Investigating the Effect of Weather Coniditions on Forest Fire Burn Area

Conner Brew, Radhika Mardikar, Sho Tsuyuki

Introduction

Deadly wildfires have wreaked havoc across the globe in recent years, and Portugal has not been exempt from environmentally destructive disasters. In 2003, Portugal experienced wildfires that were described as "the worst in living memory" that took the lives of 18 people. Early detection and prevention of fire spread is the most effective way to combat the destruction of forest fires. This report aims to use wildfire and meteorological data to understand which factors most influence forest fires and predict areas more prone to them so preventive measures can be carried out more effectively. Hence, our research question is: Did weather conditions in August/September seasons in Montesinho Natural Park in Portugal have a causal effect on the overall area burned by forest fires between 2000 - 2003?

Data and Research Design

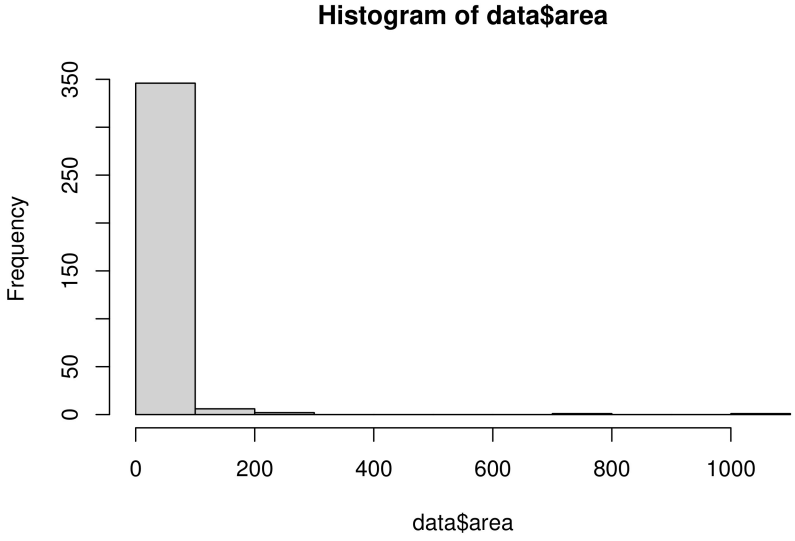
The dataset is available from the UC Irvine Machine Learning Repository. The data was collected from January 2000 to December 2003 and it was built using two sources. The first data set was collected by the inspector for the Montesinho fire. Several features were registered in the data collection – location, time, date, four FWI metrics, and the total burned area. The second data set was collected by the Bragança Polytechnic Institute. Weather observations, such as wind speed, temperature, rain, and relative humidity, recorded in 30 minute periods by a meteorological station in Montesinho Park were stored and used to generate the final dataset used in this analysis.

The features in the dataset are Fine Fuel Moisture Code (`FFMC`), Duff Moisture Code (`DMC`), Drought Code (`DC`), Initial Spread Index (`ISI`), `temperature`, rain accumulatio (`rain`), relative humidity (`RH`), and wind speed (`wind`). The `FFMC` denotes the moisture content of surface litter and is a measure of ease of ignition. The `DMC` and `DC` measure the moisture content of shallow and deep organic layers, which reflects the effect of seasonal drought. The `ISI` depends on wind speed and denotes the expected rate of fire spread. Although there are different scales for each of the FWI elements, higher values of each suggest drier conditions and therefore more severe burns.

Model Building Process

In order to understand the relationship between burned areas and weather patterns, this report will focus on most of the variables in the data set which, as mentioned above, can be largely separated into two categories: Forest Fire data (`FFMC`, `DMC` and `DC`) and weather data (`wind`, `temperature` and `RH`). After a glance at the entire dataset, it is evident that most of the forest fire damage measured in area burned occurs in the summer months of August and September. Thus the report will drop variables in all other months to study the data only in the months of August and September. Non-numeric variables such as `Day of the Week` and `Month` have been dropped since they will not be relevant for this research.

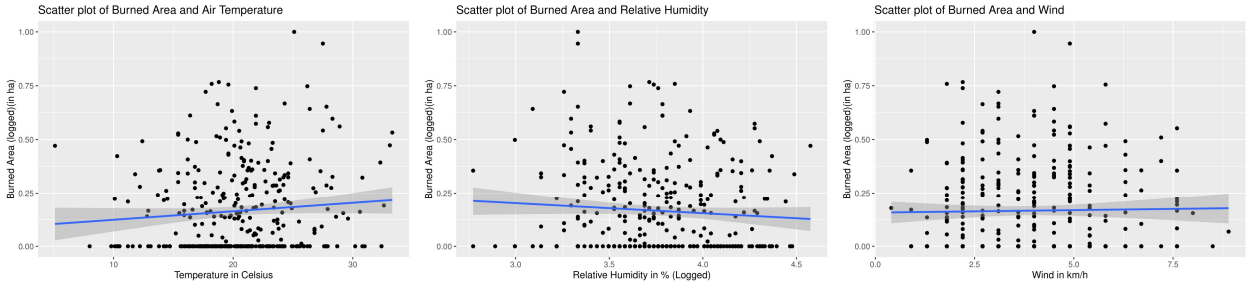
A sanity check of one of the dependent variables for our research, `area`, representing the area burned shows that the data is heavily skewed towards 0.0 and would be difficult to meet our model assumptions - thus a log transformation is necessary.



For the same reasons, we applied a log transformation to the features as follows: `rain` (the measured accumulation of rain in the 30 minutes prior to the fire), `RH`, and the `ISI`. The `FFMC` presented with a right-skewness that was resilient to log transformations, meaning that the log transformation alone was not sufficient to reduce skewness in the data. After iterative experimentation, we determined that a Box-Cox transformation most effectively normalized the feature. The Box-Cox transformation is designed to be a flexible solution which uses an optimized lambda value, selected by R through iterative processing, which best approximates a normalized distribution when applied to the Box-Cox function $\frac{y^{\lambda}-1}{\lambda}$.

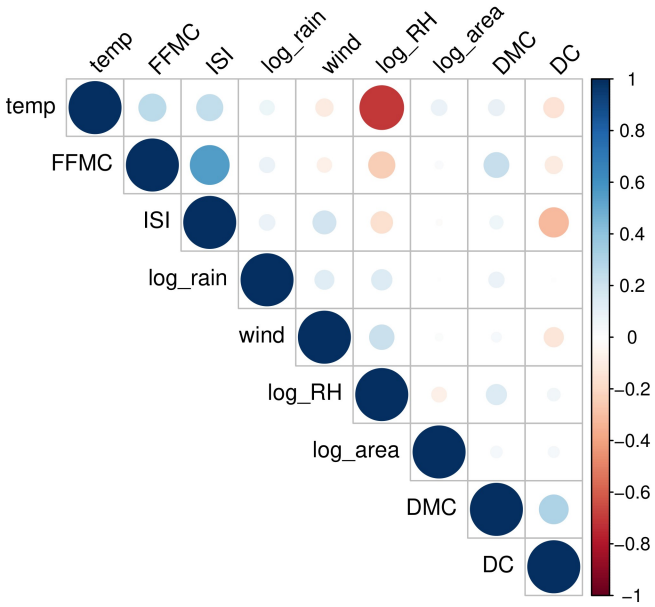
Exploratory Data Analysis/Transformations

After the initial data cleaning, we plotted some initial data that we expect to have the biggest impact on wildfires. Given the scale of the variables and the trends, we applied the appropriate transformations. For example, we plotted the log of the dependent variable `area` on the Y axis and the independent variable of interest such as `temperature`, `RH`, `FFMC`, and `wind` with the appropriate transformations on the X axis.



From the scatter plots above we can see that there may be a slight correlation between `temperature`, `RH` and `area`. Days with higher temperature and days with lower humidity (dry air) show a tilt, indicating a relationship with the size of the burned area. Wind speed however, does not show any correlation.

Finally, we looked at the correlation between all relevant variables in our dataset to make sure we understand what each variable represents and the relationship between one and the other.



The results of this exploratory data analysis indicates that `temperature` and `log_RH` have a negative correlation at -0.7, and `FFMC` and `ISI` have a correlation of 0.55. This plot also indicates that there is no perfect collinearity between the variables of interest.

We built five models to test the effects of groups of variables on the prediction of the dependent variable. The first model included only the environmental factors like wind, temperature and rain. As described above, other variables are derivatives of these and other environmental factors. The first model did not capture the trend of the dependent variable in a significant way, so other models were built by adding in the other independent variables (`FFMC`, `ISI`, `DMC`, `DC`, `X`, `Y`, `RH`) in groups. These were further evaluated.

Results

Our initial model attempt eschews all variables which are not direct measurements of weather conditions. This model includes `wind`, `temperature`, and `rain`. This model (model 1) alone did not reveal a significant indication of causation between weather features and area burned, so subsequent models (models 2, 3, 4) added more variables to seek features which may influence the amount of area burned. The longest model (model 5), which contained all the variables in the dataset, was compared to each of the smaller models in an ANOVA test to determine whether any subgroup of the variables was deterministic enough of the area burned. The results of the ANOVA test are summarized below (p-values):

Table 1: ANOVA Test Results

Anova	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Model 5	0.5196	0.4336	0.3381	0.1795

A significant p-value (<0.05) indicates that the shorter model is a better estimator than the longer model. An insignificant value indicates the longer model is better. From this ANOVA test, it is clear that the models with fewer variables are not robust predictors of the dependent variable. Model 5, the longest model, seemed to be the most promising one. The table below shows the regression table of model 5.

Following analysis of the regression table, we performed a t-test of coefficients of model 5 to determine significance of the individual coefficients.

Table 2: Results

	Dependent variable:
	log_area
boxcox_ffmc	0.127
	(0.210)
log_isi	−0.045
	(0.049)
log_RH	−0.040
	(0.052)
wind	0.008
	(0.007)
temp	0.002
	(0.004)
DC	0.0001
	(0.0002)
DMC	0.0001
	(0.0002)
X	−0.003
	(0.006)
Y	0.020*
	(0.011)
log_rain	−0.006
	(0.089)
Constant	0.057
	(0.331)
Observations	356
R^2	0.026
Adjusted R^2	−0.002
Residual Std. Error	0.205 (df = 345)
F Statistic	0.923 (df = 10; 345)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: T-Test of Model Coefficients

	Dependent variable:
boxcox_ffmc	0.127
	(0.191)
log_isi	−0.045
	(0.045)
log_RH	−0.040
	(0.055)
wind	0.008
	(0.007)
temp	0.002
	(0.004)
DC	0.0001
	(0.0001)
DMC	0.0001
	(0.0003)
X	−0.003
	(0.006)
Y	0.020
	(0.013)
log_rain	−0.006
	(0.200)
Constant	0.057
	(0.347)
Note:	*p<0.1; **p<0.05; ***p<0.01

From this table, it is apparent that no variable is a statistically significant predictor of area burned in the August/September time-frame. Intuitive logic and linear trends observed in exploratory analysis indicate that weather conditions should have an effect on forest fires. Therefore, we concluded that additional data collection is necessary to fully evaluate the effect of environmental factors on fires and amount of burned area.

Limitations

Assumptions (Statistical Limitations)

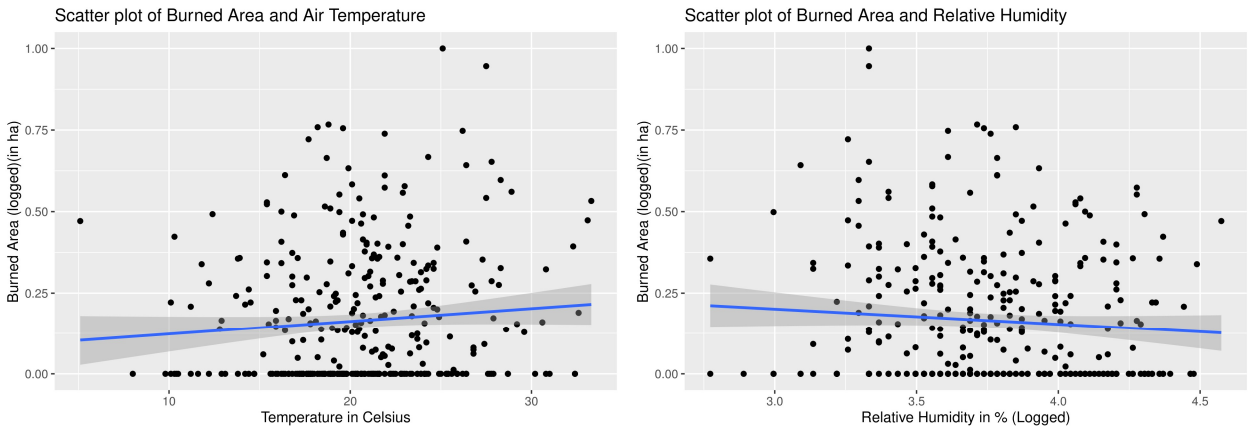
IID Data

The original dataset, which consisted of all forest fire data from 2000 - 2003 in the park, did not meet the assumption of IID data. Because of the temporal nature of the data over that time period, there was an inherent seasonality present in the data which disrupted the IID assumption. To mitigate the impact of this in our research, we narrowed our research scope to the months of August and September, focusing our analysis only on that season. Due to the adjacency of the months and similar climatic conditions, the effects of seasonality on the data were mitigated. The resulting data, which is specifically focused on forest fires in Montesinho Park in August and September months between 2000 - 2003, meet the assumption of IID.

No Perfect Collinearity

Initial exploratory analysis demonstrated correlation between some features, particularly `FFMC` and `ISI`. This is because `ISI` is, in part, modeled on the same weather conditions that support calculation of the `FFMC` for a particular area and time period. The `FFMC` is a measure of surface litter which influences the spread of fire, while ISI considers wind speed more heavily and thus is meant to better correlate with the velocity of fire spread. Therefore, while there is some degree of correlation between these modeled variables, there is no perfect collinearity and the data meets the assumption. This is also discussed briefly in the exploratory data analysis section.

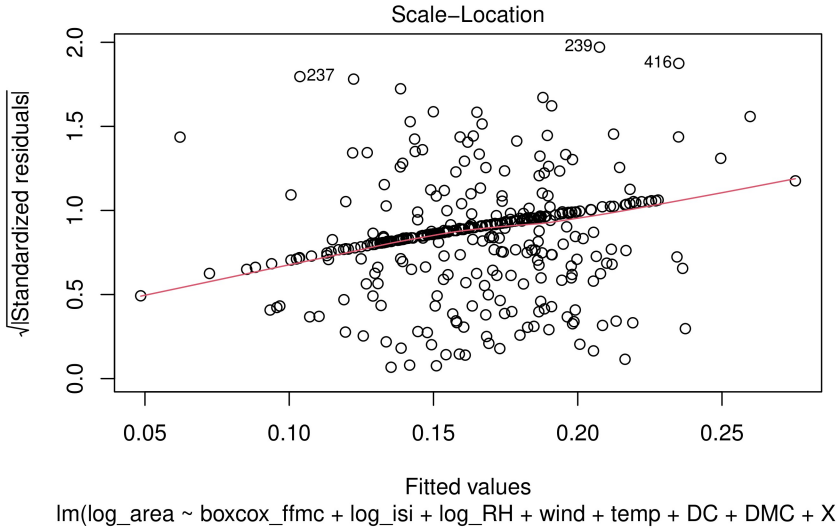
Linear Conditional Expectation



The features of `RH` and `temperature` appeared visually, in exploratory data analysis, to display a linear relationship with the target variable `area`. There is an observable positive linear relationship between air temperature and area burned, and an observable negative linear relationship between relative humidity and temperature. In addition to these observations, it is intuitively logical that factors such as low humidity,

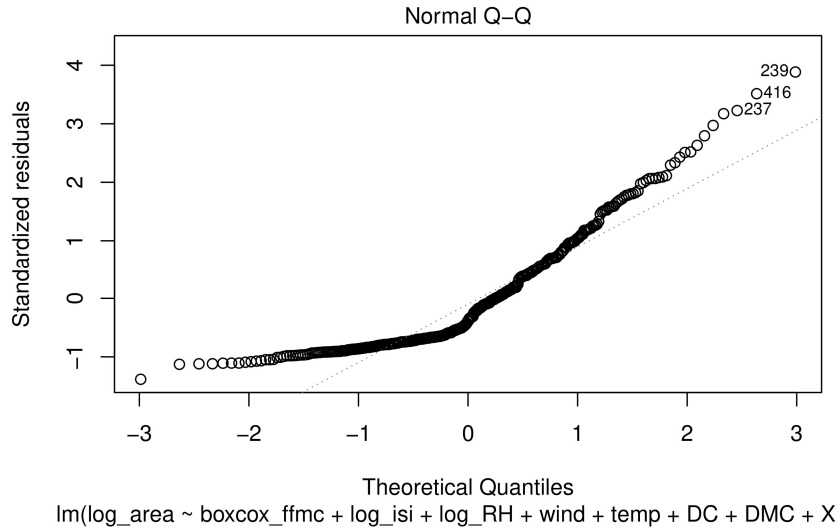
thereby reducing moisture which might disrupt a burn, would have a linear relationship with the amount of area burned. Therefore, the data meets the assumption of linear expectation.

Homoscedastic Errors



To test this assumption, we conducted a scale-location plot analysis. The scale-plot analysis, when the assumption is met, should produce a line midway through the scatter plot of fitted values and residuals (standardized) wherein the distribution of points are evenly distributed above and below the line. We tested the model and found that the points are observably evenly distributed around the line, and the model therefore meets the assumption of homoskedasticity.

Normally Distributed Errors



To test the normal distribution of errors, we plotted a quantile-quantile plot (Q-Q plot) to visualize the quantiles of residuals versus theoretical quantiles of the fitted model. To meet the assumption, the plot should show a positive linear relationship between standardized residuals and theoretical quantiles with minimal outliers. The model achieves this, and therefore meets the assumption for normally distributed errors.

Omitted Variables (structural limitations) and their directions

The modeling process was likely influenced by the omission of multiple variables which did not exist in the collected data. These variables include Air Pressure, Vegetation/Land Cover, and Slope of Terrain. The omission of these variables in the originally collected dataset likely introduced bias into the modeling process, and therefore collection of data associated with these variables should be taken into account in future modeling attempts.

Air pressure is an omitted variable in this dataset. Air pressure is an important indicator of weather conditions and affects `wind`, `ISI`, and other measurement factors that were included in the analysis. Going hand in hand with air pressure is elevation, which also affects the propensity of the given area to burn. In general, higher air pressures are associated with more dry and sunny conditions while lower air pressures are associated with rainy or stormy conditions. Based on this, air pressure should be positively correlated with the dependent variable. The direction of the bias is away from zero.

The slope of the land also introduces omitted variable bias due to the fact that it affects the sparsity or denseness of vegetation along the land and also influences the `ISI` which measures the spread of fire. The slope of land is negatively correlated with vegetation growth in that a greater slope is less likely to have more dense vegetation. Depending on other factors such as elevation, slope may also be negatively correlated with the amount of area burned by a forest fire, so the direction of bias is towards zero.

Vegetation and land cover, due to its absence, introduces omitted variable bias. Tree density has a positive linear relationship with area of fire burn, as the proximity between trees or other fire-susceptible foliage increases the capability of the fire to spread from plant to plant. Lower tree density would in turn have an effect of reducing overall burn area. Because there is a positive correlation between vegetation and burn area as well as a positive correlation between `ISI`, `FFMC`, and vegetation, there would be a resulting positive bias away from zero which overestimates the coefficient associated with the included variables.

Conclusion

The purpose of this study is to determine the causality of weather conditions on the amount of burned area in Montesinho in the August/September time frame. Based on our exploratory data analysis and statistical analysis, there is no clear effect of any of the independent variables on the dependent variable (`area`). Since this result does does not align with the linear expectation observed between the target variables and the independent variables and is also counter to intuitive logic regarding the relationship between forest fires and environmental conditions, we propose conducting this study in the future with a more comprehensive dataset.