Predicting Forest Fire Burn Area in Montesinho Natural Park, Portugal

By Vicky Wang

Introduction

Montesinho Natural Park is a protected area in the northeastern region of Portugal that boasts a wide variety of natural landmarks and fauna populations native to the country. In recent years, wildfires in Portugal have become a bigger problem, not only due to insufficient firefighting resources, but also increasing concerns over climate change (Elbein 2021). Therefore, it is important to maintain a preemptive response to wildfires by examining potential factors that influence fire strength and burn size. This report will attempt to predict forest fire burn area by analyzing a variety of predictors and also give firefighters a better understanding of conditions that could lead to major wildfires.

The predictors used to formulate the model chosen are as follows:

area- the burned area of the forest (in ha)

month (of the year)

day (of the week, Mon-Sun)

FFMC- Fine Fuel Moisture Code, from 0 (dryest) to 101 (wettest)

DMC- Duff Moisture Code, or the organic material moisture content

DC- Drought Code (0 to 1000)1

temp- in Celsius degrees

ISI- initial spread index, a unitless value of spread potential

RH- Relative Humidity (%)

wind- speed in km/h

rain- in mm/m^2

The question being considered for this data set is, “What predictors have the biggest effect on the burn size of forest fires in Montesinho Natural Park, Portugal?” Based on preliminary findings, there seems to be no clear answer, but further analyses, as shown below, provided slightly more convincing evidence for a model for the relationship between several of the predictors and “area”.

In the following report, initial data exploration is first summarized. Next, after using a tentative “best” model generated through forward selection, variable exploration is done to show how manipulating some aspects of the data might help increase the predictability of a fitted model. Then, a final conditional mean function was reported and verified after use of model comparison methods.

Initial Data Exploration

During initial exploration of the data set used, a handful of predictors were used to narrow down the scope of the analysis done. As a result, three numerical predictors and one categorical predictor were

singled out to help determine how they can contribute to burn area. Using practical reasoning, it seemed as though wind, temperature, moisture (FFMC index), and month could potentially have an impact on fire size, but the initial analysis indicated otherwise. After using R to generate a scatterplot matrix relationship between the above predictors, it was found that very few showed a significant or clear relationship with one another. The conclusion was that transformations or alterations might help with overall predictability, so in further analyses, new models were created to see if that held true.

Original analysis also showed two major outliers in the data set that had the effect of significantly skewing the variable "area" to the right. The fires had burn sizes of 700+ and 1000+ hectares, which when compared with the numerical summary of the area variable in Fig. 1 (where these outliers were removed), are *much* higher than the average area of 17.93 ha. As a result, the outliers were omitted so that the model could show a more accurate representation of data.

Transformations

When analyzing the individual histograms for numerical variables, the graph for area was found to have a significant right skew—even with the removal of the outliers. Therefore, it made sense to introduce a log transformation on "area". This transformation is shown in Fig. 3, and very clearly shows the distribution for log(Area) becoming fairly normal in shape. The remaining histograms for each numerical predictor showed either no clear pattern or an approximately normal distribution, so they were not further manipulated. Additionally, when considering a quadratic fit, no individual predictor showed a quadratic-like relationship when plotted against burn area on the y-axis. Therefore, it did not make sense to introduce a quadratic transformation to any of the tested variables.

Categorical Predictors

The data set originally provided two categorical predictors; "month" and "day" (of the week). It can be reasonably assumed that the day of the week is generally insignificant towards impacting a model fit, so no further analysis was done on the "day" variable. However, the month that a fire occurred may serve as a predictor of burn size, as Fig. 6a shows via side-by-side boxplots. Therefore, a model was created using "month" as a categorical variable. In order to check whether the "month" variable had a meaningful impact on the full model, two linear regressions were created (Fig. 6b); one with all of the predictors, and one with all but month included. This showed that the regression model that includes month is actually a better fit than the one without, since the adjusted R-squared for the first model was 0.0421, which is higher than the adjusted R-squared for model 2 (without month): 0.0075. As a result, the "month" *should* be included in the final model because it improves the overall fit of the data by considering what unique conditions certain months may bring when influencing a fire's burn area.

Interaction Term

Based on theory alone, one might assume that some of the predictors in the set might show a dependency on another. For example, it is reasonable to conclude that temperature and month might have some relationship, which is important to consider when factoring in what time of the year will show optimal temperature conditions for a wildfire. Fig. 7a highlights this association because the side-by-side boxplots show very different temperature ranges depending on the month. However, when comparing interaction-less and interaction regressions between month and temperature, the AIC values were 972.28 and 981.36, respectively (Fig. 7b), meaning the model with no interaction was a better model. Therefore, an interaction between month and temperature should *not* be included in the model because it actually decreases the predictability of the relationship (See Fig. 7c for visualization).

Final Conditional Mean Function

$$E(\log(Y_{area}) \mid X) = \beta_0 + \beta_1 X_{ISI} + \beta_2 X_{DC} + \beta_3 X_{DMC} + \beta_4 X_{wind} + \beta_5 X_{temp} + \beta_6 Z_{mar} + \beta_7 Z_{apr} + \beta_8 Z_{may} + \beta_9 Z_{jun} + \beta_{10} Z_{jul} + \beta_{11} Z_{aug} + \beta_{12} Z_{sep} + \beta_{13} Z_{oct} + \beta_{14} Z_{dec}$$

Assumptions of Ordinary Least Squares

Linearity: The linearity assumption is reasonably met by the conditional mean function because the residuals vs fitted values plot for the data is fairly random in pattern (Fig. 8a).

Constant variance: Constant variance is essentially met because the variability in errors around the regression line is fairly constant for almost all values of the predictor variable (Fig. 8a).

Normality: Since the QQ plotted points generally follow the pattern of the QQline, the normality assumption can be considered to be reasonably met (Fig. 8b).

0 Mean: The error distribution as shown by the residuals vs fitted values plot is approximately centered around 0 (Fig. 8a).

Independence: There is no evidence in the data description to suggest that the fires were independently sampled, so it is difficult to verify if independence assumptions are met. Each forest fire might represent an individual observation, but it is also plausible that one forest fire was somehow caused by a different fire but they were recorded as individual observations.

Random Sample: Every fire over the duration of the sampling period was recorded and included in this data set. This is not 100% guaranteed to be the best representation for all fires that ever occurred in Montesinho Park, but for the sake of the model, is most likely close enough.

Multicollinearity and Overfitting

As shown in Figure 5, since all VIF values for the model are below 0.90 and less than 10, there is no evidence of multicollinearity. Furthermore, there is no evidence of overfitting because there are 267 samples in the data set, which is greater than 10 times the number of predictors (6) used.

Interpretations

The interpretations for each of the β values in the model are as follows:

- β_0: The p-value for "Intercept" (0.002) tests if the burn area for fires recorded in the month of February is different from 0.
- β_1: The p-value for "ISI" (0.274) tests if the value of the slope for the ISI index is different from 0. It is greater than 0.05, so the β_1 coefficient is not significant.
- β_2: The p-value for "DC" (0.023) tests if the value of the slope for the DC index is different from 0. It is less than 0.05, so it is statistically significant.
- β_3: The p-value for "DMC" (0.001) tests if the value of the slope for the DMC index is different from 0. It is less than 0.05, so it is statistically significant.
- β_4: The p-value for "wind" (0.401) tests if the value of the slope for wind is different from 0. It is greater than 0.05, so the β_4 coefficient is not significant.
- β_5: The p-value for "temp" (0.181) tests if the value of the slope for temperature is different from 0. It is greater than 0.05, so the β_5 coefficient is not significant.
- $\beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}$: The coefficients for β_6-β_{14} test for a difference in the *intercepts* between the reference category of February and a month corresponding to the β values. All months but "oct" are not statistically significant at 0.05 significance.

In terms of coefficients that are both practical and statistically significant, only the "monthoct" coefficient is meaningful. It shows a p-value of less than 0.05l meaning it is statistically significant, and fires that occur in the month of October typically correspond with a much larger increase in burn area. All others are essentially negligible in their ability to affect burn size OR do not have a statistically significant p-value at the 0.05 significance level.

Model Fit

Multiple R-squared Value: 0.0923. Since the R-squared value is very low (almost 0) for this model, that means that only 9.23% of the variance in the model can be explained by the regression relationship between area and ISI, DC, DMC, wind, temp, and month.

F test p-value: 0.0339. Since the p-value for the F test of our model is very low, at 0.05 significance, this model is statistically significant and therefore is a good general predictor of burn size using the mentioned predictors.

Conclusion

By using forward selection and evidence based on testing transformations and interaction models, the "best" predictive model was found for the data set with Montesinho Natural Park wildfires. Given the raw data for this set, there is no absolute guarantee that the recorded values taken from each observation are relevant to predicting overall burn size, but there does seem to be some sort of predictive power for the model. Interestingly, in the *Model Fit* section, summary output for the final model showed a low multiple R-squared value but a statistically significant F test p-value. This suggests that while the general trend for the model is statistically significant, the model should not be used to predict specific values or conditions because the variability about the regression line for individual observations is too great. Perhaps in including better predictors and more samples, a better model can be generated. Even though analysis did not provide the most conclusive of results statistically, creating such a model is still important because it helps give, at the very least, an idea about what and how factors are involved with forest fires.

Sources

Elbein, Saul. "What Portugal's Hellish Wildfires Can Tell Us about Forest Futures." Science, 3 May 2021, nationalgeographic.com/science/article/how-to-live-with-mega-fires-portugal-forests-may-hold-secret.