Group number 48. Rongjun Gao, Wanchai Grossrieder, Niels Willems

Analysis of potential factors influencing the moisture content of trees

1. Introduction

The moisture content is an index to evaluate the amount of water in the xylem elements under tension. According to the study in the original literature, four factors have been suggested to influence the moisture content: the species, branch, location, and transpiration. The species correspond to the species of the tree: four of them have been selected: Loblolly pine, Shortleaf pine, Yellow poplar, and Red gum. For each of these trees, there are five branches selected with 3 locations on them: central, distal, and proximal. The last factor considered is the transpiration type, it can either be rapid, due to some hot, dry, and sunny conditions, or it can be slow. Due to the cool, moist, cloudy conditions. Our study aims to propose a model that describes the dataset best using ANOVA and find the most significant variance.

2. Exploratory Data Analysis

As a start, we plot the different variables of the dataset by doing exploratory data analysis. For each variables we check the number of data point we have (Table 1 and Figure 1). We also collected the mean value and standard deviation of moisture under each variable, as well as the Pearson correlation coefficients between moisture and variables (Table 2). It is shown that “Species”, “Location”, and “Transpirations” are negatively correlated to moisture while the “Branch” is slightly positive-related.

Table 1 Univariate numerical and graphical, number of data points per variable

Species	From 1 to 4, 30 data points for each species
Branch	From 1 to 5, 24 data points for each branch type
Location	From 1 to 3, 40 data points for each location
Transpiration	From 1 to 2, 60 data points for each transpiration type

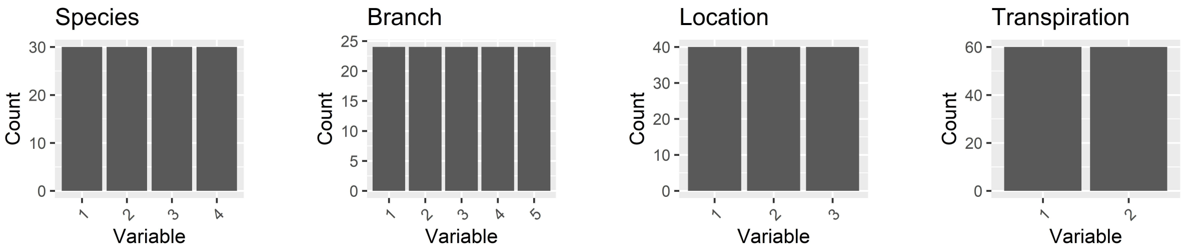


Figure 1. Counts per variables, all the variables have the same number of points.

Table 2 Bivariate numerical and graphical, mean and sd of moisture for each variable

	Species 1	Species 2	Species 3	Species 4
Mean moisture	1262.033	968.600	1196.433	1121.867
Std moisture	121.9524	158.1319	106.1335	203.2128

	Branch 1	Branch 2	Branch 3	Branch 4	Branch 5
Mean moisture	1126.208	1119.875	1145.917	1196.083	1098.083
Std moisture	125.8432	171.2497	134.3519	298.1752	144.8570

	Location 1	Location 2	Location 3
Mean moisture	1190.075	1143.800	1077.825
Std moisture	231.1053	150.2388	151.5244

	Transpiration 1	Transpiration 2
Mean moisture	1194.550	1079.917
Std moisture	158.5941	194.6353

coef_corr	Species	Branch	Location	Transpiration
Moisture	-0.1163	0.0152	-0.2475	-0.3096

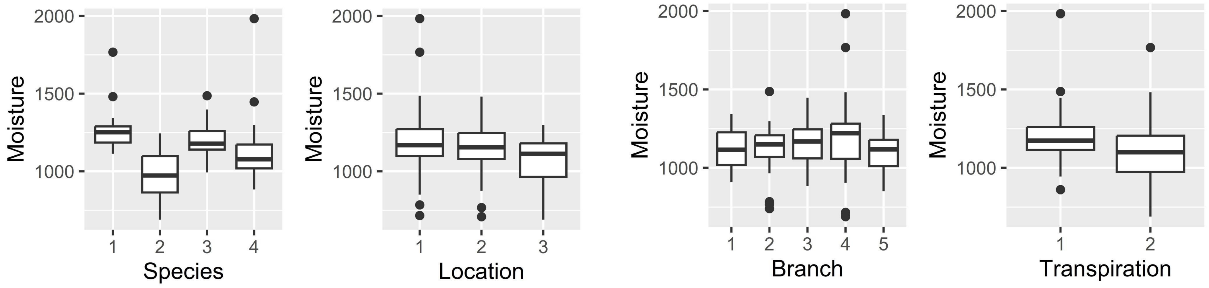


Figure 2. Box plot of the different factors influencing the moisture content. For each of them, we separated the plot into different values.

By comparing first the different boxplots for each of the factors (Table 2, Figure 2). From the plots, we can see that for the species, the moisture content has mean values that are different. On the other hand, it seems that the transpiration type or the location or the branches has a relatively poor effect on the moisture content as the mean across the different variables. While for the branch it seems that there is no effect. By investigating how the data is generated, we discovered that the branch variable represent several branches taken from the same tree. Therefore, we decided to discard this variable. Most of the confidence intervals are overlapping in the variables of the transpiration type. This is less the case for the branches and the locations on the branches. But more importantly, this is not the case for the species. This suggests that the moisture content mostly depends on the tree species.

To assess the feeling of the exploratory data analysis we perform an ANOVA. We define the α-level at 5%. Taken separately, it shows that the strongest factor on the moisture content is mostly the species type as predicted, for each species the p-value is significant and is far below the α level. From the table 1, we also conclude that the location of the branch, as well as the transpiration type, have a significant effect on the moisture content, as both are below the α-level set.

3. Model fitting and assessment

Model Fitting

We fit the model using Least Squares argument, where we take an arbitrary value of 0.05 as a threshold for our p-value (=Pr(>F)).

We can model the data using a regression as the moisture content to be linear to the variables representing species, location on the branch and transpiration type selected with some error term. Our regression looks like this:

$$Moisture \ = \ \beta_0 + \beta_1 * Loblolly \ + \beta_2 * Shortleaf \ + \beta_3 * YellowPolar \ + \gamma_1 * Proximal \ + \\ + \gamma_2 * Central + \psi_1 * Transpiration_2 + \epsilon$$

$\beta_0 \ = 1119.8, \ \beta_1 \ = 140.2, \beta_2 \ =- \ 153.3 \ , \ \beta_3 \ = 74.5 \ , \gamma_1 = \ 112.3, \gamma_2 = \ 65.0,$
$\psi_1 =- \ 114.6$

Where each variable is an indicator function (i.e, it can only take values in {0,1} based on if the property of the same name is verified). Note that we only need #df variables for each factor, and not #df+1, as we have a constant variable β_0.

Table 3. Analysis of variance of the data, Model 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Loblolly Pine	1	623002	623002	34.647	4.12e-08 ***
Shortleaf	1	726186	726186	40.386	4.53e-09***
YellowPoplar	1	83403	83403	4.638	0.033390*
Proximal	1	167535	167535	9.317	0.00283**
Central	1	87054	87054	4.841	0.02982*
Transpiration 2	1	394224	394224	21.924	7.95e-06 ***
Residuals	113	2031884	17981		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

A second model can be proposed that takes into consideration the combined effect of the different factors. This is summarized in Table 4.

Table 4 Analysis of variance of the data, Model 2
(unsignificant variable relations are not displayed to increase readability)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Loblolly Pine	1	623002	623002	41.395	4.86e-09***
Shortleaf	1	726186	726186	48.251	4.47e-10***
YellowPoplar	1	83403	83403	5.542	0.02061*
Proximal	1	167535	167535	11.132	0.00121**
Central	1	87054	87054	5.784	0.01809*

Transpiration_2	1	394224	394224	26.194	1.58e-06***
Loblolly*Transpiration_é	1	259972	259962	17.273	7.03e-05***
Proximal*Transpiration_2	1	68378	68378	4.543	0.03560*
Residuals	96	1444806	15050		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

We notice that all the previous factors are still relevant, and only take into account the factors with p-value less than 0.05. Therefore, the regression model that takes into account this interaction is:

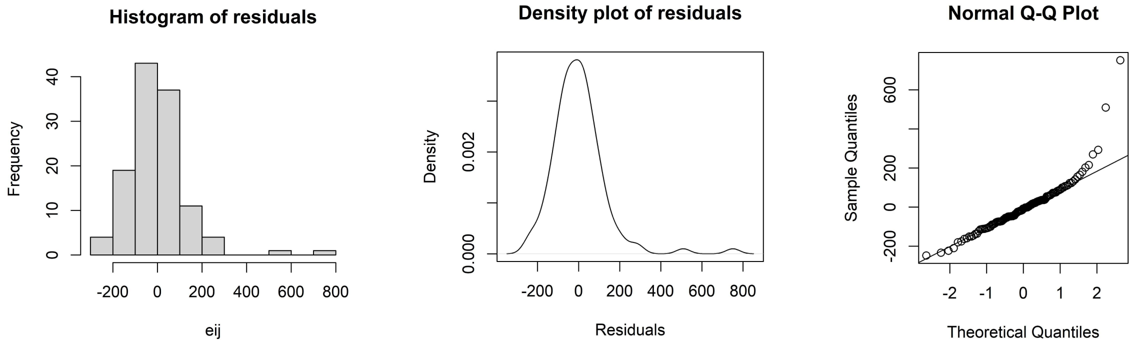
$$Moisture \ = \ \beta_0 + \beta_1 * Loblolly + \beta_2 * Shortleaf + \beta_3 * YellowPoplar + \gamma_1 * Proximal + \\ + \gamma_2 * Central + \psi_1 * Transpiration + \iota_1 * Loblolly * Transpiration + \\ + \ \iota_2 \, Proximal * Transpiration + \ \epsilon$$

Where $\beta_0 = 1115.0, \beta_1 = 81.0 \, , \beta_2 =- 143.6 \, , \ \beta_3 = 95.6 \, , \gamma_1 = \ 269.0, \gamma_2 =- 2.6 \, ,$

$\psi_1 =- 81.6 \, , \iota_1 = 78.4, \ \iota_2 =- 286.4$

Model Assessment

The histogram of the residuals shows that the mean of the residuals is around zero. The model has therefore an expectation of the error term of zero. This is also what shows the density plot of the residuals and the normal QQ-plot. (Figure 3) The distribution of the error terms is normal as the line on the plot is straight, even if it seems to have long tails. These are due to several observations that are outliers and present in particular at the top right of the QQ-plot. We can also see them around 600 on the histogram and density plot of the residuals. Moreover, the residuals seem randomly dispersed around 0 in both our models (see figures 4 and 5), which is evidence in favor of uncorrelatedness of error terms. To conclude, the models are reasonably good.



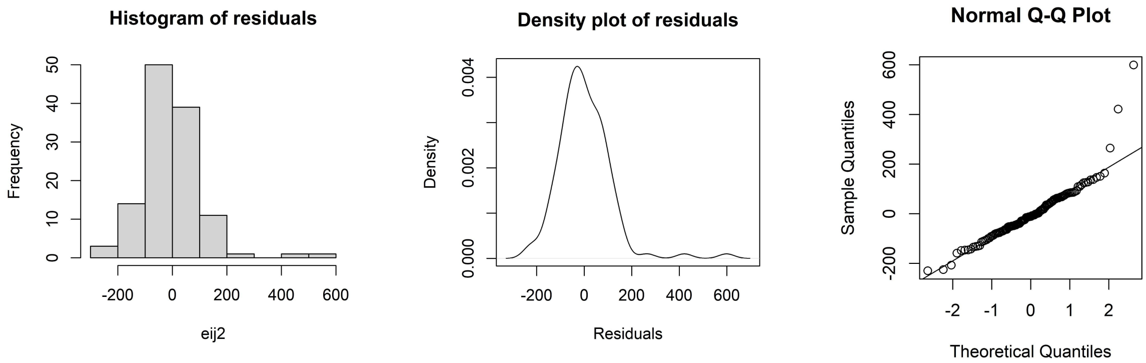
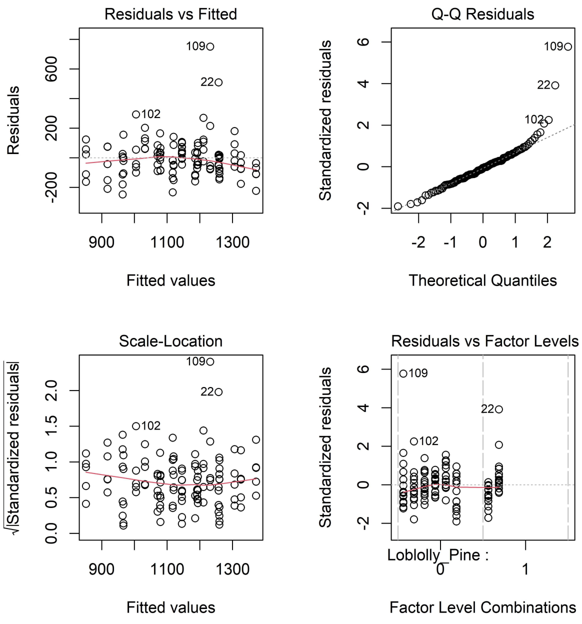


Figure 3 QQ normal plot of residuals, model 1 is at the top, model 2 is at the bottom.



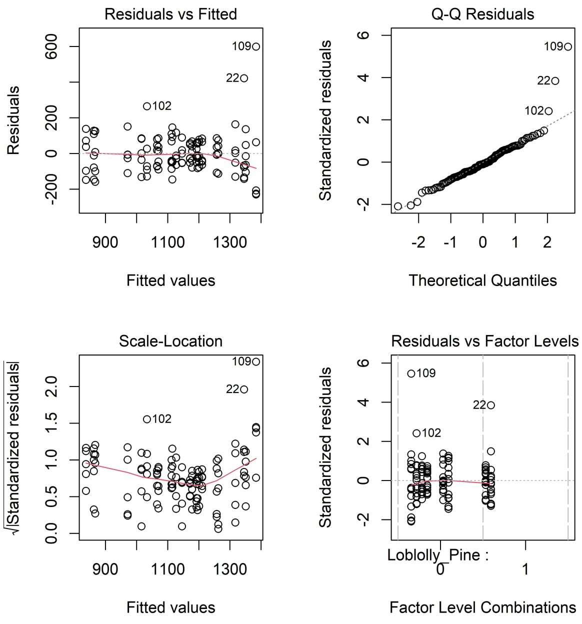


Figure 4. Assessment: QQ normal plot of residuals vs. fitted (Model 1)

Figure 5. Assessment: QQ normal plot of residuals vs. fitted (Model 2)

4. Final estimated model

We know describe how we select our final model : The residual vs fitted plot has a straight line, suggesting that a linear model is reasonable for both models. The scale-location plot should show a straight line, meaning that the residuals are randomly scattered around the line, suggesting that the model is homoscedastic. In this case, model 1 and model 2 have the same result suggesting no improvement in the homoscedasticity, the model 1 seems even better. Both models have the same QQ plots and the same outliers appearing. Also in general the residuals of both models are distributed around the zero for the residual vs factor level plots. This means that the model is representative of the relationship between the factors and the moisture. The spread of the points and the pattern is equally distributed around the line. This suggests that the model is reasonable. Hence, as the second model does not bring much information, the model retained is the model first one, with the regression.

5. Conclusion

To investigate the effects of species, locations, transpiration, and branches on moisture of trees, we first proposed a linear regression model, where the moisture depends on the variable listed. Intending to improve this model, we further tested Model 2 by including the interaction of combined factors. Among the combinations, the species-transpiration and location-transpiration interactions showed a low p-value and was added to our model. However, by plotting the QQ normal plots of residual/residual vs. fitted, we did not observe a significant difference between the two models, moreover both models shows good homoscedasticity. From residual vs factor level plot the model 1 is representative of the relationship between the factors and the moisture. Based on our analysis, the effect of the interactions do not have a large effect in Model 2. Thus, we would conclude that Model 1 performs well enough with a simpler equation, reminded here :

$$Moisture\ =\ \beta_0 + \beta_1 * Loblolly + \beta_2 * Shortleaf + \beta_3 * YellowPolar + \gamma_1 * Proximal + $$
$$+ \gamma_2 * Central + \psi_1 * Transpiration_2 + \epsilon$$

$\beta_0\ = 1119.8,\ \beta_1\ = 140.2, \beta_2\ =- 153.3\ ,\ \beta_3\ = 74.5\ , \gamma_1 = 112.3, \gamma_2 = 65.0,$
$\psi_1 =- 114.6$

Still, we should notice that the line in the residual vs. fitted plot is not perfectly straight and close to 0 in both models, so our model is not perfectly accurate.