# Forecasting Forest Fires

Pablo Ruiz Barnada

March 2023

# 1 Introduction

In the recent decades, the world has seen how climate conditions have been changing across every continent. These changes do not only affect temperature, but present a challenge to the biosphere that has adapted over centuries to a certain climate. In particular, forests, that are a key component in the development of other species and a valuable resource for humans [**1**], suffer the changing conditions the most. Alterations in climate produce an increasing risk of forest fires, endangering wildlife and negatively affecting the ecology and the economy. Throughout this report, an analysis will be carried out on the main factors that affect the severity of overall damage caused by forest fires.

# 2 Data exploration

The data used for the analysis corresponds to the Montesinho Natural Park, in Portugal. The dataset contains information from 517 forest fires that happened between 2000 and 2003, and are classified into three levels according to the severity of the damages caused: "none", "moderate" and "severe". Additionally, four other measures of the weather at the time of the fires are recorded: temperature ($^{o}$C), relative humidity (%), wind speed (km/h) and rainfall (mm/m$^2$).

An initial check on the dataset is carried out to find outliers in the measurements of each variable (Figure 1). Even though there are a few potential observations that can be classified as outliers, they likely belong to forest fires that happened under extreme weather conditions, as the severity of the damages correspond to what could be expected given the measurements, and not misrecorded data. For instance, the fires under abnormally high relative humidity lead to no damages, while stronger gusts of wind are associated to moderate and severe damages to the forests. Hence, these observations will be kept and used for further investigation in the coming sections.

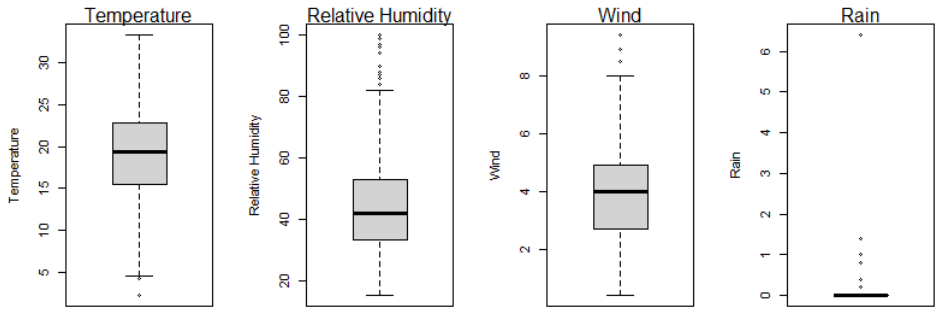


Figure 1: Boxplots for the distribution of each variable.

The analysis of the dataset reveals that most fires recorded carried no damage (approximately 48%), followed by a moderate damage (34%), and severe fires (18%). By examining the weather conditions in which these fires occur, no clear differences are observed (Figure 2). In fact, the distributions for temperature, humidity and wind speed are similar across the three levels of severity. It can be argued that temperatures follow a distribution similar to normal but with heavier tails, with the highest median found in moderate fires, and the highest mean in severe fires. In regards to the humidity, the distributions are (positively) skewed, with the lowest mean and median observed in severe fires, and then followed by moderate and no-damage fires. The distributions for wind speed appear to be bimodal in the three cases, with the first hump being larger for no-damage fires, and the second one more relevant to moderate and severe fires, which have the highest median and mean. These results are indeed what were expected: lower relative humidity can be associated to dry environments, while wind is associated to the spread of the flames. Temperature is relevant to the creation of a fire, but is not as directly related to its severity as the other weather features.

The examination of the rainfall observations in the dataset reveal the expected yet extreme result, as there is rainfall present in only 1.5% of the fires. Breaking this down, 6 times the rainfall belonged to fires that lead to no damage, 1 time to moderate fires and 1 time to severe fires. How to use this variable will be discussed in the coming sections.

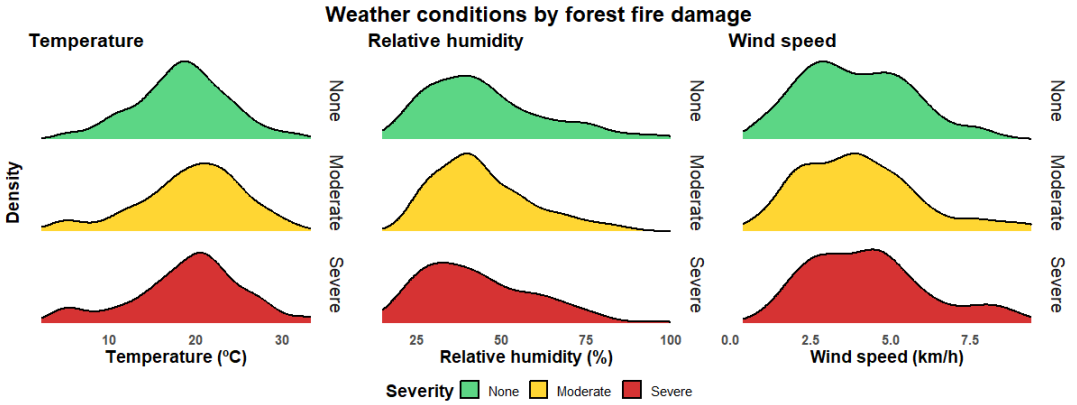


Figure 2: Densities of each variable.

Next, the relationships between the variables are studied (Figure 3). When taking the full dataset, the only remarkable relationship is found between temperature and relative humidity, which are negatively correlated (-0.527). This phenomenon is also observed when splitting the dataset by the severity of fires. Particularly, the correlation between temperature and humidity is higher with no-damage (-0.57) and moderate fires (-0.581), and slightly lower for severe fires (-0.38). This phenomenon is logical, as the relative humidity naturally decreases with higher temperatures, given the specific humidity [**2**]. Aside from this, some negative correlation is also present in moderate and severe fires between temperature and wind (-0.251 and -0.435, respectively). Again, this is not unexpected, as it is logical that a damaging fire does not need extremely high temperatures, provided there are guts of wind that can spread the flames.

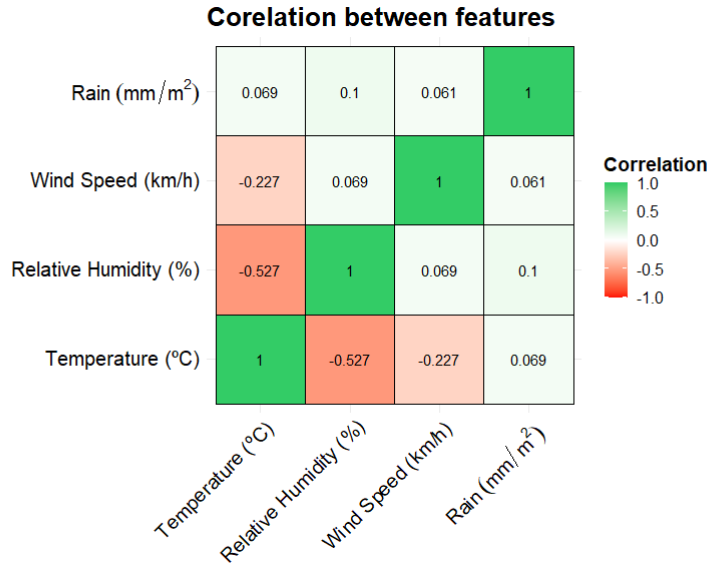


Figure 3: Correlation between variables, taking the complete dataset.

# 3 Methodology

In this section, the methods used for the analytical part of the report are presented. Firstly, a Multivariate Analysis of Variance (MANOVA) test is carried out, in order to check whether the measurements for each severity level are statistically different between one another. The null hypothesis for the MANOVA test is that the group means are the same, while the alternative hypothesis is that they are different. The MANOVA test yields a p-value of 0.1664, implying that group means are not statistically significantly different at any reasonable level, and the selected variables on their own might not be good to discriminate between groups.

Moreover, if the individual variables' ANOVA tests are performed, it is concluded that the mean of each variable is not statistically different between groups. Figure 4 shows the scatterplots of the main features in the dataset, distinguishing by severity of the fire: there is no clear distinction between groups, as anticipated by the variance tests above.

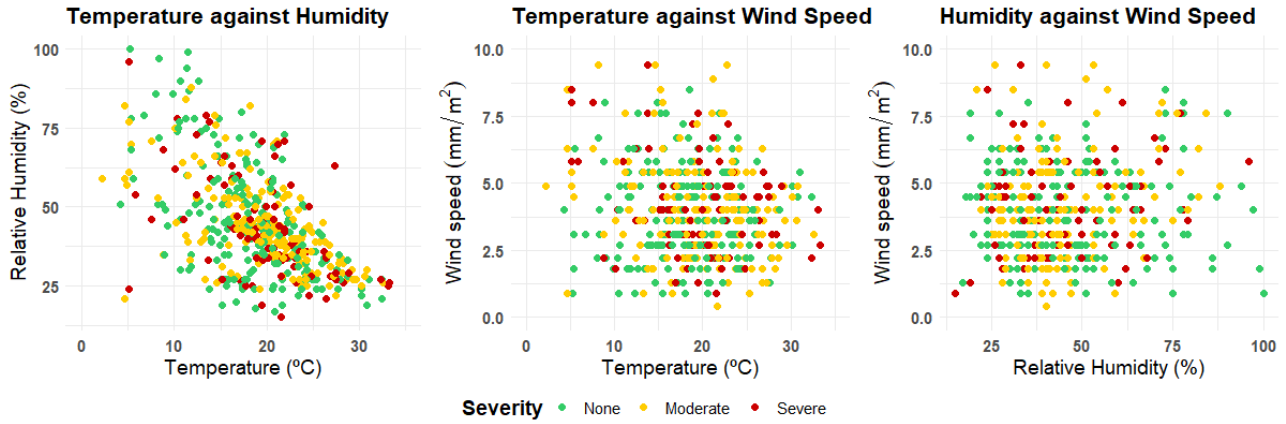


Figure 4: Scatterplots of the relationship between temperature, relative humidity and wind speed, per damage severity.

Next, the normality of the features recorded is studied. The multivariate normality is tested twice, once taking the four available variables, and a second time not taking rain, which may not be included in model creation. The null hypothesis for these two tests is normality. The p-values obtained are approximately 0, so in both cases we can reject the null hypothesis, hence rejecting normality (as expected from the distributions in Figure 2). Moreover, the Q-Q plots (Figure 5) exhibit a non-linear behaviour, with humidity and rain clearly deviating from the reference line, and temperature and wind doing it slightly in the tails.

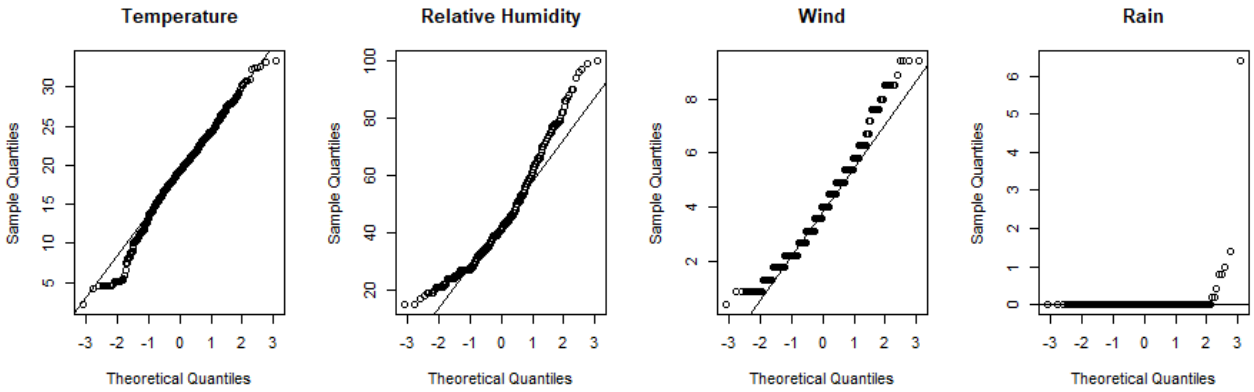


Figure 5: Q-Q plots of each feature in the dataset.

With all these information taken into account, the aim is to produce a model that can classify a set of measurements of weather conditions to their respective expected damage that a fire could cause.

# 4 Model Development and Selection

Two different models were developed with the objective of predicting the severity of a new fire, given certain weather conditions. The models created used Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Each model was fitted twice, once passing all the features recorded (complex model) in the dataset, and a second time dropping the variable "rain" (simpler model).

Firstly, the two versions (with and without rain) of LDA models are fitted. Each of them is fitted using the pre-defined training subset, obtaining similar results. Linear Discriminant 1 (LD1) explained about 60% - 62%

of the between group variance, being correlated mainly with wind, and with smaller effect of either rain (complex model) or temperature and humidity (simpler model). LD1 could be interpreted as a factor accounting for the risk of fire propagation, as wind guts typically do so. LD2 was (negatively) correlated with temperature, with some influence of wind in the simpler model, but the real-life explanation is not easily interpretable. The overall performance of these models was evaluated using Leave-One-Out-Cross-Validation (LOOCV). The forecasting accuracy was similar for the two models, achieving 47.8% of correct predictions each, and tending to forecast "no fire damage" more often than any other severity.

Lastly, the QDA models, which do not assume homogeneity of the covariance matrices, were developed. These models were able to balance their predictions, and forecasted moderate and severe forest fires more accurately than LDA. The overall performance exhibited a slight improvement, with each model being able to correctly predict about 48.5% of the fires after LOOCV. In particular, the biggest difference appeared from the improved predictability of moderate fires, approximately 15% better than in LDA models, and also an approximated 6% improvement for severe fires.

# 5 Results and Out-of-Sample Predictions

The previous section collects the results for each model developed. It is observed that the better performing models are based on QDA. These two models had a similar predicting power, about 48.5% each. However, the LDA models can predict about 47.8%, which means that, in practice, they have the same forecasting ability, and differences may just arise from variance in the data. Hence, the chosen model to carry out out-of-sample performance will be the LDA model that only takes three covariates (temperature, relative humidity and wind). This model is preferred for two reasons: firstly, eliminating a variable did not produce a loss in performance; secondly, for similar performances, the simpler model is favoured as it is computationally less demanding, and inference is faster. Not taking rain also eliminates the chance of obtaining wrong measurements and possibly outliers, as rain does not often happen while there is a fire, and if so, it is not present throughout all the time period that the fire is live. The LDA model with three variables is less computationally expensive than the other models because it has fewer variables and assumes that covariance matrices are equal, unlike QDA.

Hence, predictions can be done to estimate the risk of possible forest fires, given certain weather conditions. For instance, a hypothetical fire where the weather conditions are $27^{o}$C, 20% relative humidity and wind speed of 15 km/hour (with no rainfall) will be predicted to cause severe damages, with a probability of 41.47%, against 39.68% of being moderate. Analytically, this prediction makes sense, as it is a warm, dry and windy day, with the ideal conditions for causing and spreading fire. Moreover, the confidence in this classification could be of around 48%, accroding to the results of LOOCV. On the other hand, the model's predictions are of no damages most of the times, which is the main cause of misclassified fires. However, the predictions for moderate and severe damages tend to be accurate, so the chances of having forecasted correctly the damages of this hypothetical fire can be higher.

# 6 Discussion

This report has expounded different models to forecast the gravity of forest fires. After an initial exploration of the Montesinho Natural Park's fires dataset, it was concluded that the most typical weather measurements (temperature, relative humidity, wind and rain) are usually weakly correlated, except in the case of temperature and relative humidity, which are inversely correlated (correlation = -0.527). This is logical, as higher temperatures dissipate specific humidity, hence reducing relative humidity. Some negative correlation is also found between temperature and wind for moderate and severe fires. Further analysis showed that the weather conditions across the three classes of fire severity are not statistically different, so grouping fires in terms of their produced damages is potentially a difficult task, given the variables available.

Different predicting models (LDA, QDA, with and without rain) were developed, and all achieved a similar performance, obtaining about 47%-48% correct predictions of damage levels. Thus, the simpler, LDA model was chosen, as it is less demanding than QDA and suffers no forecasting losses. This model did not take rain into consideration for two main reasons: the predicting performance was the same after removing this variable, and the calculations are (slightly) less computationally demanding. The main difficulty of the LDA model was that it over-predicted fires with no burning damages. An example of the out-of-sample forecasting ability of the model is presented in Section 5, for given weather conditions during a hypothetical fire.

# References

[1] European Comission, *Forest*, url: https://environment.ec.europa.eu/topics/forests_en. Last accessed on 02/04/2023.

[2] Natural Weather Service, *Discussion on Humidity*, url: https://www.weather.gov/lmk/humidity. Last accessed on 02/04/2023.