ALGERIAN FOREST FIRE ANALYSIS

BY SAHAL SAJEER KALANDAN[1,a]

[1] Candidate for Honours Statistics , [a] ssajeer@uwaterloo.ca

1. Introduction.

1.1. *Motivations.* Forest fires continue to be a growing threat to ecosystems all over the world. In the year 2023 alone, Samborska and Ritchie report that approximately 400 million hectares of forestation was lost through fire, of which 7.2 % came from Africa alone. Algeria in particular remains incredibly sensitive to these fire due to the nature of its arid climate, terraneous landscape and high wind-speeds, a trait commonly shared by other countries in the Mediterranean basin [Belgherbi, Benabdeli and Mostefai (2018)].

Criticism continues to grow towards the approaches taken by the government, specifically in regard to how methods of prevention are implemented. However, recent years have shown some progress with collected data being analyzed towards this task. The Centre of Development of Advanced Technologies have taken the initiative to collect data (see Section 2.1) comprising of different meteorological features and forest fire indices [Abid (2019)].

This data was then set up for the task of predictive modelling through a decision tree[1] to determine the occurrence of a forest fire in 2 different forested regions in Algeria [Abid and Izeboudjen (2020)]. It proves to be instrumental in resolving the issues that Belgherbi, Benabdeli and Mostefai discuss, but it does not address the matter of analyzing how *important* these factors are, as well as how they determine the level of *risk* towards the occurrence of a forest fire.

1.2. *Research Topic.* In this report, the data provided by Abid will be repurposed instead for the task of regression. A kNN[2]-based system is introduced to facilitate the conversion of the binary occurrences of forest fires into risk probabilities in order to answer the question of *how well the covariates present in the dataset can explain the occurrence of forest fires*.

2. Data.

2.1. *Description of Features.* The data used in this study features 243 observations collected from June 2012 to September 2012 in two heavily forested regions of Algeria: Bejaia and Sidi Bel-Abbas (labeled accordingly). Besides a description of the date the observations were taken, the features involved can be split into two categories: meteorological features and fire weather indices (see Table 2).

All covariates corresponding to the meteorological features were measured discretely (with the exception of Rain). Fire weather indices are treated as continuous covariates.

[1] Decision Tree: A predictive algorithm that splits the data into branches based on feature values, where nodes represent features, and branches represent a decision rule. Target outcomes are determined by the leaves of the tree.

[2] kNN: k-nearest neighbours is a simple algorithm that predicts the value of a target variable by averaging or voting on the k closest data points in a feature space

2.2. *Description of Response.* The data initially provided from Abid classify outcomes in 2 categories: 0 for the occurence of a fire for a given observation, and 1 otherwise. In order to construct a continuous response variable suited for the task of regression, the response variable will be calculated by finding local neighbourhoods[3] (based on the features) of any given observation, and then considering its surrounding observations in those neighbourhoods (see Eq. 5).

For each neighbourhood, a probability of how likely a fire is to happen will be computed via the number of observations that correspond to a fire versus the total number of observations in that neighbourhood. This is first done by constructing $n = 100$ bootstrap samples of the observations in order to ensure that the resulting probability estimates are "smooth" so that the distribution of the response can be treated as a continuous random variable.

Then, for any given observation in each bootstrap sample, once we consider a certain number of neighbourhoods, the average of those probabilities will be associated with that observation as the response, and this process will be repeated for each observation and each bootstrap sample. In doing so, the response variable is transformed from binary to continuous, making it viable for regression.

In deciding which features would work best the following plot corroborates with the results discussed by Abid and Izeboudjen (2020), detailing how Rain has no impact on influencing the decision tree[4]. Hence, it was excluded in the construction of the response.

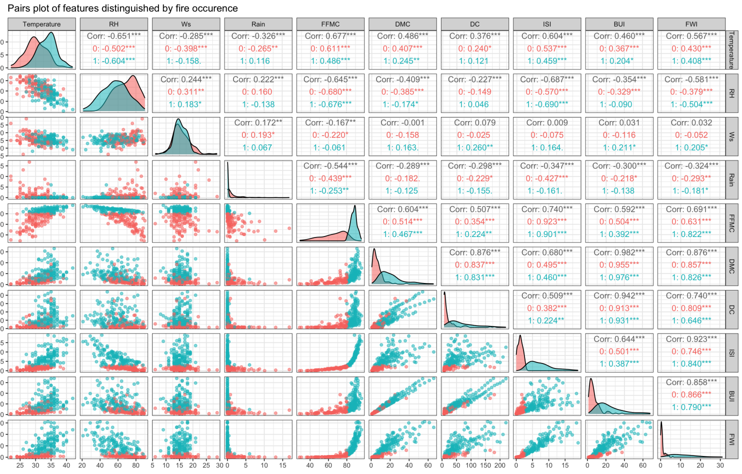


Fig 1: Scatterplots (Lower-Triangular), Correlation and associated relationship (Upper-triangular), and Density plots categorized by outcome (Diagonal)

Empirical CDF[5] plots revealed that the best value of σ to choose for the cutoff is 1 as it had the most amount of variation between the risk probabilities (see Fig 5). The calculated risk probabilities were constructed so that the higher the probability, the higher the risk of fire (see Eq. 5). To further refine the analysis, redundant observations of probabilities of 0 and 1 were omitted so that only patterns in variation could be analyzed, leaving a total of 164 observations to work with.

[3] Neighbourhoods are determined by a Gaussian Kernel with a cutoff value (σ) corresponding to how many standard deviations away neighbours can be from a given observation for them to still be considered in the same neighbourhood, and then weighted based on their influence (ex. closer values are more influential)

[4] It is reasonable to assume since the climate of Algeria is generally receives little rainfall in the summer time; it does not have enough variability and can hinder the influence of how the response would be determined

[5] CDF: Cumulative Distribution Function

2.3. *Initial assumptions.* In combination of the exploratory analysis and construction of the response, it is reasonable to assume that due to a lack of collinearity[6], the meteorological data will be more useful in analyzing meaningful trends in how the risk of fire occurences can be predicted.

3. Methods.

3 approaches were considered when tackling regression[7]: Multiple Linear regression, Ridge Regression, and Generalized Additive Models. Each were setup so that first all features were considered, then only the meteorological features, and finally only the fire weather indices. In each model description, only the general form shall be provided. All models undergo the assumption of residual normality and homoskedasticity[8]. Description of notation used can be found in Table 3.

3.1. *Multiple Linear Regression.*

$$\widehat{y^{\text{resp}}} = X\beta + \epsilon_i \tag{1}$$

such that:

$$\hat{\beta} = \arg\min_{\beta} \|y^{\text{resp}} - \mathbf{X}\beta\|_2^2 \tag{2}$$

3.2. *Ridge Regression.*

$$\widehat{y^{\text{resp}}} = \mathbf{X}\beta + \epsilon_i \tag{3}$$

such that:

$$\hat{\beta} = \arg\min_{\beta} \|y^{\text{resp}} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{4}$$

3.3. *Generalized Additive Models.*

$$\widehat{y^{\text{resp}}} = \alpha\mathbf{1}_n + \mathbf{X}\beta + \epsilon_i \tag{5}$$

such that:

$$\mathbf{X}\beta = (f(x_1), \ldots, f(x_n))^T \tag{6}$$

$$\hat{\alpha}, \beta = \arg\min_{\alpha,\beta} \|y^{\text{resp}} - \alpha\mathbf{1}_n - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^{p} \int \left\{ f_j^{''}(x_{ij}) \right\}^2 dx \tag{7}$$

4. Results.

Despite the prior assumptions made about collinearity, it appears that in all 3 cases, either the inclusion or the model solely dependent on the fire weather indices outperform the models that solely consist of provided meteorological features (based on the GCV[9] and Adjusted R^2[10]). The model that performs the best based on 1 is the model that uses the entire set of features (date, meteorological features, fire weather indices, and region).

A deeper investigation into the residual analysis of each model (see Fig **??**) showcases some form a linear trend in the residuals vs. fitted values, as well increasing and decreasing variance. Both of these occurrences suggest that not only do the models fail to capture some of the information provided by the constructed risk probabilities, but they also fail to entertain the assumption that the variance remains constant throughout.

[6] The strength of the association between any 2 features

[7] with respect to the L2-norm

[8] Constant variance in residuals. Otherwise, it is reasonable to assume that the model is biased towards the data

[9] Generalized Cross Validation Error estimates the predictive performance of the model by balancing model bias with model variance

[10] Measure of how well variation in features explain the variation of the response

TABLE 1
GCV Scores and Adjusted R^2 for Different Models and Feature Groupings

Model	**All Features**		**Meteorological Features**		**Fire Weather Indices**	
	GCV	**Adj.** R^2	**GCV**	**Adj.** R^2	**GCV**	**Adj.** R^2
Multiple Linear Regression	N/A	0.864	N/A	0.150	N/A	0.854
Ridge Regression	0.021	0.860	0.122	0.148	0.022	0.851
Generalized Additive Models	**0.010**	**0.935**	0.119	0.177	0.011	0.927

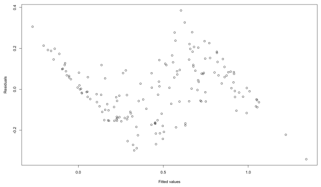


Fig 2: Residuals of Multiple Linear Regression Model

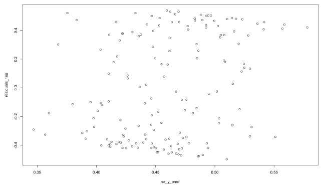


Fig 3: Residuals of Ridge Regression Model

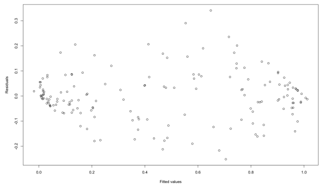


Fig 4: Residuals of GAM Model

Ridge regression appears to have the least amount of trend present in the residuals, and maintains some notion of homoskedasticity as compared to the other plots. Intuitively, this is consistent with what is expected from the purpose of the fire weather indices: they serve to describe features that directly correlate to the occurrences of forest fires.

5. Conclusions. In an effort to explore how calculated risk probabilities of fire occurrence can be explained by different measures provided in the Algerian Forest Fire dataset, ridge regression using the fire weather indices proves to be the most effective not only in predictive power, but also in best representing the relationship of the features based on model diagnostics.

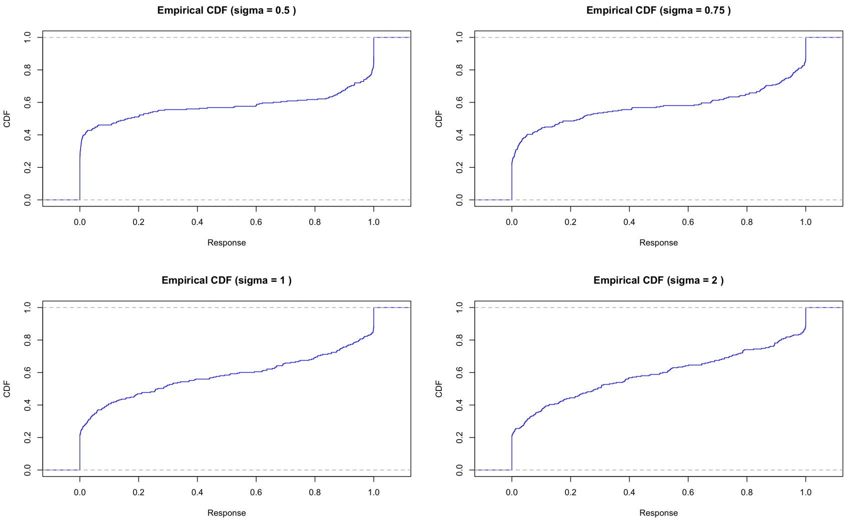
What remains interesting is the ineffectiveness of the meteorological features in being able to explain the occurrences of forest fire, and how inclusion of date and region seem to leave only more factors unexplained in the residuals. Future studies should directly analyze the influence of each individual feature, and nested interactions between the features to see what may be hidden beyond regression and decision trees. One possible recommendation is understanding the importance of how clustering of points in the GAM residual plot suggest that a more in-depth time-series analysis could be effective to see if there were underlying patterns in the summer of 2012 influencing the occurrences of these fires.

This report only scratches the surface of the potential that predictive modeling can hold to solve issues of forest fires across the world. While the task of regression holds limitations to the modelling accuracy of the risk probabilities, it sets a hopeful precedent for how forest fires can be prevented before they even begin.

APPENDIX: RESPONSE CALCULATION

```
For each bootstrap sample $b$ from $1$ to $B$ :
  $\{\mathbf{X}_b, \mathbf{y}_b\} = \{\mathbf{X}_{i_b}, \mathbf{y}_{i_b}\} \quad \text{where} \quad i_b \sim \text{Uniform}(1, n)$
  For each observation $i$ in the dataset, find the $k$-nearest neighbors from the bootstrap sample:
    Calculate the weights for each neighbor based on the distance:
      $w_{ij} = \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{X}_j\|_2^2}{2\sigma^2}\right)$
    Normalize the weights:
      $\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{j=1}^{k} w_{ij}}$
    Calculate the response probability $\widehat{y^{\text{resp}}}_{ib}$ for each observation $i$ :
      $\widehat{y^{\text{resp}}}_{ib} = \sum_{j=1}^{k} \tilde{w}_{ij} \mathbf{1}\{y_{j_b} = 0\}$
Average the response probabilities over all bootstrap samples to get the final response $\widehat{y^{\text{resp}}}_{ib}$:
  $\widehat{y^{\text{resp}}}_i = \frac{1}{B} \sum_{b=1}^{B} \widehat{y^{\text{resp}}}_{ib}$
```

APPENDIX: FIGURES



Empirical CDF (sigma = 0.5)

Empirical CDF (sigma = 0.75)

Empirical CDF (sigma = 1)

Empirical CDF (sigma = 2)

Fig 5: Empirical CDF of response with $n = 100$ bootstrap samples

TABLE 2

Feature descriptions and value intervals for the regions

		Values Interval		
Attribute Name	**Description**	Sidi-Bel-Abbes Region	Bejaia Region	Both Regions
Meteorological Features				
Temperature	Temperature in Celsius degrees	24 to 42	22 to 37	22 to 42
RH	Relative humidity in %	21 to 90	45 to 89	21 to 90
Ws	Wind speed in km/h	6 to 29	11 to 26	6 to 29
Rain	Outside rain in mm/m^2	0 to 16.8	0 to 8.7	0 to 16.8
Fire Weather Indices				
FFMC	Fine Fuel Moisture Code	28.6 to 96.0	36.5 to 92.5	28.6 to 96.0
DMC	Duff Moisture Code	0.7 to 65.9	1.1 to 110.2	0.7 to 291.3
DC	Drought Code	6.9 to 220.4	7 to 860.6	7 to 860.6
ISI	Initial Spread Index	0.0 to 19.0	0 to 25.7	0 to 56.1
BUI	Buildup Index	1.1 to 68.0	1.1 to 109	1.1 to 252.4
FWI	Fire Weather Index	0.0 to 31.1	0 to 31.1	0 to 77.5

TABLE 3

Notation and their relevance in the context of the models

Notation	**Description**
$\mathbf{X}$	The design matrix containing all 164 observations and the respective predictor variables. In the context of Generalized Additive Models (GAMs), $\mathbf{X}$ is defined as the spline design matrix with pre-determined with 10 knots.
β	The vector of coefficients applied to the predictor variables. In Multiple Linear Regression and Ridge Regression, it will determine the linear contribution of the features to the risk probabilities. In the case of GAMs, β they form coefficients for the B-spline basis functions, which are used to form the smooth functions $f_j(x_{ij})$.
λ	The regularization parameter used in Ridge Regression. It controls the strength of the penalty applied to the coefficients β such that a higher λ values increase the penalty, and identifies features that may lack relevance in predicting risk probabilities.
$f(x_i)$	The vector of cubic B-splines representing smooth functions of the predictor variables in GAMs. Each $f(x_i)$ is enforced with a point constraint to ensure smoothness. The B-splines will attempt to identify if any non-linear relationships exist between the features and the risk probabilities.
ϵ_i	Residual term after each model fit, under the assumption that $\epsilon_i \sim N(0, \sigma^2)\ \sigma^2$.
α	The intercept term in GAM describing the mean risk probability

APPENDIX: TABLES

REFERENCES

ABID, F. (2019). Algerian Forest Fires. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5KW4N.

ABID, F. and IZEBOUDJEN, N. (2020). Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)* (M. EZZIYYANI, ed.) 363–370. Springer International Publishing, Cham.

BELGHERBI, B., BENABDELI, K. and MOSTEFAI, K. (2018). Mapping the risk forest fires in Algeria: Application of the forest of Guetarnia in Western Algeria. *Ekológia (bratislava)* **37** 289–300.

SAMBORSKA, V. and RITCHIE, H. (2024). Wildfires. *Our World in Data*. https://ourworldindata.org/wildfires.