Fitting an ANOVA model to assess the influence of segment location on moisture in branches of trees

Seif Ben Mustapha, Mohamed Dhraief, Alice Patin, Julian Schnitzler

Introduction

This report takes over on the scientific question of the paper of J.J. McDermott published in 1941. The paper was studying the effect of different parameters on tree moisture content. The authors looked at how tree species as well as the type of the transpiration rate, the branch segment chosen and its cutting method could predict moisture. Because of the tension of water inside the branch the method of cutting of the branch segments is expected to influence the measured moisture content. In the below work, we optimized the model of prediction of moisture based on ANOVA and on the Akaike Information Criterion.

The dataset consists of the following columns:

- *species*: The species of the tree, with
  - 1: Loblolly Pine
  - 2: Shortleaf Pine
  - 3: Yellow Poplar
  - 4: Red Gum
- *branch*: Branch of the tree
- *location*: Location of cutting on the branch, with
  - 1: Central
  - 2: Distal
  - 3: Proximal
- *transpiration*: Transpiration Type, with
  - 1: Rapid, i.e. cutted on a hot, dry, sunny day
  - 2: Slow, i.e. cutted on a cool, moist, cloudy day
- *moisture*: measured in % of dry weight

Exploratory Data Analysis

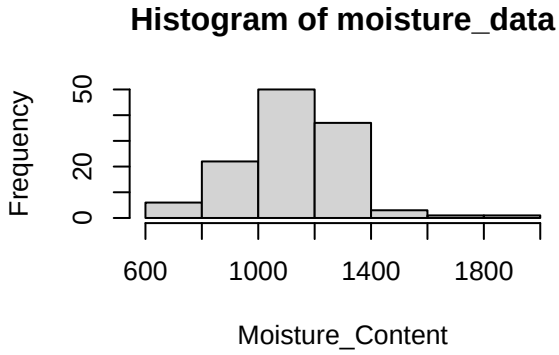
Univariate Analysis

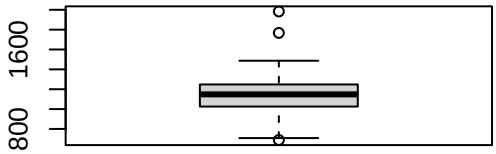
We can see that in the dataset Species, Branch, Location and Transportation types are categorical values, while Moisture Content is a numerical value. We will discuss first the categorical attributes and then numerical attribute (Moisture Content).

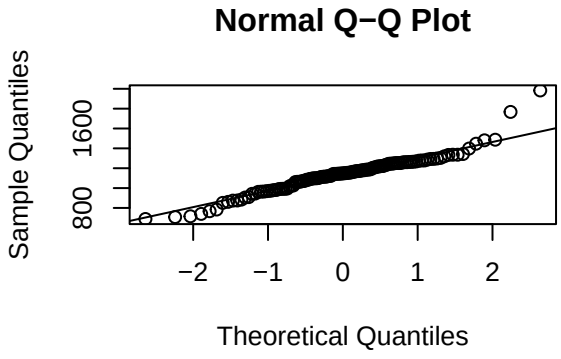
- *Categorical values*:

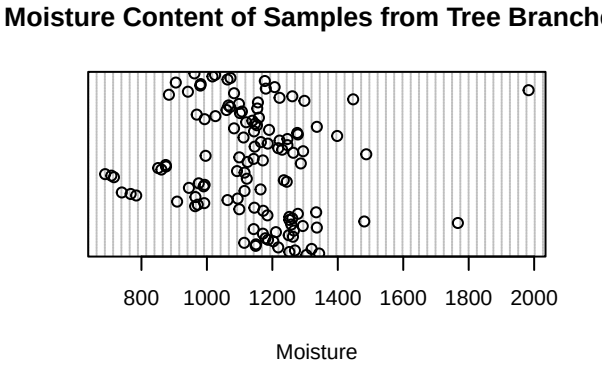
All of the variables Species, Branch, Location and Transportation are uniformly distributed. The difference between those attributes is how many values can each attribute can have.

- *Numerical values*:









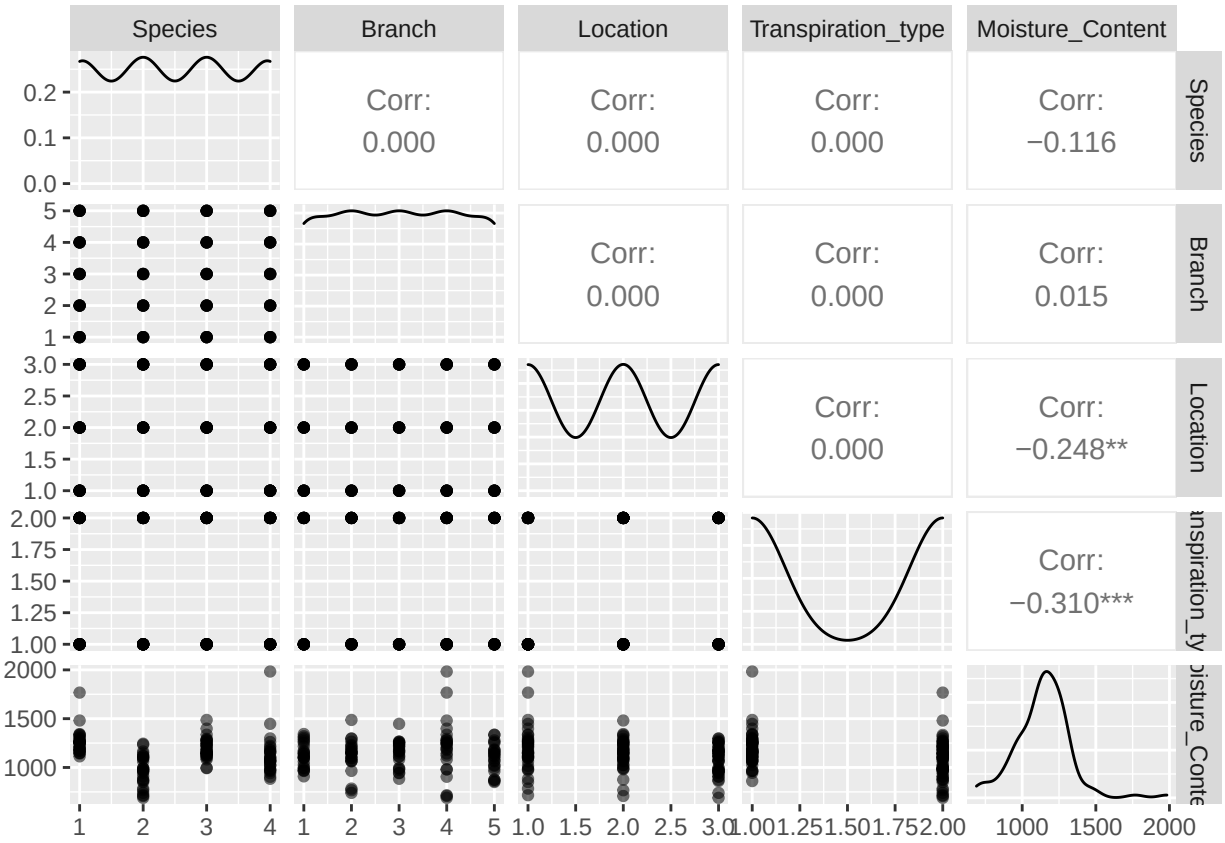
Moisture Content has a more complex distribution. Here are the main analytical metrics:

- *Min*: 689
- *1st Quartile*: 1026
- *Median*: 1148
- *Mean*: 1137

- *3rd Quartile*: 1247
- *Max*: 1983

We can see from the normal q-q plot and the histogram that this distribution is right-skewed. If we look deeper into the spread of the Moisture Content, the feature has a Standard deviation of 185, an Interquartile range of 221 and a Median Absolute Deviation of 152. This also conveys how dense is this attribute close to both its median and mean.

Bivariate Analysis



The pair-wise plot presents both a numerical and graphical analysis. It contains in the upper-right triangle the pairwise correlation coefficients and in the lower-left triangle the pairwise distributions. We notice a significant correlation of -0.248 between the location of cutting on the branch and the moisture content and a significant correlation of -0.310 between the transpiration type and the moisture content and no significant correlation between other variables.

Model Fitting

In order to find the model that fits the most accurately with the moisture data, we first created a model considering each parameter separately: species, branch (sample number),

location (location of the segment on the branch which corresponds to the cutting method) and transpiration (transpiration rate).

$$moisture \sim species + branch + location + transpiration$$

Then, we constructed another model taking into account the interaction of the parameters two-by-two and the interaction of all the parameters together:

$$\begin{aligned} moisture \sim\ & species + branch + location + transpiration + species : branch + species : location \\ & + species : transpiration + branch : location + branch : transpiration \\ & + location : transpiration + species : branch : location : transpiration \end{aligned}$$

Finally, we compare the two models.

In order to choose the "best-fit" model, meaning the model explaining the most the variance in the data, we used the "stepAIC" function. This function allows to simplify the model by decreasing the number of features without loosing too many pieces of information. For each version of the model, the Akaike Information Criterion (AIC) is computed. The AIC represents the amount of information loss by the model. It is the inverse of the information explained by the model. The aim is to choose the model with the lowest AIC value, meaning that the model explains more information. The absolute value of AIC has no significance but we compare its value between two models.

The models are fitted using the Least-Squares Approach, which minimizes the sum of the squares of the residuals between the predicted values and the observed values.

Looking at the different steps of the stepAIC function, we see that that the first model containing all the interactions has an AIC of 1247.37. At the last step, the model with the lowest AIC of 1236.98 was the one containing only an interaction term between species and transpiration. The "best-fit" model is

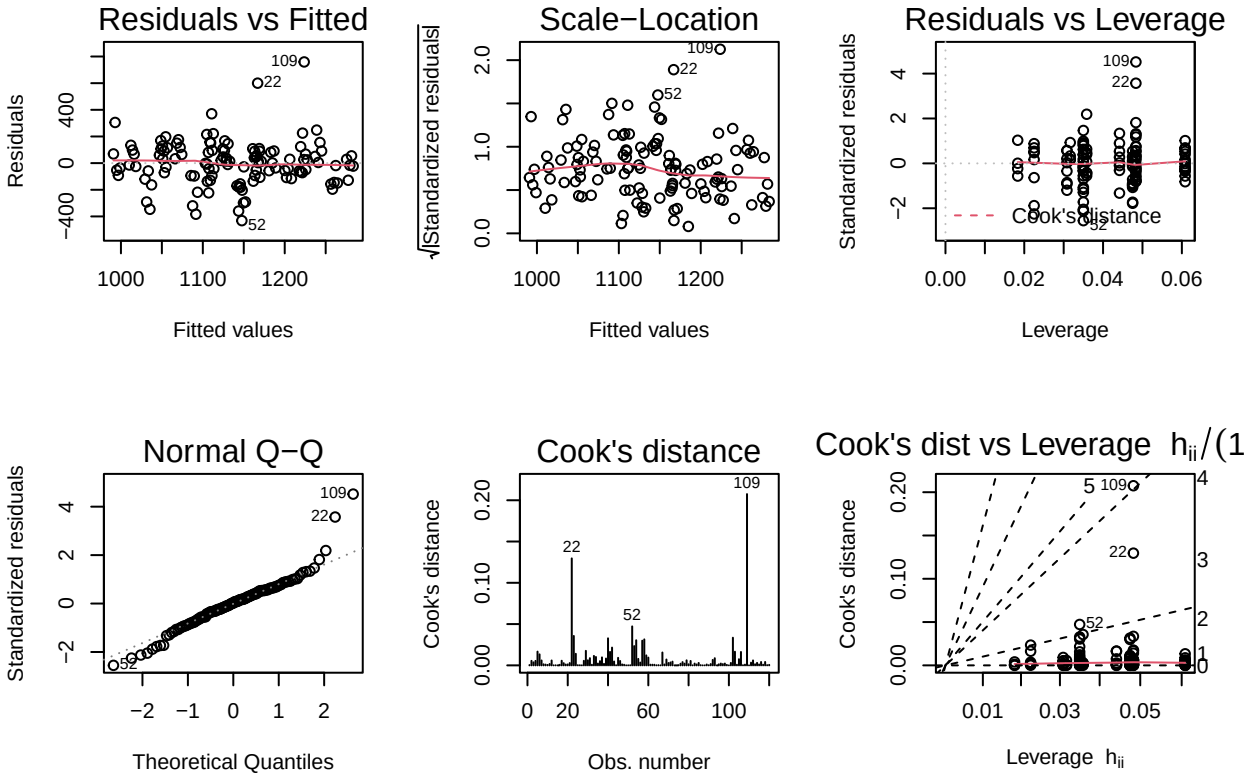
$$moisture \sim species + location + transpiration + species : transpiration$$

After removing outliers as described in the Model Assessment section, we receive a AIC of 1179.8 in the last step of stepAIC, giving us the best model as

$$moisture \sim species + location + transpiration$$

Model Assessment

In the following, we assess whether the assumptions underlying our model hold, i.e. that the errors have mean 0, the errors are homoscedastic, the errors are uncorrelated and are normally distributed.

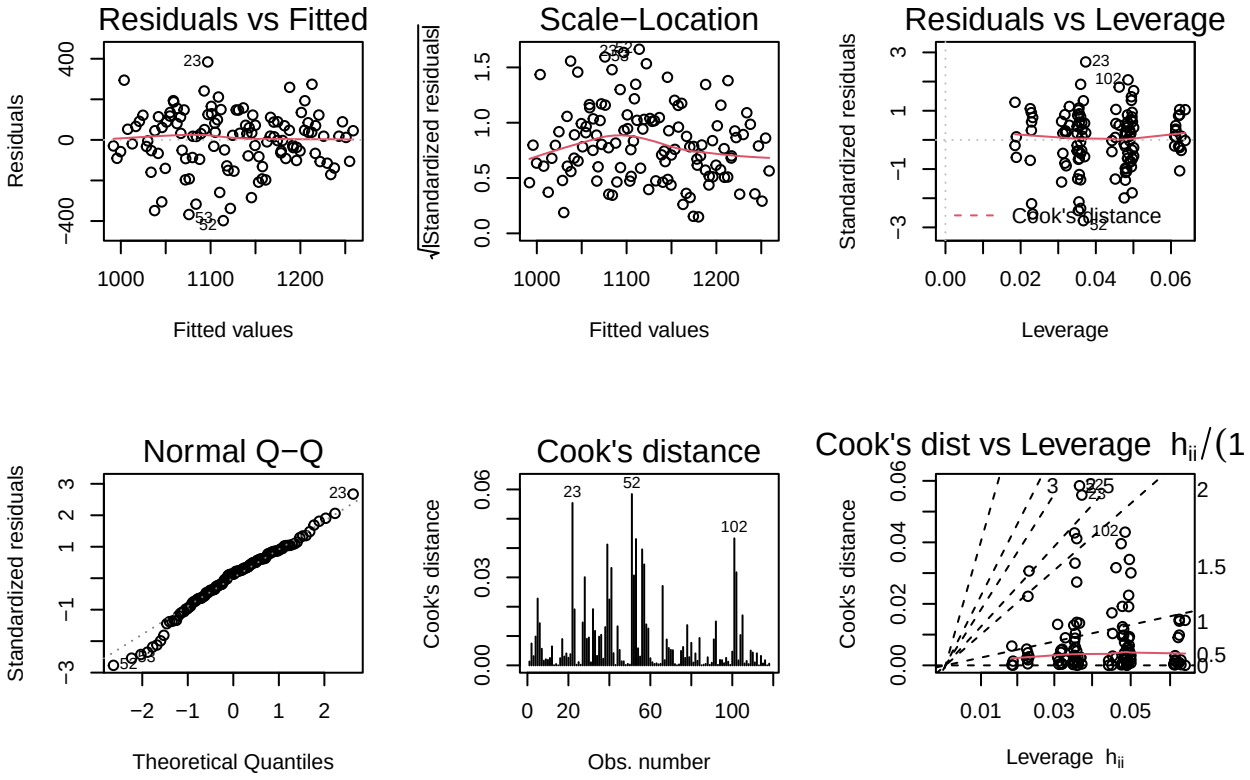


We observe that there are two outliers that might harm our model assumptions: Rows 22 and 109. Since these two points are the only two points that appear as outliers in all 4 graphs, we will remove those two from our dataset. For all other observations, it seems that they follow the model assumptions approximately.

In the Residuals vs Fitted graph, we can observe that the errors are homoscedastic, i.e. have the same variance, since the datapoints distribute in an similar manner above and underneath the line, independent of the fitted value. We can see an exception for 109 and 22.

The design of the experiment as written in the original paper indicates that the samples were chosen randomly, so we can consider all features to be uncorrelated.

From the Normal Q-Q plot, we can see that the residuals follow the line closely, i.e. the standard residuals do not exceed 3, except for rows 109 and 22. Hence, except for those two points, we consider the assumption of normally distributed errors to hold.



Indeed, removing those outliers seem to make the whole dataset support the model assumptions.

Our final, fitted model without removing outliers is

$$\hat{y} = 1274.25 + 58.87\ species - 56.13\ location + 15.60\ transpiration - 52.09\ species \cdot transpiration$$

When removing outliers, we get as our final model

$$\hat{y} = 1422.18 - 20.94\ species - 38.34\ location - 111.85\ transpiration$$

Conclusions

In comparison to the model of the paper, we see that our model fits better the moisture content of trees than the one from the authors, as the AIC of the model of the paper is 1583.49 and ours is 1236.98. Since two observations of the dataset seem to be outliers, when removing those we receive an AIC of 1179.8. If we take a step back, our model confirmed that as expected by the authors, the segment location on the branch is influencing the measured moisture content, probably due to the difference in water tension inside the segment. Our

model also showed that species and transpiration are not only influencing moisture individually but the interaction between them has to be considered in order to predict the most precisely possible the moisture content.