MATH-493 Applied Biostatistics - Project 1

Team 28: Loïc Busson, Riccardo Cadei, Khalil Merzouk
Model: ANOVA
Dataset : `treemoist.dat`
Term: Spring 2021

1 Introduction

In 1941, J.Joseph McDermott pursued a scientific study in order to find the effect of the method of cutting on the moisture content of samples from tree branches. Before designing his experiments, he had to find a way to measure this moisture content. To do this, he used the following process:

- measure w_w which is the weight of the tree branch sample directly after being cut (wet);
- measure w_d which is the weight of the tree branch sample after it has dried;
- define the moisture content M as proportional to the ratio $\frac{w_w - w_d}{w_d}$.

This process makes sense as only the moisture is responsible for the change of weight of the tree branches. For more readable results, McDermott decided to define the moisture M as 10 times the ratio of weights ($\frac{w_w - w_d}{w_d}$), and convert this value into a percentage. This way, a moisture $M = 1200\%$ means that $w_w = 2.2 \cdot w_d$.

Now that he had a way to measure the moisture of a sample, he designed an experiment to gather data and answer the following question: what are the effect of the location of the cut and the transpiration conditions on the moisture content of tree branches of different species? To answer this question, he gathered 5 branches of 4 different tree species (Loblolly Pine (LP), Shortleaf Pine (SP), Yellow Poplar (YP) and Red Gum (RG)) and designed an experiment to study the influence of the following variables on the moisture content:

- species of tree: LP, SP, YP, or RG;
- location of the cut: Central, Distal or Proximal,
- transpiration conditions: Rapid (Hot, dry, sunny day) or Slow (Cool, moist, cloudy day),

Experiment: For all species of trees (4), he cut each branch (so 5 per species) in three different locations (central, distal and proximal) within 30 seconds. This way, he had $4 \cdot 5 \cdot 3 = 60$ different samples to work with (15 samples of each species, of which 5 are central, 5 are distal and 5 are proximal). Then, for each of these 60 branches, he made them undergo two types of transpiration: the first one being in a dry and hot sunny day where the transpiration was rapid, and the second one was in a wet and cold cloudy day where it was slow. For each transpiration conditions, the moisture value is computed thanks to the process described before. This experimental process gave the $60 \cdot 2 = 120$ data points (lines) in the `treemoist.dat` dataset.

In this paper, we will use the dataset gathered by McDermott to shed light on statistically significant factors (among location of the cut, transpiration conditions and species of tree) regarding the Moisture level using the ANalysis Of VAriance (ANOVA).

2 Exploratory Data Analysis

2.1 Individual effects of each factor on moisture content

We first start by having a look at the individual effects of each factor. In particular, for each feature we analyze the boxplots of the Moisture level, one for each value of the feature. Each boxplot briefly summarizes the distribution of the Moisture level, given a certain value of a factor. These plots are reported in Figure 1.

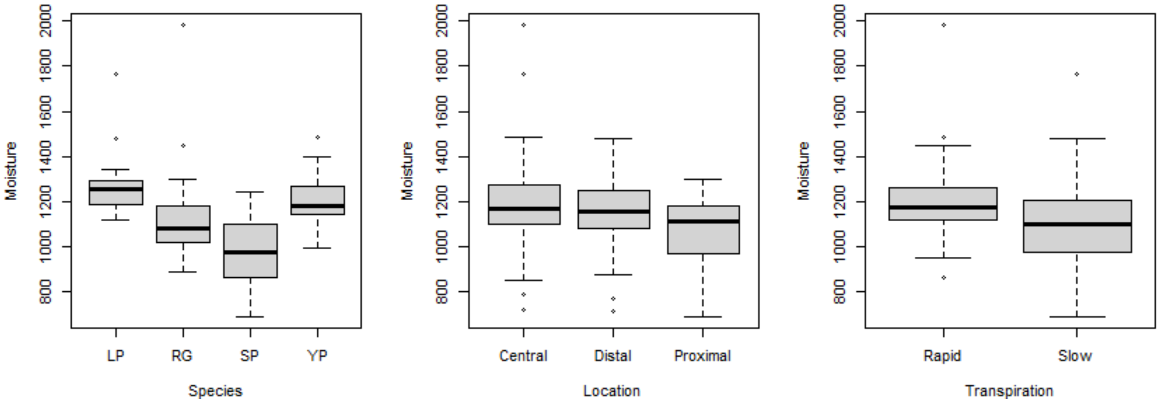

Figure 1: Individual Boxplots

In the *Species* boxplots we can see that the values of all the quartiles are quite different among the 4 types. We also know that the means are quite different and so we suspect that the *Specie* type has an significant influence on the *Moisture*. Similar differences are shown in the *Transpirations* boxplots. In particular we can see that the median of the *Slow* class is smaller than the first quartile of the *Rapid* class. Finally, in the *Location* boxplots, the differences are less evident: the *Central* and *Distal* boxplots are very similar to each other and just the *Proximal* one is slightly shifted downwards. However further analysis should be conducted to draw conclusion and generally, just the different behavior of a class (*Proximal*) is enough to suspect that also the *Location* class influence the *Moisture* level. The boxplots also show the variance and the outliers for each type of each factor.

2.2 Interaction effects between factors on moisture content

Now that we have analyzed the individual effect of each factor, we will now look at how the interactions between each couple of factors affects the moisture content. As we have 3 individual factors, we will have $\binom{3}{2} = 3$ interactions plots.

The plot in the middle of Figure 2 (involving the *Species* and the *Transpiration* variables) shows an interaction between these two factors that cannot be ignored in our model. Indeed, for most species (RG, SP, YP), rapid transpiration is equivalent to having around 200 *more* moisture than slow transpiration, whereas for the LP species, the two lines crossed, hence making rapid transpiration having 50 *less* moisture than slow transpiration! This phenomenon of *line crossing* means that the effect on moisture content not only changes quite a lot (from +200 to -50), but also is reversed depending on species! Hence, this plot indicates really strong interaction between the two factors *Transpiration* and *Species*. For the rest of the plots, the curves are mostly similar. The gap between them stays almost constant, indicating that the two factors do not really influence one another.

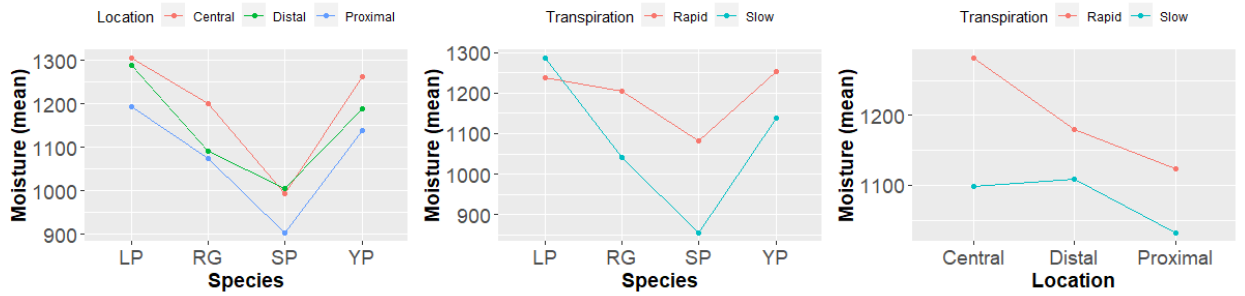

Figure 2: Interaction plots

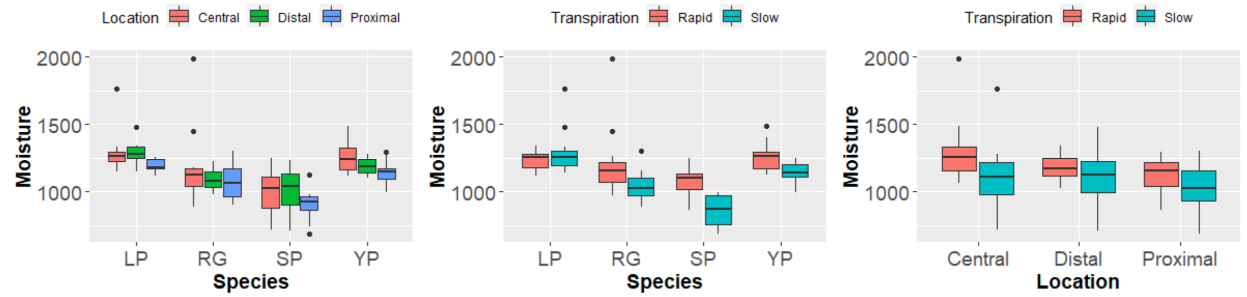

Figure 3: Boxplots per couples

If we now look at Figure 3 where are represented the interactions as boxplots, we see that the little changes causing lines to not be perfectly parallel might be caused by outliers, which are present in almost all the features and have a strong influence in the interaction plots as the dataset is relatively small (and as the "mean" measure is sensitive to outliers). Such ouliers can change the mean (and thus the interaction curve) quite a bit and make it seem like there is a interaction (when in reality there is none).

Therefore, from the above exploratory data analysis, we predict the following variables to be significant in our model: *Species* + *Location* + *Transpiration* + *Species:Transpiration*. If our graphical reasoning were correct, the best ANOVA model should be the one using only the variables stated above. In the following, we will consider that a feature is *statistically significant* if and only if its p-value for the F-test is lower than 0.05.

3 ANOVA

3.1 Models

We first performed the simplest ANOVA model considering only the individual effect of each factor. We remember that the *Branches* is not a factor but only the number of the data point collected in a certain *Specie*, in a specific *Location*, with a certain *Transpiration* condition. In formula:

$$Moisture \sim Species + Location + Transpiration \tag{1}$$

This model can see both as a baseline for future more complex models, but also to check the relevance of each factor. The result is shown in the following table:

	Df	Sum Sq.	Mean Sq.	*F* value	*p*-value
Species	3	1432590	477530	26.557	$4.48e-13$
Location	2	254589	127294	7.079	0.00127
Transpiration	1	394224	394224	21.924	$7.95e-06$
Residuals	113	2031884	17981		

As expected from the Section 2.1, the F-test on the 3 factors confirms that all of them are statically significant factors to model the *Moisture* level.

If we also considered the *Branches* as a factor in this model, it would get a p-value on the F-test greater than 0.8: this confirm that the *Branches* shouldn't be considered as a factor to model the *Moisture* level.

We decided then to improve this baseline proposing a second model where we considered also all the interactions (in pairs). In formula:

$$\begin{aligned} Moisture \sim & Species + Location + Transpiration + \\ & Species : Location + Species : Transpiration + Location : Transpiration \end{aligned} \tag{2}$$

The result is shown in the following table:

	Df	Sum Sq.	Mean Sq.	*F* value	*p*-value
Species	3	1432590	477530	30.285	$4.39e-14$
Location	2	254589	127294	8.073	0.000556
Transpiration	1	394224	394224	25.002	$2.39e-06$
Species:Location	6	47261	7877	0.500	0.807398
Species:Transpiration	3	305952	101984	6.468	0.000473
Location:Transpiration	2	70368	35184	2.231	0.112596
Residuals	102	1608305	15768		

Similarly we used the F-test on each factor to filter only the statistical significant ones. All the individual factors remain significant, and as expected from Section 2.2, only the interactions between the *Species* and *Transpiration* seems to bring an added value to the model. We can also observe that considering also the interactions in pairs, the sum of squared errors decreased noticeably.

Finally, we trained the ANOVA model considering only the relevant features just filtered from the model with all the interactions (and already selected in Section 2.2). In formula:

$$Moisture \sim Species + Location + Transpiration + Species : Transpiration \tag{3}$$

The result is shown in the following table:

	Df	Sum Sq.	Mean Sq.	*F* value	*p*-value
Species	3	1432590	477530	30.435	$2.10e-14$
Location	2	254589	127294	8.113	0.000517
Transpiration	1	394224	394224	25.125	$2.07e-06$
Species:Transpiration	3	305952	101984	6.500	0.000433
Residuals	110	1725933	15690		

All the factors are statistically significant (all the p-values of the F-tests are even smaller than 0.001), the model is simpler than the previous one and the sum of squared errors doesn't increase so much. We therefore decided to consider this model as the definitive one.

3.2 Hypothesis testing

The ANOVA tests assume that the groups are independent, normally distributed and that the variance among them should be approximately equal. In order to accept the results from the previous Section, we have first to check if these hypotheses hold.

Independence: Given by the experiment settings.

Normality: We checked the normality assumption of the dependent variable (*Moisture*) through the QQ Plot reported in Figure 4. It shows that the normal assumption mainly holds only for the central values, but globally the distribution is a bit right skewed. The Shapiro-Wilk test also confirms the normal assumption doesn't perfectly hold and a logarithmic transformation of the output doesn't seem very helpful.

Homoscedasticity: A comparison of the variabilities among the different groups was already showed in the box plots in Figure 1 and Figure 3. Even if the size of each box plot are slightly different we can generally say that homoscedasticity among group could hold. This is also supported by the plots *Residual vs Fitted* and *Scale-Location* in Figure 4, which don't present any clear pattern among the residuals, their spread is roughly equal at all fitted values and centered to zero.

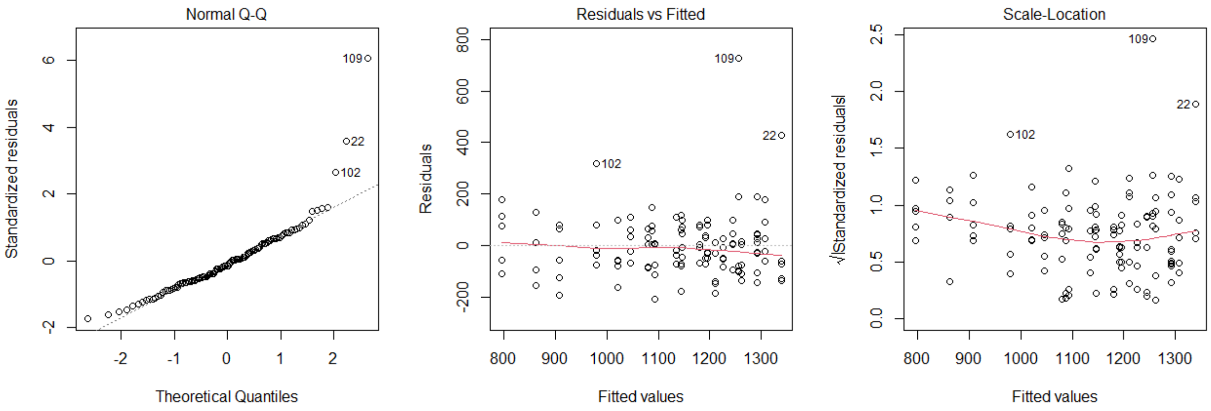

Figure 4: From the left: Normal QQ plot, Residual vs Fitted values plot and Scale-Location Plot

4 Conclusion

Our ANalysis Of VAriance concludes that there is a statistical significance to say that *Species*, *Location* and *Transpiration*, and also the interaction between the *Species* and *Transpiration*, influence the *Moisture* level, and more specifically that, for each factor, there is at least a class with a mean value different from the others. However, the ANOVA hypothesis of normality doesn't hold perfectly and it is also the case for the homoscedasticity hypothesis. A bigger dataset of independent and balanced data points should be considered for further investigation of these hypotheses and preventing the undesired effect of few outliers.

References

- J.J. McDermott. 1941. *The Effect of the Method of Cutting on the Moisture Content of Samples from Tree Branches*, American Journal of Botany, Vol.28, 6, pp

- Brian S. Everitt and Torsten Hothorn, *A Handbook of Statistical Analyses using R*, 2005.