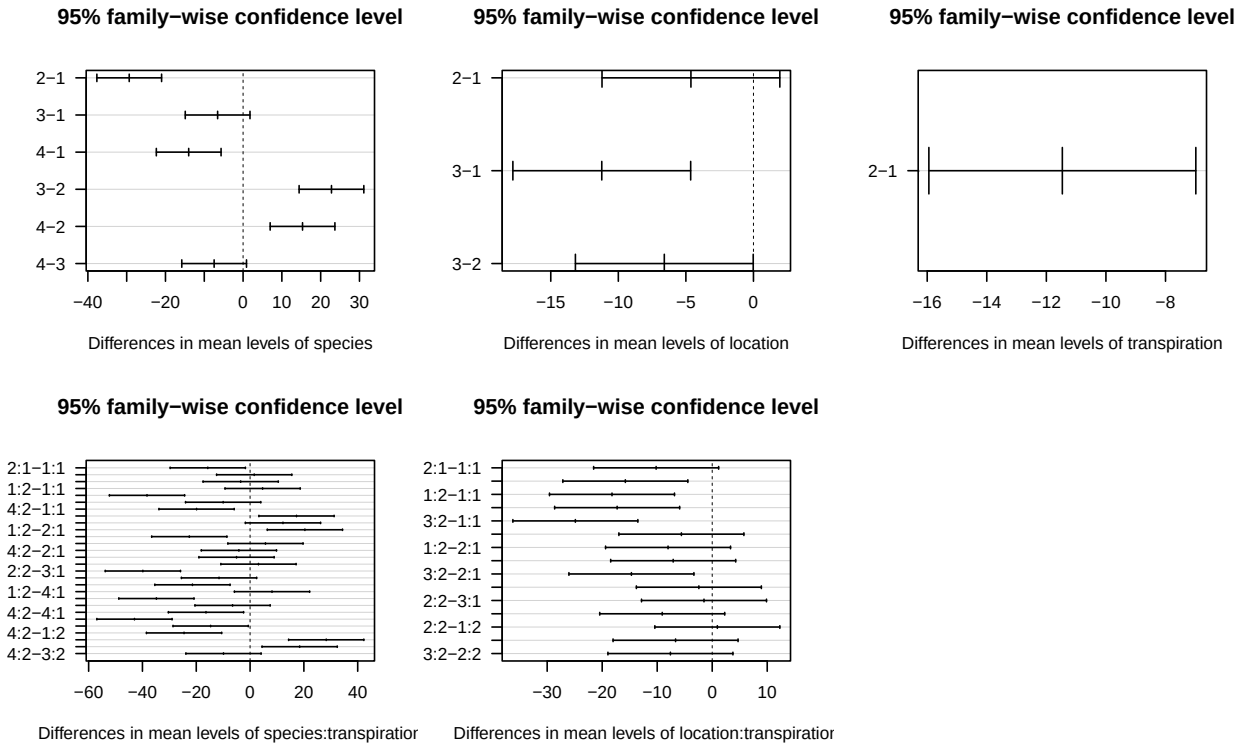
- Errors have mean 0.
- Errors are homoscedastic (same variance).
- Errors are uncorrelated.
- Errors are normally distributed.



Above we first plot the residuals and the fitted regression lines. We can clearly observe the fitted line has a mean value of 0 and a slope of 0, indicating that errors are centered around 0 with the average error being 0. We then show the QQ plot of the residuals on the second figure. It indicated that the normality may be violated due to the clear deviation on the right tail. However, we further conduct Shapiro-wilk test and accept its normality with a p-value $p = 5.2e - 10$. We plot the Residuals vs Fitted plot on the third figure to examine the homoscedasticity assumption. The fitted curve is nearly a straight line with slope 0, supporting the homoscedasticity property.

3.3 Interpretation : post-hoc test

The Tukey's Honestly-Significant-Difference (TukeyHSD) test enables us to know which groups are different from one another. The following output shows the pairwise differences 95% confidence interval between the 5 types of varieties and 7 types of fusarium strains : If the interval does not include zero then the difference is significant.



From the post-hoc test results, we see that there are statistically significant differences (p-value < 0.05) between the following groups

- species : groups 1-2, 1-4, 2-3, 2-4
- location : groups 1-3
- species:transpiration: 2:1-1:1, 2:2-1:1, 4:2-1:1, 2:1-2:1, 2:2-2:1, 2:2-3:1, 4:2-3:1, 2:2-4:1, 4:2-4:1, 4:3-4:1, 3:2-1:2, 3:3-1:2, 3:2-2:2, 3:3-2:2
- location:transpotation: 3:1-1:1, 1:2-1:1, 2:2-1:1, 3:2-1:1, 3:2-2:1

To conclude, the Tukey post-hoc test revealed a statistically-significant difference between the species 2 (Shortleaf Pine) and all the others and between the location 1 (Central) and 3 (Proximal).There are also significant difference between groups different in the combination of species and transpiration, as well as the location.

4. Conclusion

To compare the tree moisture in different types of species, locations and transpiration statistically, we implement a 3-way ANOVA following such sequence: data preliminary exploration, model assumption checking, model selection, and fitting, model assessment, and interpretation. In conclusion, we find a statistically-significant the difference in the tree moisture by species, location, transpiration, and combination of species with transpiration, the combination of locations with transpiration. However, due to the nature of the given

dataset, such as the small number of observations, we can't statistically conclude if the interaction between these terms is significant or not. To solve this issue, more data is needed to reach a credible result.

Appendix

Here are the coefficients for the final model.

β_0	132.53
β_1	-15.72
β_2	1.54
β_3	-3.49
β_4	-10.19
β_5	-15.79
β_6	-2.09
β_7	-27.25
β_8	-16.20
β_9	-21.05
β_{10}	11.13
β_{11}	9.13