Project A1 : Fungal invasion of the Apple fruit

Auriane Cozic, Thibault de La Taille, Eloi Littner, Audrey Menaesse

May 1st, 2020

Introduction

Apples are among the best-preserved fruits. However, after a while fungi strains are inevitably starting to grow. In F.G. Gregory and A.S.Horne study, the conditions of infection, the progression of the invasion and the final stage of invasion were studied for two varieties of apple : the Cox's Orange Pippin from Burwell, Cambridgeshire and Bramley's seeding from 6 different localities. The data set under study describes the fungal invasion of 35 apples. It contains 7 different variables, regarding:

- the characteristics of the apple (variety, weight, radius)
- the strain of the fungi, separated in 7 types (A, B11, B111, C1, C21, C3 and D)
- the measurements of the infection (days after infection, fungal radial advance, rate of fungal advance).

The aim of our study is to compare the infecting power depending on the conditions of infection and the resistance to invasion of different varieties of apples and their characteristics.

Exploratory data analysis

Data description

To begin with, we want to get familiar with the data. The variable *variety* contains several metadata information. Therefore, in order to have explicit variables in our model, we decided to split it into two new variables : the year when the experiment was conducted and the storage temperature of the apples.

The study is looking at 2 different varieties of apples and 7 fungi strains. It comes from 5 different experiments started in 1924 or 1925. 4 of them were performed at the same temperature (12°), but one was at (3°). The duration between the infection and the measurements is different for each experiment.

Outliers

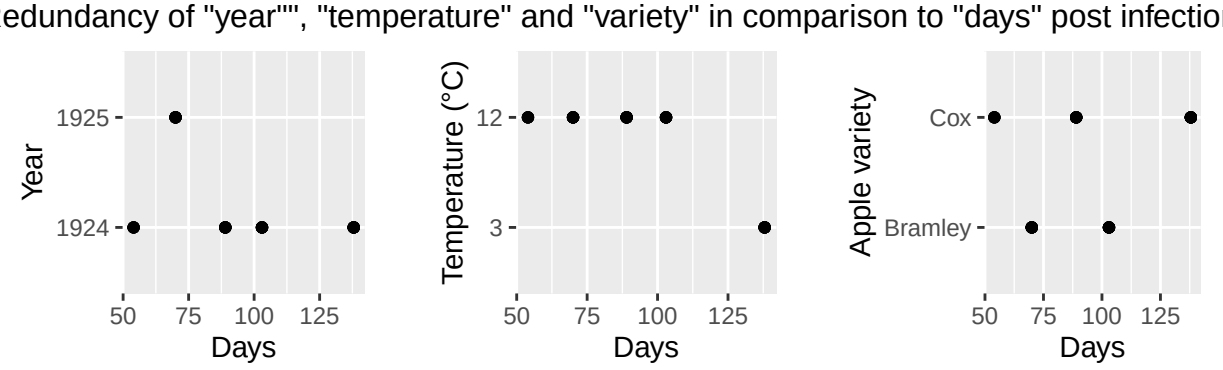
Exploring the data, 3 outliers can be detected. The Cox apples *18* and *21* have a fungal radial advance superior to the apple radius. As it does not have any physical sense or meaningful explanation, they were removed. The infection did not develop in one apple *17*. The fungal radial advance is very lower to any other apple. As it the only case, it reveals an experimental error and was also removed.

Correlations and redundancy

Then, we tried to identify correlations between variables prior to trying to model the infection power, as the presence of highly correlated explanatory variables would complexify the model without bringing any valuable information. After some exploration, we noticed that the apples' "weight" and their "radius" were highly correlated (based on Pearson's correlation coefficients), as shown below. This makes sense based on the physical relationship between these two, and we will hereafter ignore the "radius", as the apple "weight" would in our opinion be more reliable in an infection model. What's more, we can see a good correlation between the "fungal radial advance" and the "rate of advance", which is logical since the latter is computed by dividing the former by the number of "days" since infection; and since we are more interested in an infection power, independently of the time elapsed, we chose the "rate of advance" as our variable of interest and ignored the "fungal radial advance".

Table 1: Pearson's correlation coefficients

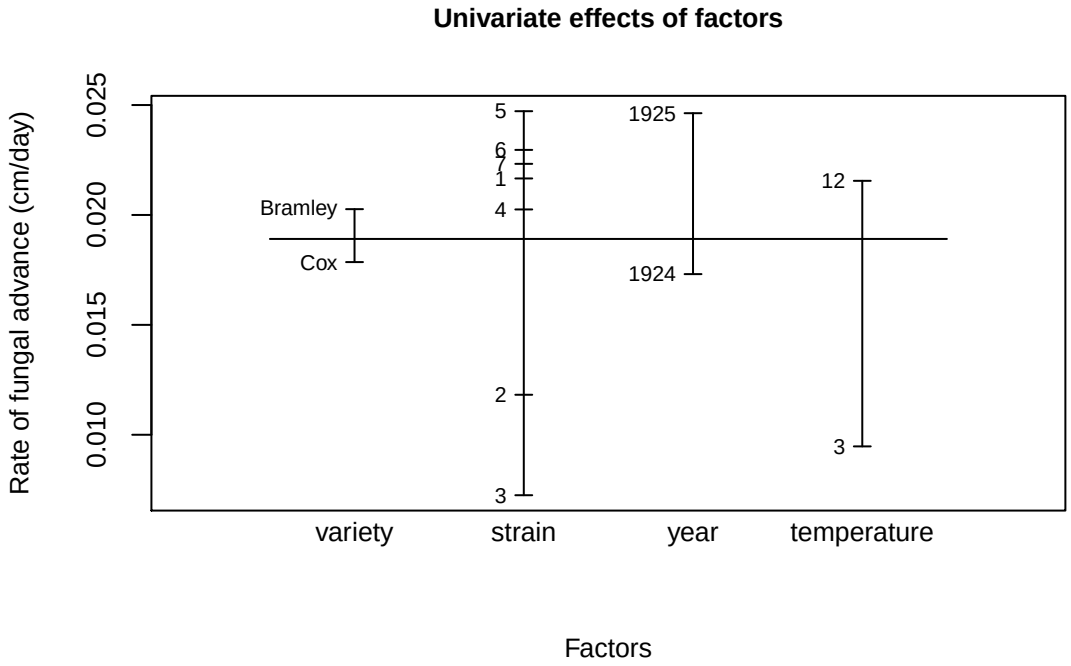
	weight	radius	fungal radial advance	rate of advance
weight	1.00	0.97	0.17	0.24
radius	0.97	1.00	0.18	0.22
fungal radial advance	0.17	0.18	1.00	0.81
rate of advance	0.24	0.22	0.81	1.00



Furthermore, we can see that the factor "days" actually contains information on the "year", "temperature" and "apple variety" variables, which is coherent with how the different batches of apples were monitored. Because of that, we removed the "days" variable from our study as

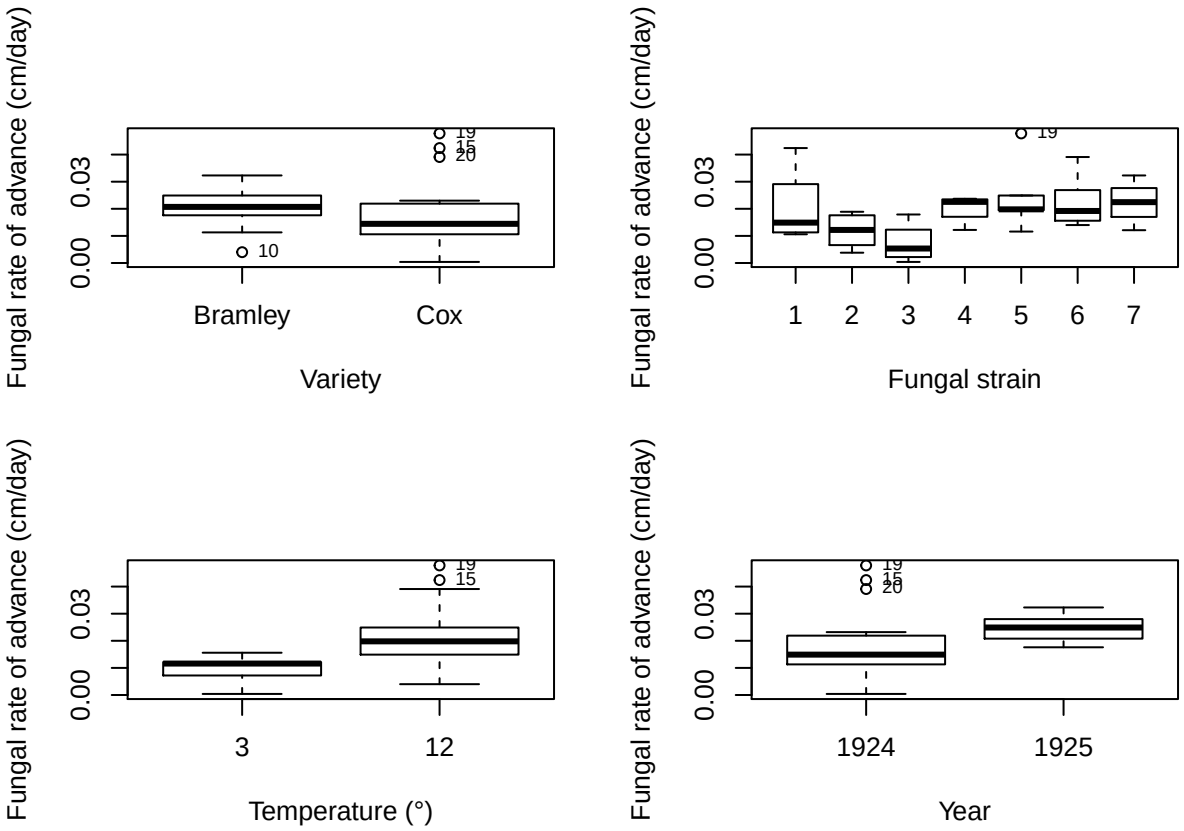
these correlated factors would affect our model.

#Exploratory data analysis



The strain, temperature and year seem to have a strong influcence on the rate of fungal advance, while the variety has low impact.

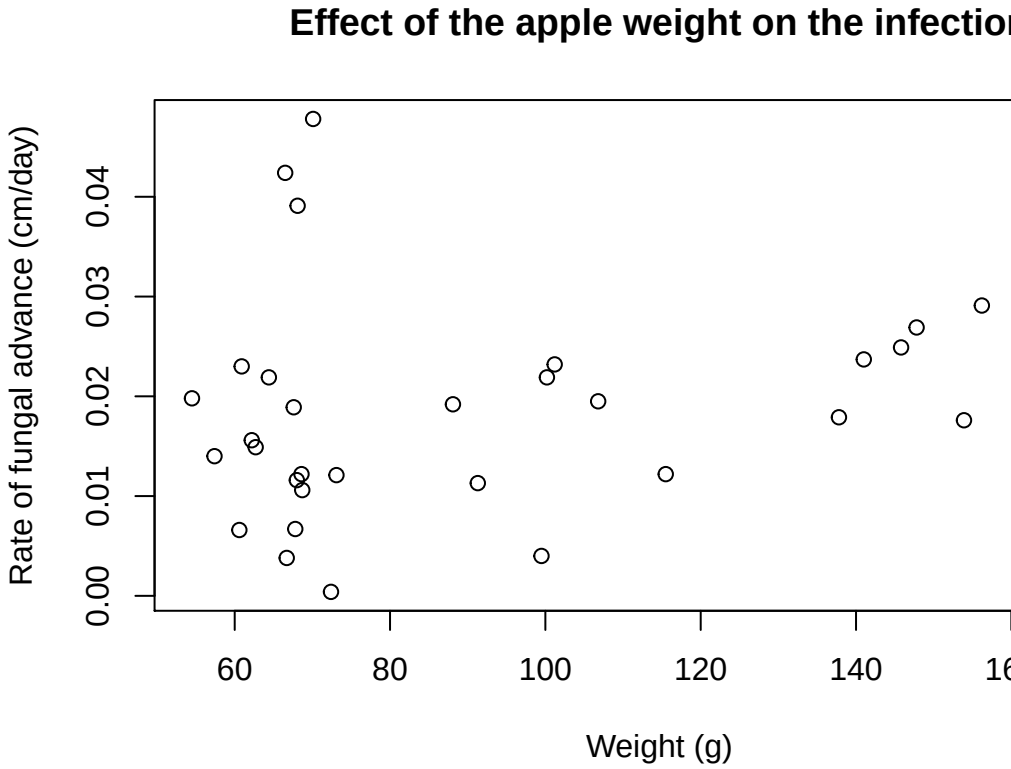
###if enough space, graphs This can be confirmed by looking at the data points



The fungi strains of type V (2 and 3), and a cold temperature, seem to have a reduced infecting power. It can also be noticed that, fungi strain 1, 12° temperature and year 1924 experiments display higher variability, which is confirmed by their standard deviations. This is due to the fact that each of these categories gather only few data (less than 5). Any generalisation of the results is thus complicated.

##Do not show

3 outliers apples stand out having particularly high Fungal rate of advance (15, 19 and 20), and one has a slightly lower one (10).



###graph only if enough space
Finally, we can not see any abnormal variability regarding the rate of fungal advance depending on the weight.

#Infection modeling ##Full data The exploratory data analysis enabled us to disentangle the non-redundant, independent, parameters for modeling the progression of fungal infection in the apple. More precisely, we identified the infection rate of advance as the most reliable reporter of the infection strength and the variety, *Fusarium* strain, apple weight, storage temperature and year of experiment as potential explanatory variables.

To select the most meaningful variables, we first fitted a full linear model, using all the independent parameters cited above. Then, we used the stepwise AIC (Akaike information criterion) method for model selection. This method serially checks if deleting or adding a variable to the model improves its AIC score until a minimum is reached : the lower the AIC score, the better the model.

This method, when applied on the full data (with the outliers) yielded the following model with AIC score -317.68 :

$$RateOfAdvance = \beta_0 + \beta_1.weight + \beta_2.\mathbb{1}_{variety=Cox} + \beta_3.\mathbb{1}_{temperature=12} + \sum_{i=2}^{7} \beta_{4,i}.\mathbb{1}_{strain=i} \quad (*)$$

Only "year" was removed from the explanatory variables, showing that no seasonality effect can be inferred from the data. Looking at the Cook's distance diagnostic plot confirmed the fact that samples 18 and 21, previously identified as non-sense data, have a strong influence

on the model. Hence, we decided to do a further stepwise model selection using filtered data as described in the explanatory data analysis.

##Data without outliers

This time, the final model also included the variable "year", but with a tremendously high p-value ($Pr(> |t|) = 0.21363$). To remain consistent with the model obtained when the whole dataset was used, we chose to run a final stepwise model selection without the "year" variable. Removing "year" did not change fundamentally the obtained AIC (-303.1 instead of -303.52 when "year" was included) and even slightly reduced the model's global p-value (0.002044 instead of 0.002403). Furthermore, an ANOVA of the two models yielded no significancy for the addition of year. Our final model is then also described by $(*)$, with the following parameter values. To gauge the consistence of our model and the influence of outliers, we added the parameter estimate when fitted on the full data (last column, "FullData"). No important change could be observed.

	Estimate	Std. Error	t value	Pr(>|t|)	FullData
(Intercept)	-0.0177	0.0113	-1.5698	0.1307	-0.0197
varietyCox	0.0155	0.0057	2.7347	0.0121	0.0174
strain2	-0.0106	0.0049	-2.1680	0.0412	-0.0106
strain3	-0.0133	0.0052	-2.5443	0.0185	-0.0147
strain4	-0.0001	0.0052	-0.0271	0.9786	0.0047
strain5	0.0031	0.0049	0.6267	0.5373	0.0031
strain6	0.0022	0.0049	0.4514	0.6561	0.0021
strain7	0.0002	0.0052	0.0295	0.9767	0.0053
weight	0.0002	0.0001	2.7618	0.0114	0.0002
temperature12	0.0143	0.0038	3.7655	0.0011	0.0172

Assessment of the model

##Residuals

In order to assess the model found previously, we first check the QQ-plot of the fitted data.

The tail of residuals is composed of a very few data points which are significantly far from the model. These two outliers (15 and 19) were already noticed during the exploratory data analysis, as having particularly advanced infection stage.

The study of the residuals is also important to assess the model. Indeed, if the residuals are not carrying any information relevant to the model, their mean should be zero and no pattern should appear, which is the case in the following plot.

Based on these plots, we can confirm that the data set is then well represented by the linear model.

##Homoscedasticity

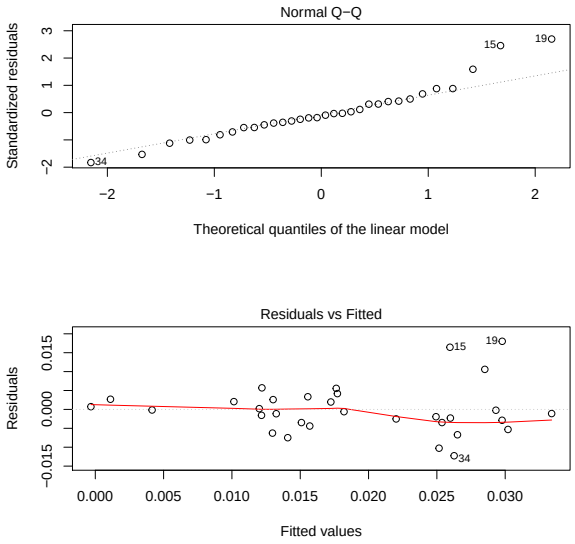
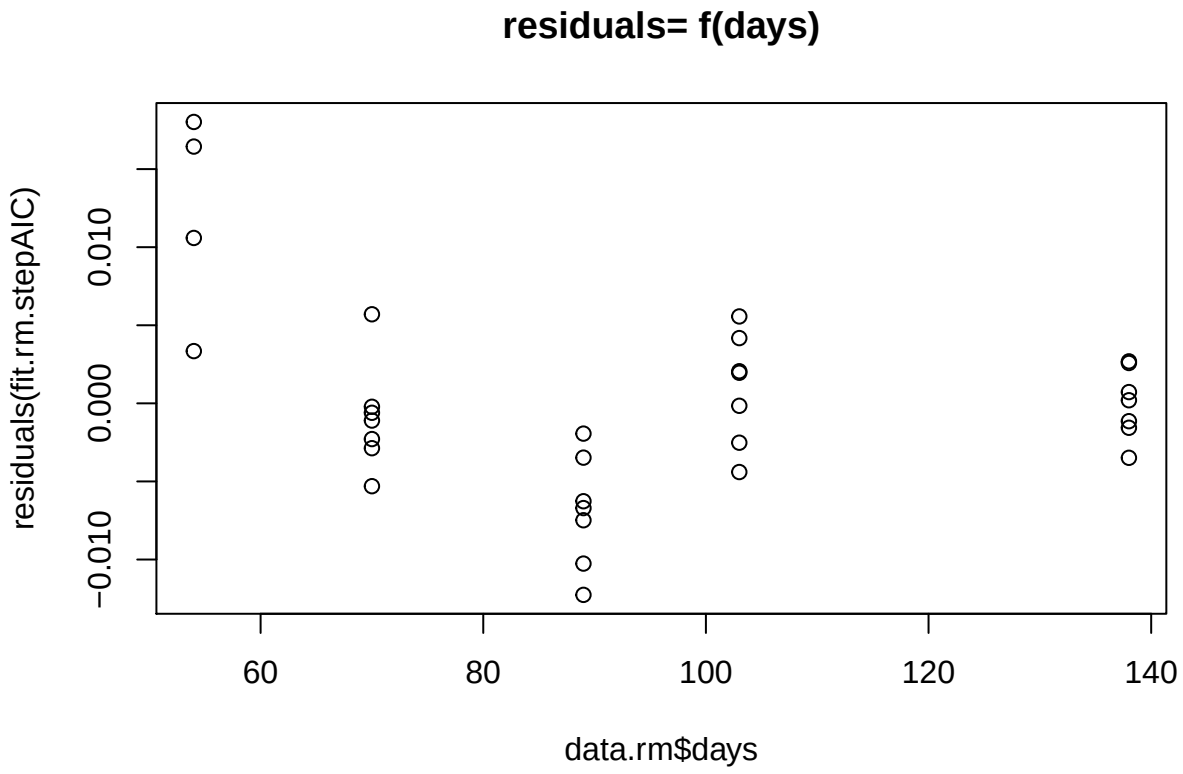


Figure 1: Diagnostic plots

To check that the dataset is not biased, we verified that the variance is not dependent on the conditions of the experiment. Dependencies of the residuals on days were plotted, as *days* variable differentiate the apples based on the experiment it was part of.



No pattern can be observed on the graph, which confirms the homoscedasticity of the dataset.

#Discussion / Results (short)

We have identified 4 characteristics influencing the course of infection.

Variety of the apple. Infection progressed more slowly in Cox apples than in Bramley apples. This reveals that they have different resistance to the fungal invasion and that one should prefer Cox's Orange Pippin than Bramley's seeding to avoid infections in its fruits.

Strain of the fungi The fungi strain influenced the rate of fungal advance. Especially, the two B strains (B11 and B111) are less invasive than all others.Different fungi strain can thus have different infecting power.

Temperature Some environmental conditions play a role in infection, as the temperature. Very cold temperature (3°) delays the course of infection, compared to cold temperatures (12°). However, only one experiment was conducted at 3° with 7 apples. This represents very few amount of data, which explains the high variance associated. It is thus difficult to affirm any generalities about the ideal temperature, but we can assess that this variable has a significative impact.

Weight of the apple The weight of the apple, and by extension its size has some impact on the speed of the infection. The heavier, the more rapidly the infection progresses. IHowever, this influence is moderate. We suggest it is due to the fact that the infection is less limited by the boundary conditions, and that this variable is not related to the resistance of the apple itself.

The variable year is not relevant to explain the infection power. This means that the experiments were reproducible, and support the affirmation that the dataset is unbiased.

###Conclusion #TODO : finish it The most impactant variable is the day of the infection : the infection is strongest (progresses more rapidly) when the apple is infected early. (Pas tres sure de comment interpreter ça ?)

The strains B (strains 2 and 3) are very less virulent.

Finally, low temperature leads to more resistance of the apple.

#Question to the teacher Is our method ok or only ANOVA ? Do you validate the part when we remove year ?

#R?partition

Corr?lations + Conclusion : Thibault EDA sauf corr?lations : Audrey Fit du model : Eloi Residuals + discussion : Auriane

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.