AN APPLICATION OF REGRESSION MODELS FOR ANALYSIS AND FORECASTING OF FIRE EXTENT

Luigi Passariello (Senior Technologist at INGV*), Giuseppe Passariello (Data Scientist at MAPACOM**), Michele Passariello (Data Scientist at MAPACOM**), Alessandor D'Apice (MAPACOM**)

* National Institute of Geophisics and Vulcanology (INGV), Rome, Italy
** Ma.Pa.COM.

Abstract

Every year wildfires cause destruction that results in the loss of human and animal lives and economical and ecological losses. In addition, unpredictable forest fires make it very difficult to plan suppression actions in light of saving monetary resources while effectively fighting all the active fires. By creating a model that uses meteorological measurements such as wind and relative humidity, it should possible to prepare better in order to fight wildfires more effectively and ideally decrease the amount of physical and economic damages. A dataset of forest fires identified in Montesinho Natural Park, located in the mountainous northeast of Portugal, was used for application of our model. While the analyzed data were not collected real time, our design can be applied to real-time fire management. The study identifies specific actions combating wildfires with low cost equipment, suggesting when more specialized equipment like satellite imaging and smoke scanners provide the required information to fight fires, especially the largest ones.

Key-words: Mediterranean region; economic damages; suppression; prevention; global change.

INTRODUCTION

Every year wildfires cause destruction that results in the loss of human and animal lives and economical and ecological losses. In addition, unpredictable forest fires make it very difficult to plan suppression actions in light of saving monetary resources while effectively fighting all the active fires. By creating a model that uses meteorological measurements such as wind and relative humidity, it should possible to prepare better in order to fight wildfires more effectively and ideally decrease the amount of physical and economic damages. A dataset of forest fires identified in Montesinho Natural Park, located in the mountainous northeast of Portugal, was used for application of our model. While the analyzed data were not collected real time, our design can be applied to real-time fire management. The study identifies specific actions combating wildfires with low cost equipment, suggesting when more specialized equipment like satellite imaging and smoke scanners provide the required information to fight fires, especially the largest ones. More specifically, the objective of this study is to build a model that predicts the burned area of the forest given the explanatory variables of measurements collected. Additionally, a temporal model was developed to understand which day and weather patterns offer the highest risk conditions. We believe the achievement of the project objectives requires answering the following issues: (i) during which months are forest fires most common? (ii) On which days of the week are forest fires most common? (iii) How do the measured data affect forest fires?

MATERIALS AND METHODS

Our dataset has 517 observations and 13 columns. None of the variables have any missing values. All of the data are numeric except for day and month. It is also interesting to note that our response variable (area), has a large range and the difference between the median (0.52) and mean (approximate 12) seems to suggest there maybe outliers in our data. There are 247 observations were our target variable (area) is 0. This is not not completely appropriate in a basic assumption characteristic of simple linear models, and for this reason, we avoided considering them. Below is the description and abbreviations of the variables of interest in the data set (Table 1). These abbreviations will be used throughout our work.

Table 1. Variables considered in this study.

VARIABLE NAME	DEFINITION
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
month	month of the year: “jan” to “dec”
Day	day of the week: “mon” to “sun”
FFMC*	Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
DMC*	Duff Moisture Code index from the FWI system: 1.1 to 291.3
DC*	Drought Code index from the FWI system: 7.9 to 860.6
ISI*	Initial Spread Index from the FWI system: 0.0 to 56.10
temp	temperature in Celsius degrees: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
wind	wind speed in km/h: 0.40 to 9.40
rain	outside rain in mm/m2 : 0.0 to 6.4
area	the burned area of the forest (in ha): 0.00 to 1090.84

There are not missing Attribute Values: The FWI, or Fire Weather Index system, is an estimator developed in Canada for assessing fire risk. It ranges 0-20 and considers weather and fire conditions withing a critical period prior at the start of the fire. The FFMC is a composition of rainfall, humidity, temperature and wind data; DMC is rain, humidity and temperature, DC is rain and temperature, and finally ISI is a fire behavior index. One element of FWI, the build-up index, was not included as this is an indicator of the bulk fuel on the ground and not ascertainable from meteorology. In order to look further into the relationships between our response variable, area, and each of the predictor variables (see also appendix for further graphical details), we imposed a linear fit (with 95% confidence band in grey) to each of the pairs of variables to understand the underlying relationships. We see some outliers in the data. Figure 1 (left) shows a map of Montesinho Natural Park divided into a 9x9 raster matrix; wildfires were recorded based on the position inside the matrix, using MAPS library functions, as exemplified in Figure 1 (right).

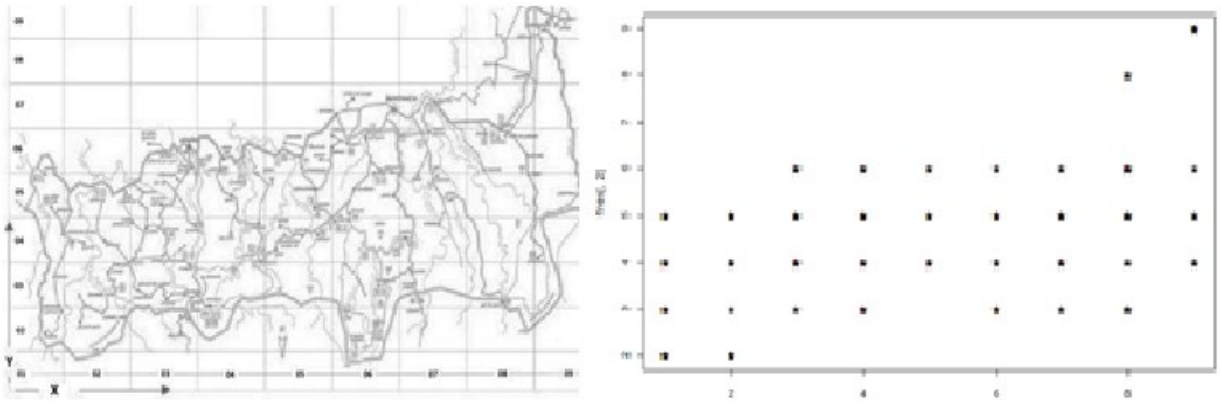


Figure 1. A map of the study area (left) and the position of each recorded forest fires in respect with the regular grid developed in this work (right).

RESULTS AND DISCUSSION

A descriptive approach

Concerning the issues (i) and (ii) posed in the introduction chapter, we made a descriptive analysis of the available data for Montesinho Natural Park (Figure 2).

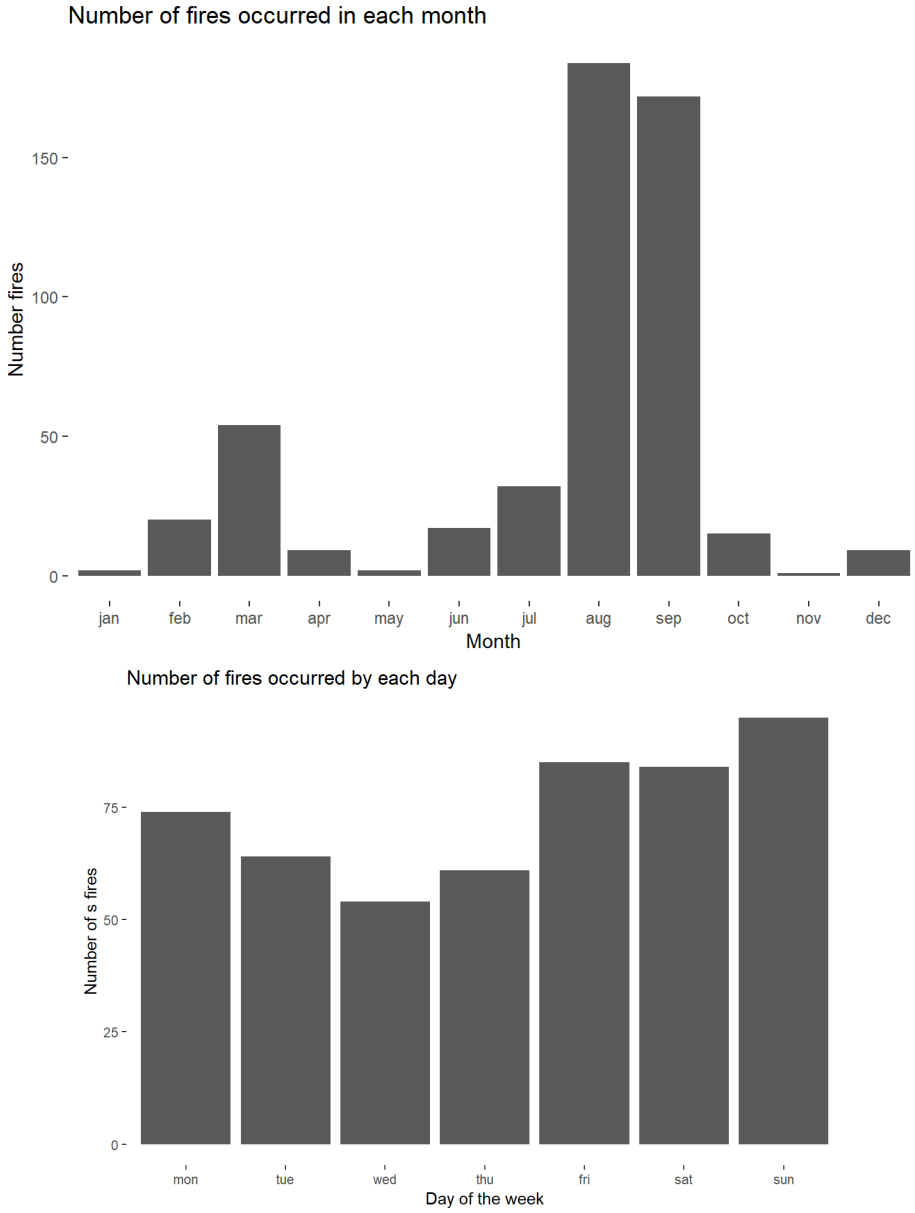


Figure 2. A descriptive analysis of selected wildfire variables in the study area.

The majority of fires that occurred happened during the months of August and September. Weekend days (Friday, Saturday and Sunday) were the most common days for fires' start. Additionally, the rainfall variable seems to be not particularly important since the study area in Portugal did not receive much rainfall. Therefore, to prevent accidents in the Natural Park it is advisable to keep it constantly monitored with the presence of firefighters and immediate resources to manage wildfires. A strengthened enhancement is required in the months of August and September. As regards the day of greatest incidence, the data do not allow a significant distinction to be made between the various days of the week, although there is a greater

concentration around the weekends. In this perspective, factors affecting the distribution of wildfires are related not only to air temperatures but also to the presence of visitors. Concerning the issue (iii), namely the influence of the measured variables on fire size (i.e. total burnt area), we first understood what information can be obtained from a graphical data analysis. We investigated the factors influencing burned area as a possible indication of fire severity. The scatter plots in Figure 2 allowed determining the initial relationship between each individual variable and log-transformed area. The log transform (log(x+1)) of the area was taken in order to make the data less skewed (the data is heavily skewed towards zero acres being burned), while not invalidating the analysis. Each plot includes a least-squares line to show whether the relationship is positive or negative. From these plots it can be seen that there is no strong trend between any one of the variables and log-area.

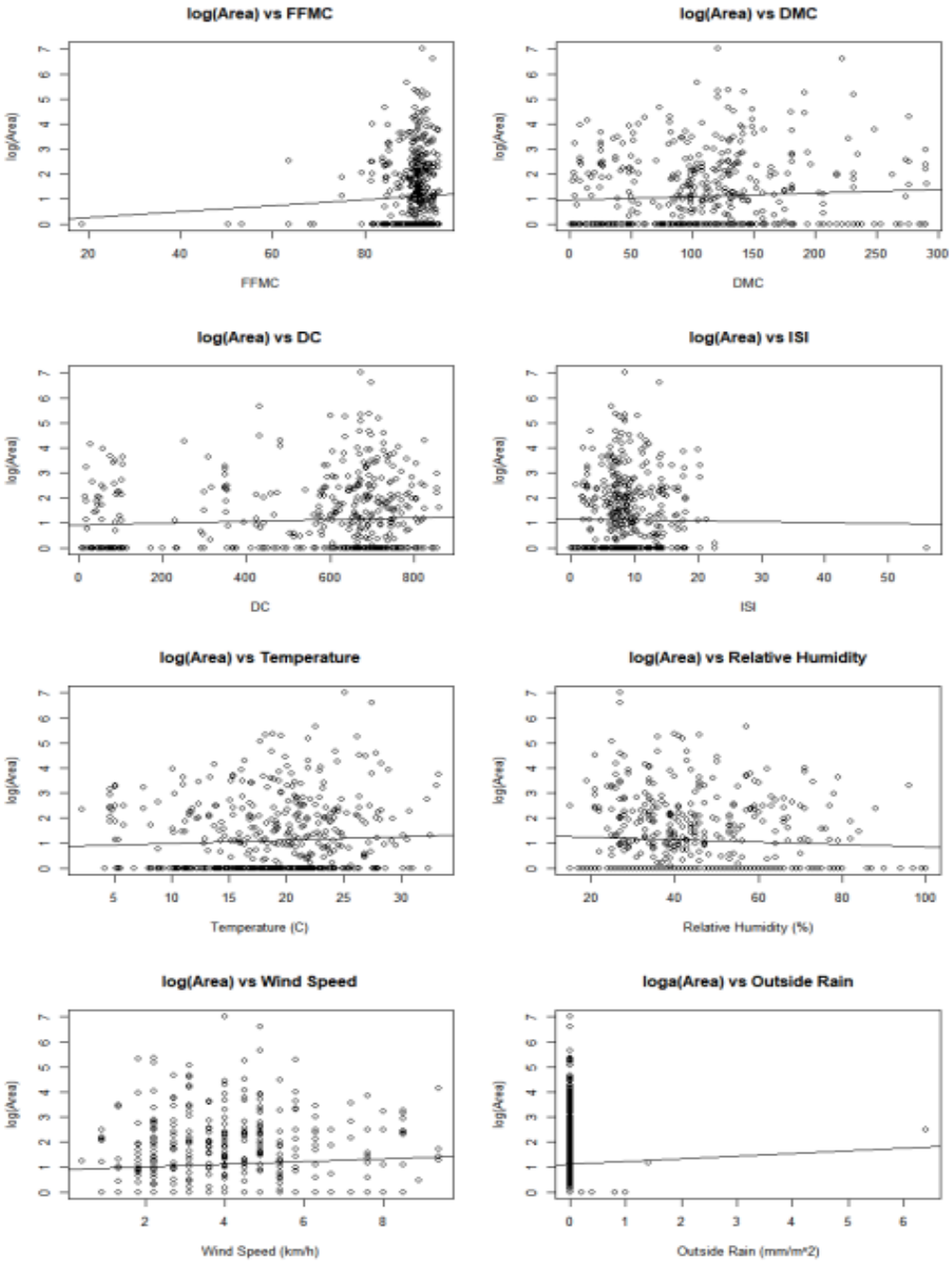


Figure 2. Correlation between area (log) and predictors, see Table 1.

Concerning pair-wise correlations between predictors (data not shown), Temp was positively correlated with several other variables and Temp and RH were negatively correlated. These results suggest the existence of

some multicollinearity within the predictor's dataset (Figure 3). All the variables when plotted by month show some kind of relation with the month. The temp variable shows a pattern of peaks during the summer months. We can also see that the DC variable, which stands for a sort of 'drought code' and is intended as a measure of how dry conditions are, is high during the late summer months.

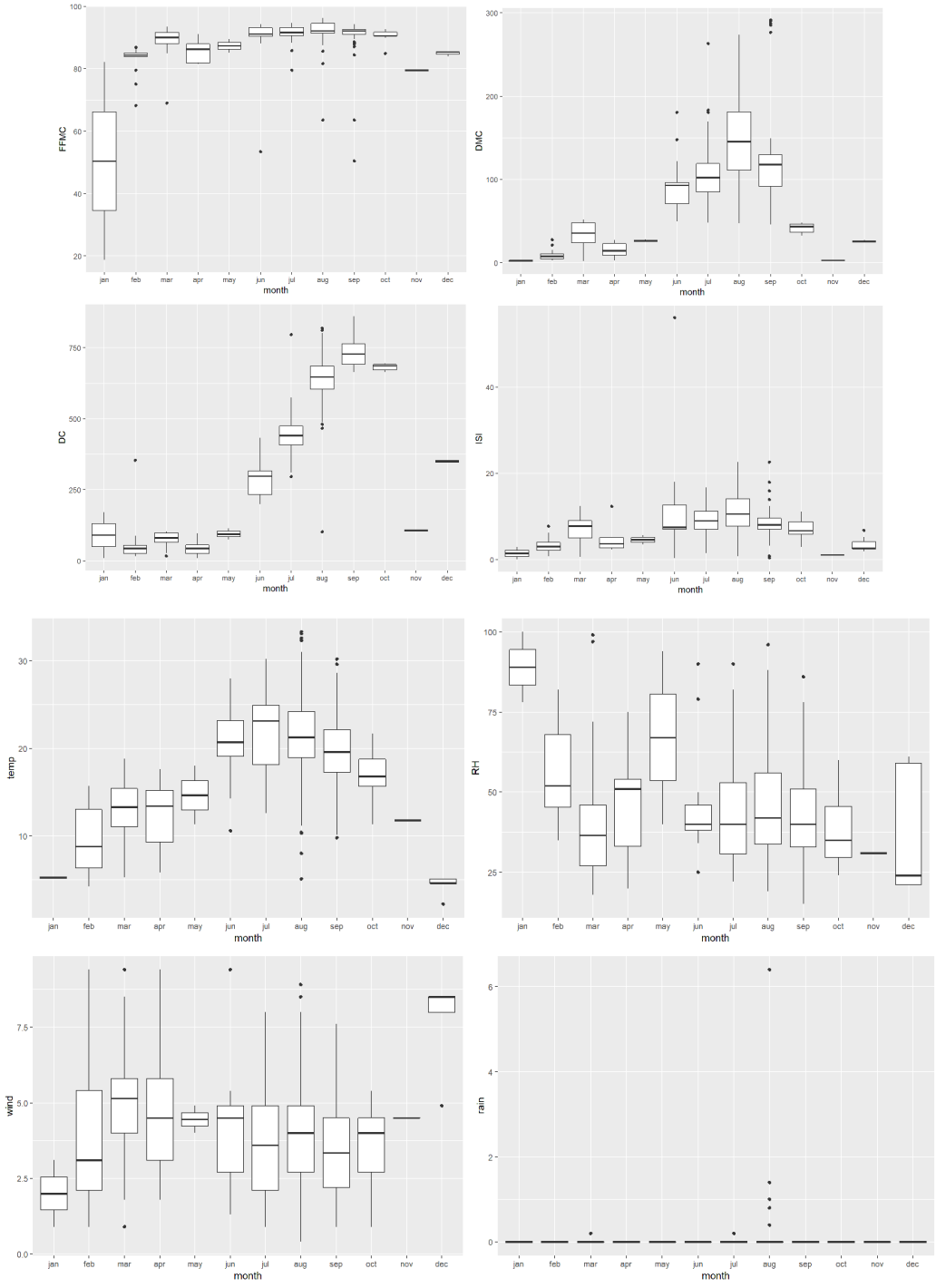


Figure 3. Scatterplots of predictors by month of the year, see Table 1.

By looking at the solid black lines in the centers of the box plots that medians for each variable seem to be quite consistent across days of the week (Figure 4). The size of the boxes is also consistent across days, suggesting that the ranges of data values are similar. The number of outlier points and the length of the box whiskers representing high and low points vary from day to day. However, there do not seem to be any patterns that suggest that the variables differ by day of the week, despite the fact that the number of forest fires appears to be higher on weekends. Though week-day plots did not reveal anything of

particularly interesting, month-wise box plots got some information which looks quite relevant and the variables which revealed such differences might be capable of explaining why there are such differences in a further investigation (Figure 4). Based on these preliminary findings, we performed a linear regression to determine the best predictor variables relevant to predict forest fires'area for each time of the year.

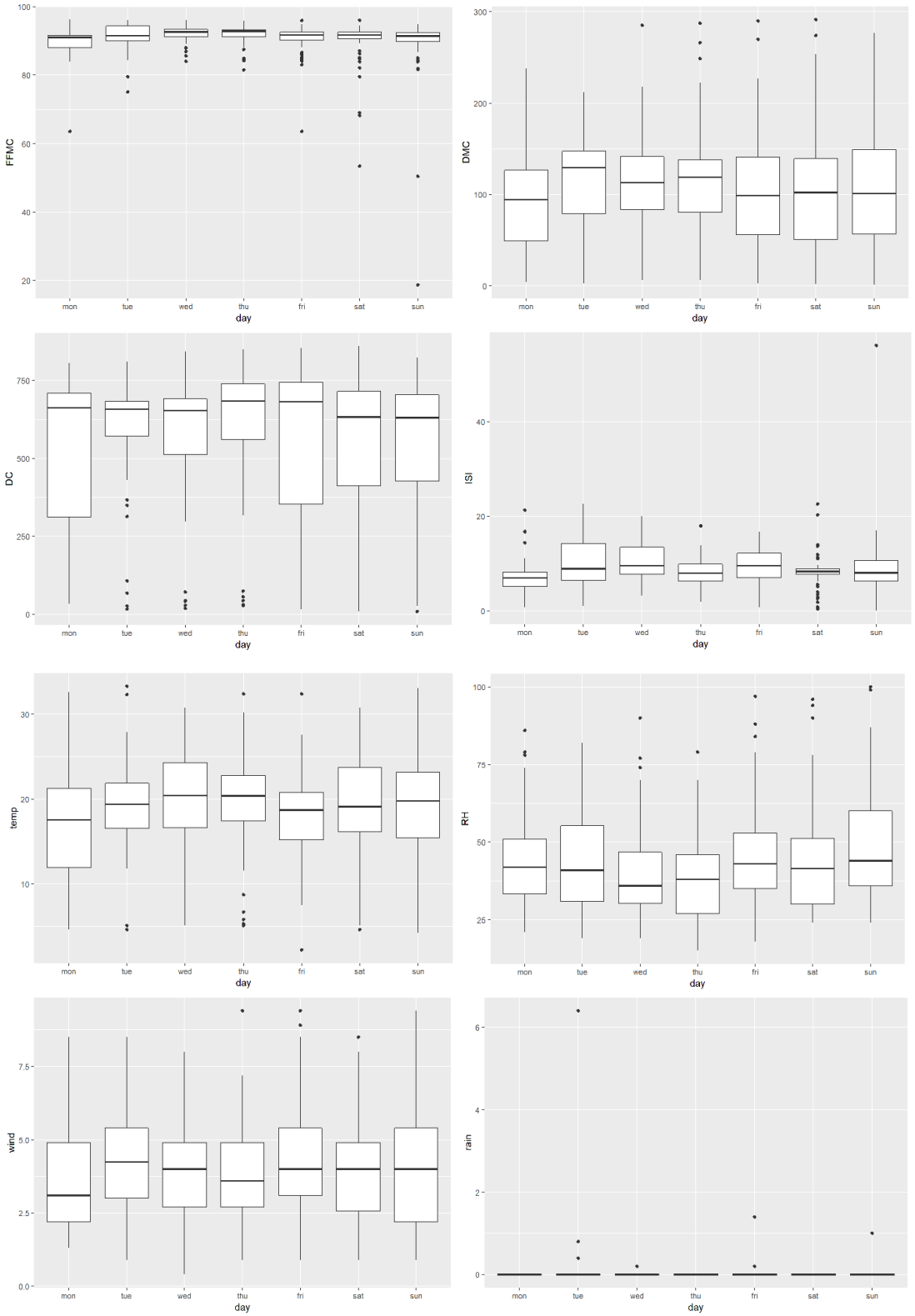


Figure 4. Scatterplots of predictors by day of the week, see Table 1.

The distribution of the dependent variable (area) is strongly asymmetric. However, if we consider the log-transformation of this variable (by addying a fixed coefficient of 1 to all observations), we get a much better fit into a normal distribution (Figure 5); this allows the use of more traditional regression fitting models.

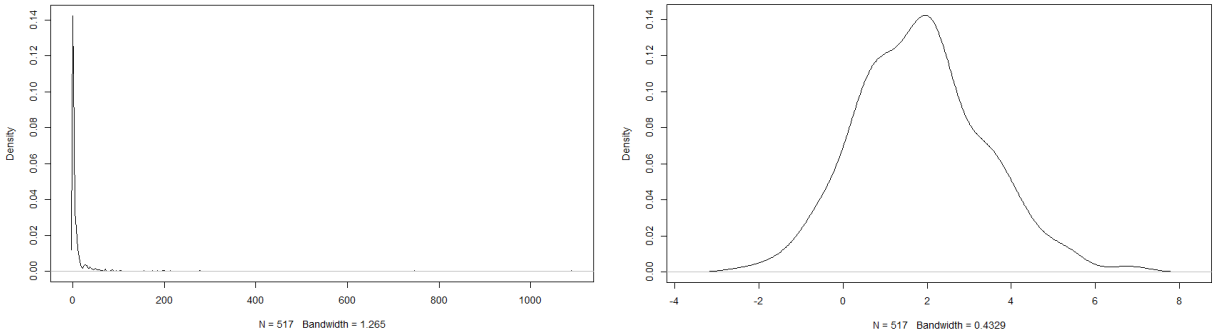


Figure 5. The statistical distribution of the dependent variable in this study (burnt area); left: before logarithmic transformation; right: after logarithmic transformation.

Model specification

To understand what factors affect the extent of burnt area, descriptive statistics alone are not sufficient. Finding the relationships between the burned area and climatic factors, the areas in which fires occur most frequently and the periods in which fires occur most frequently, also means finding a way to use these measurements in real time to classify fire risk conditions and to predict which territorial damage (burnt area) will occur based on the measured variables. To pursue this result, linear regressions were used, considering 517 observations with 29 predictors and a target variable (burnt area), denoted with Y, based on the following specification:

$$y_{\text{score},i} = \alpha + \beta_1 x_{\text{hs},i} + \beta_2 x_{\text{IQ},i} + \beta_3 x_{\text{work},i} + \beta_4 x_{\text{age},i} + \epsilon_i, \quad i = 1, \cdots, n.$$

Following the above generalization, we used the model below:

$$\textit{Area} = \theta_0 + \theta_1\,(\textit{FFMC}) + \theta_2(\textit{DMC}) + \theta_3\,(\textit{DC}) + \theta_4(\textit{ISI}) + \theta_5(\textit{RH})$$
$$+ \theta_6(\textit{temp}) + \theta_7(\textit{wind}) + \theta_8(\textit{rain}) + \theta_{9+i}\,(\textit{day}(i=0\text{-}6)) + \theta_{16+j}(\textit{month}(j=0-11))$$

where $\theta_0, \ldots, \theta_{16+j}$ are predictors of burnt area. Considering days 1 to 6 instead of 1 to 7 (and months from 1 to 11, instead of 1 to 12) allows avoiding unwanted collinearity between predictors. We therefore considered 6 and 11 dummies, respectively classifying each fire by day and month of the year. Dummies were basically binary variables with 0 and 1 indicating absence or presence of the related phenomenon. Based on such a specification, the $\hat{\beta}$ vector was estimated by minimizing the sum of least squares as follows:

$$\hat{\beta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}$$

where X is our (517 x (p + 1)) matrix and p (29 or a subset) is the number of predictor variables and the first collumn of X is a column of 1s. Based on R software, we used the lm() function to estimate our $\hat{\beta}$. Goodness of fit was measured using

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)}.$$

This was used to determine how close the values of $\hat{Y}$ are to the observed values Y. Our fitted values $\hat{Y}$ is

$$\hat{Y} = X\beta$$

Our approach was based on classical linear regressions (LM) implementing a frequentist approach and a refined, Bayesian Linear Regression following BAS library on R software. We used LM and some graphical functions to understand that further pre-processing can be done on the data to improve the goodness-of-fit of the model itself (namely, R^2 and Mean Square Error). We have thus used two datasets, one for training and a separate one for test. We analyze the burned area regression with respect to the selected predictors. Below we highlight only the case of 'temp' regressor, since the results are similar for most of the other regressors (Figure 6). The general formula of multiple linear regression is as follows:

$$y_{\text{score},i} = \alpha + \beta_1 x_{\text{hs},i} + \beta_2 x_{\text{IQ},i} + \beta_3 x_{\text{work},i} + \beta_4 x_{\text{age},i} + \epsilon_i, \quad i = 1, \cdots, n.$$

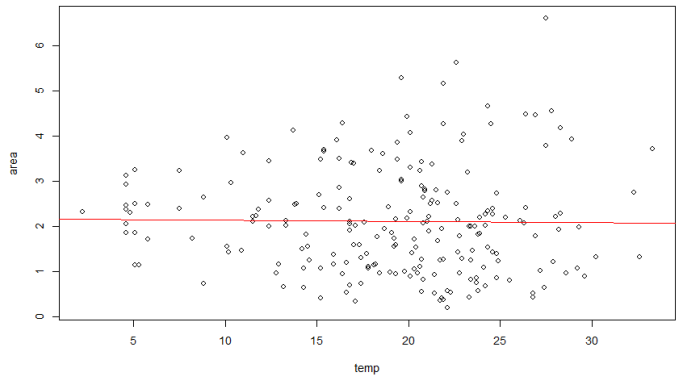


Figure 6. The pair-wise relationship between temperature and area.

Goodness-of-fit of this model is very poor (R^2 = 0.0002; MSE = 1.3756). The qqplot of the residuals is shown as follows (Figure 7).

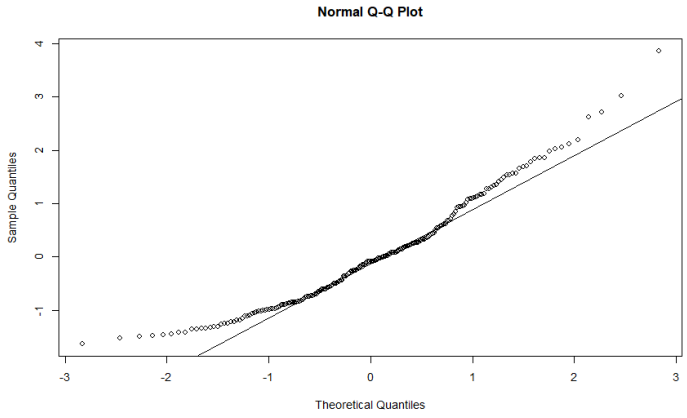


Figure 7. Normal Q-Q plot evaluating the results of the linear regression, see Figure 6.

Help comes from the analysis of the residuals plot of each regressor which highlights an anomaly (Figure 8). We do not know how the first 100 data with area <> 0 were imputed, but they certainly show anomalies. We then decide to delete the first 100 observations from the original set and recalculate the data sets for training and testing as follows. By recalculating the linear regression of burnt area against all the regressors (with the procedure set above) we obtained an improved R^2 (0.165) and MSE (1.20). The qqplot of the residuals was also improved substantially (Figure 9).

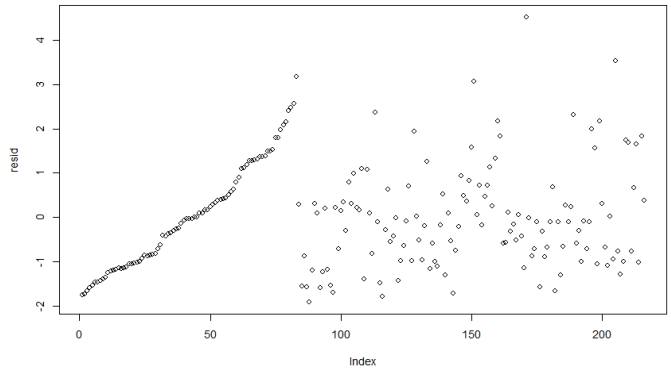


Figure 8. Anomaly plot referring to the observations in Figures 6 and 7.

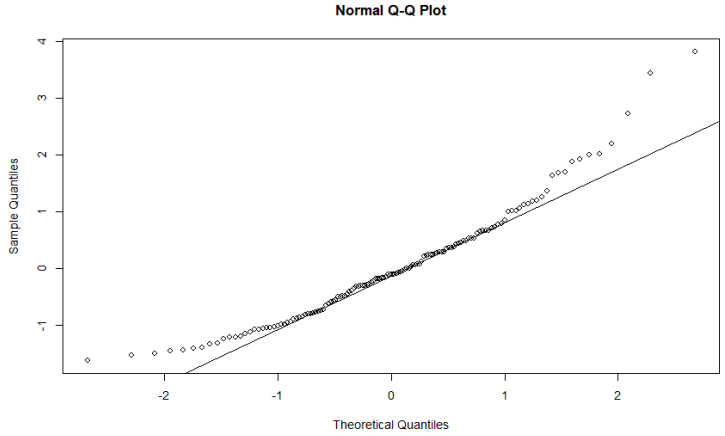


Figure 9. Normal Q-Q plot evaluating the results of the linear regression, see Figures 6 and 7, after changes suggested by results from Figure 8.
With this perspective in mind, the scatter plot also looks better and smoother (Figure 10).

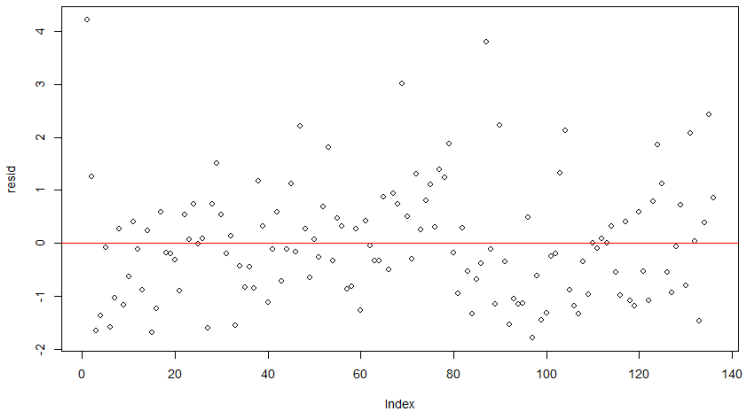


Figure 10. Raw plot of Figure 9 regression.

We illustrated the results of the prediction in Table 2.

```
Call:
lm(formula = area ~ ., data = fires_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0992 -0.7565 -0.0517  0.5777  3.7984

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.727597   6.448075   0.113  0.91036
X            0.014615   0.058465   0.250  0.80307
Y           -0.048815   0.121125  -0.403  0.68771
month2      -2.781154   1.237085  -2.248  0.02654 *
month3      -3.131986   1.263917  -2.478  0.01472 *
month4      -2.223421   1.434276  -1.550  0.12394
month5      -1.840711   1.715658  -1.073  0.28565
month6      -3.412957   1.230167  -2.774  0.00649 **
month7      -3.208261   1.280133  -2.506  0.01365 *
month8      -2.379459   1.353456  -1.758  0.08149 .
month9      -1.157378   1.463221  -0.791  0.43064
day2         0.445668   0.475067   0.938  0.35022
day3         0.211755   0.480744   0.440  0.66045
day4         0.473189   0.442139   1.070  0.28684
```

Table 2. Results of the regression analysis (see Figures 8 and 9).

All the pre-processing we implemented on the data based on the motivations we discussed above, led to improvements that we assessed in terms of R^2 and MSE and which we summarize in Table 3.

	raw data loaded	Area <- log(Area+1)	cutted cols with Area=0	nuemric conversion od month & day	using treduced training setonly	cutted first 100 samples
R2	0.0457	0.0742	0.1291	0.1291	0.1516	0,2112
MSE	3859	1,8069	1,3718	1,3718	1,3559	0.99

Table 3. Improvements in specific diagnostics of the models adopted in the present study.

Other data processing was tried but produced no significant improvements (data not shown). As a general comment, we would state how R^2 and MSE obtained from linear regressions were an initial representation of data. Based on such results, LM seems to do not sufficiently explain what was going on with real-time data. These results were expected based upon the visual scrutiny of the scatter plots. Therefore, we tried to use more complex models (variable selection) in order to draw conclusions about the data. However, these models were not likely to provide a robust solution given the uncorrelated nature of the data. More specifically, for Bayesian inference, we specified a prior distribution for the error term ϵ_i. We assume that ϵi is independent, and identically distributed with the Normal distribution as follows:

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$$

where σ^2 is the commonly shared variance of all observations. The ex-ante distribution should be specified for all the coefficients β_j. We therefore assumed that β' coefficients follow the multivariate normal distribution with covariance matrix $\sigma^2\Sigma_0$. We further imposed the inverse Gamma distribution to σ^2, to complete the hierachical model's specification as follows:

$$\begin{aligned} \beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \mid \sigma^2 &\sim \text{Normal}((b_0, b_1, b_2, b_3, b_4)^T, \sigma^2\Sigma_0) \\ 1/\sigma^2 &\sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

This gives us the multivariate Normal-Gamma conjugate family, with hyperparameters bj, Σ_0, ν_0, and σ^2. For this ex-ante distribution, we specified the values of all the hyperparameters, as reported in Appendix A. This elicitation can be quite difficult, especially when we do not have enough prior information on the variances, covariances of the coefficients, and other ex-ante hyperparameters. Therefore, we are going to adopt the non-informative reference prior, which is a limiting case of this multivariate Normal-Gamma ex-ante assumption. The reference ex-ante in the multiple linear regression model is similar to the reference we used in the simple linear regression model. The ex-ante distribution of all the coefficients β' conditioning on σ^2 is the uniform one, and the ex-ante of σ^2 is proportional to its reciprocal:

$$p(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \mid \sigma^2) \propto 1, \qquad p(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Under this reference, the marginal posterior distributions of the coefficients β' are parallel to the ones in simple linear regression. To gain more flexibility in choosing ex-ante distributions, we used bas.lm function in R software environment, allowing to specify different models. We use the BAS package to fit a Bayesian Linear Regression. BAS is mainly designed to fit a hierarchy of nested models (including and excluding the various covariates) in order to perform model selection and model averaging. Here we use BAS just to fit one single (given) model. For the LM-BAS we used two different assumptions. We first fit the model using a g-prior (Zellner) with constant = 20. This is done by specifying the string "modelprior = Bernoulli(1), include.always = ~ ., n.models = 1". Model prior is the ex-ante over the various models; in this case, using a Ber(1) distribution, we included all the covariates with probability 1. This corresponds to a prior which gives mass 1 to the complete model. The command n.models =1 specifies that we take into consideration only the first model (in this case the full model, due to the form of the prior). Finally "include.always = ~ ." command specifies the fact that in the model we include the intercept β_0. Omitting the prior description on β_0, recall that the (centered) Zellner prior is

$$\begin{aligned} \beta &= (\beta_1, \ldots, \beta_k) \\ \beta|\sigma^2, X &\sim \mathcal{N}_k(0, \alpha\sigma^2(X^tX)^{-1}) \\ \sigma^2|X &\sim \pi(\sigma^2) = \sigma^{-2} \qquad \alpha > 0. \end{aligned}$$

We now choose a suitable mxitrue of Zellner prior, more specifically the Zellner-Siow prior. A ZS prior is simply a mixture of Zellner prior, i.e.

$$\begin{aligned} \beta &= (\beta_1, \ldots, \beta_k) \\ \beta|\sigma^2, X &\sim \mathcal{N}_k(0, \alpha\sigma^2(X^tX)^{-1}) \\ \sigma^2|X &\sim \pi(\sigma^2) = \sigma^{-2} \\ 1/\alpha &\sim \pi_0 = Gamma(1/2, n/2) \end{aligned}$$

By choosing other forms for π_0, one obtains other Zellner type priors, some of them are implemented in BAS. The above bas.lm function uses the same model formula as in the lm. It first specifies the response and predictor variables, a data argument to provide the data frame. We continued using the same datasets as above, one for training and another one for test. Since we will only provide one model, which is the one that includes all variables, we place all model prior probability to this exact model. This is specified in the modelprior = Bernoulli (1) argument. Because we want to fit using all variables, we use 'include always = ~ .' to indicate that the intercept and all predictors are included. The argument n.models = 1 fits just this one model. Regarding Posterior Means and Posterior Standard Deviations, we note that LM-BAS is similar to the OLS regression process; we can extract the posterior means and standard deviations of the coefficients using the coeffunction. From the last column in this summary, we see that the probability of the coefficients to be non-zero is always 1. This is because we specify the argument include.always = ~ . to force the model to include all variables. Notice that, on the first row ,we have the statistics of the intercept β_0. The posterior mean of β_0. is 1.98, which is completely different from the original y-intercept of this model under the frequentist OLS regression. As we have stated previously, we consider the "centered" model under the Bayesian framework. Under this "centered" model and the reference prior, the posterior mean of the Intercept β0 is now the sample mean of the response variable Y_{score}. We can visualize the coefficients βj, using the plot function. We use the subset argument to plot only the coefficients of the predictors (Figure 11).

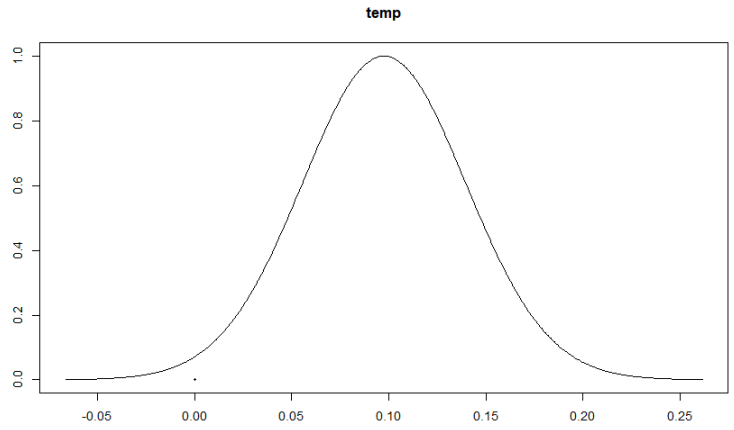


Figure 11. Ex-ante distribution of a sample predictor.

The obtained distributions all center the posterior distributions at their respective OLS estimates μβj, with the spread of the distribution related to the standard errors $se_{\beta j}$. Recall, that bas.lm uses centered predictors so that the intercept is always the sample mean. We also report the posterior means, posterior standard deviations, and the 95% credible intervals of the coefficients of all predictors, which may give a clearer summary. The BAS library provides the section confint to extract the credible intervals from the output cog.coef. Since we are only interested in the distributions of the coefficients of the predictors, we may use the parm argument to restrict the variables shown in the summary. All together, we can generate a summary table showing the posterior means, posterior standard deviations, the upper and lower bounds of the 95% credible intervals of all coefficients β_j. As in the simple linear regression, the posterior estimates from the reference prior, that are in the table, are equivalent to the numbers reported from the lm function in R, or using the confident function in the OLS estimates. These intervals are centered at the posterior mean μβj with width given by the appropriate t quantile with n−p−1 degrees of freedom times the posterior standard deviation $se_{\beta j}$. The primary difference is the interpretation of the intervals. Except for some values such as month 2, there is generally an improvement in values with the LM-BAS model. Maybe we could improve this model so that the model will accomplish a desired level of explanation or prediction with fewer predictors. We can also report the posterior means, posterior standard deviations, and the 95% credible intervals of the coefficients of all predictors, which may give a clearer and more useful summary. The BAS library provides the method confint to extract the credible intervals from the output cog.coef. If we are only interested in the distributions of the coefficients of predictors, we may use the parm argument to restrict the variables shown in the summary (Table 4).

	2.5%	97.5%	beta
Intercept	1.926731477	2.3404160932	2.133573785
X	-0.099948732	0.1284658808	0.014258575
Y	-0.284232716	0.1889833475	-0.047624684
month2	-5.129873894	-0.2967672304	-2.713320562
month3	-5.524563926	-0.5866286560	-3.055596291
month4	-4.970942818	0.6325604591	-2.169191180
month5	-5.147223840	1.5555934713	-1.795815184
month6	-5.732752454	-0.9266756818	-3.329714068
month7	-5.630654812	-0.6293658372	-3.130010325
month8	-4.965298741	0.3224518086	-2.321423466
month9	-3.987442979	1.7291442360	-1.129149371
day2	-0.493209035	1.3628060589	0.434798512
day3	-0.732507129	1.1456882930	0.206590582
day4	-0.402038122	1.3253329070	0.461647393
day5	-0.898533620	0.7627963510	-0.067868634
day6	-0.498668122	1.2756685264	0.388500202
day7	-0.924092594	0.8133948042	-0.055348895
FFMC	-0.110188339	0.1640947024	0.026953182
DMC	0.001351548	0.0137095892	0.007530569
DC	-0.010010479	0.0007747079	-0.004617886
ISI	-0.162919476	0.0239054904	-0.069506993
temp	0.038605531	0.2108198010	0.124712666
RH	-0.013255544	0.0375928045	0.012168630

Table 4. Results of Bayesian regression.

BAS details automatically split named month and day (11 and 6 respectively), too. The prior coeff distributions for two different priors considered, are showed in Figure 12.

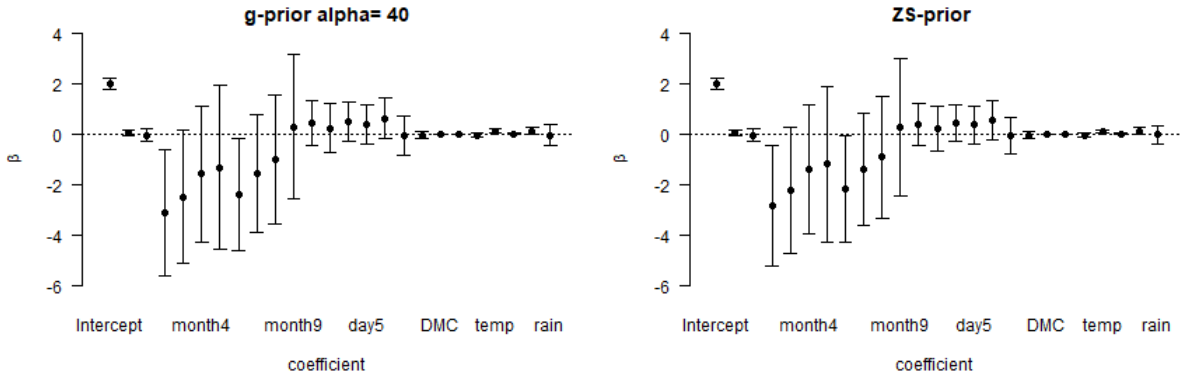


Figure 12. Analysis of regression details, see Table 4.

True values vs fitted and predicted values were illustrated in Figure 13. Black points are true vs fitted in the learning dataset and red points are true vs predicted in the additional part of the data set (not used for fitting the model – test dataset). For the prediction we used test dataset: fires_test (or newdata). Results (Table 5) are plotted in the following figure.

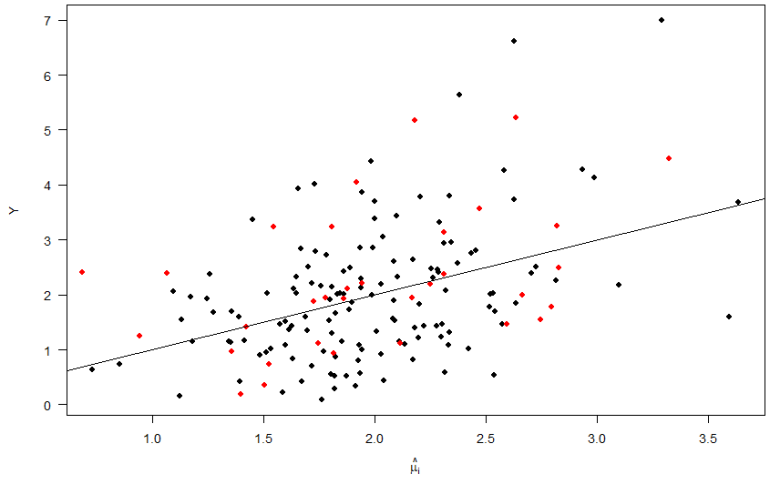


Figure 13. Estimated vs true values in the regression analysis in Figure 12.

	post mean	post SD	post p(B != 0
Intercept	1.984129	0.103161	1.000000
X	0.025312	0.057669	1.000000
Y	-0.044880	0.121115	1.000000
month2	-3.123828	1.270191	1.000000
month3	-2.479538	1.321465	1.000000
month4	-1.568462	1.360689	1.000000
month5	-1.330510	1.643007	1.000000
month6	-2.411068	1.122268	1.000000
month7	-1.573887	1.179804	1.000000
month8	-1.014535	1.283764	1.000000
month9	0.291172	1.442415	1.000000
day2	0.430756	0.440258	1.000000
day3	0.231084	0.484032	1.000000
day4	0.503818	0.386803	1.000000
day5	0.395916	0.394647	1.000000
day6	0.600018	0.406374	1.000000

Table 5. Marginal posterior summaries of coefficients.

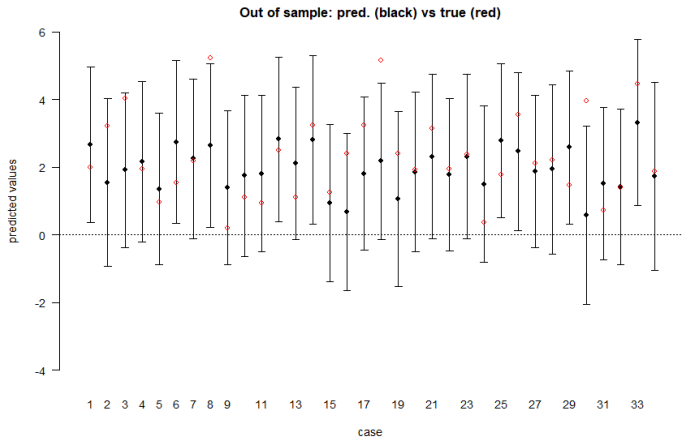
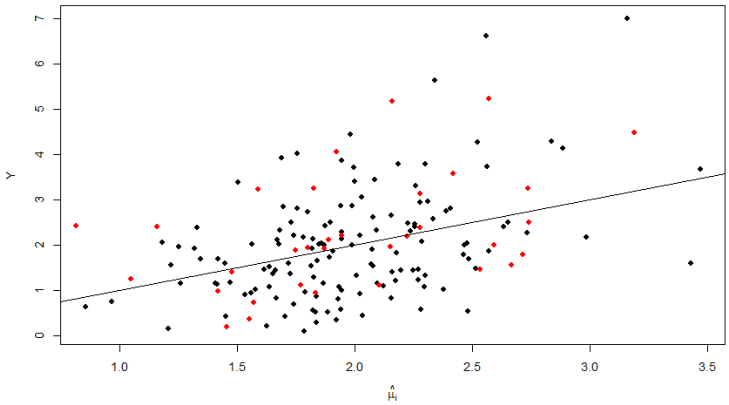


Figure 14. Out of sample plot.

We repeated the same elaboration using ZS-prior, looking for a better posterior and prediction (Figure 15).



1

```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  1 models
           post mean  post SD    post p(B != 0)
Intercept   1.984129   0.103161   1.000000
X           0.022781   0.054710   1.000000
Y          -0.040392   0.114900   1.000000
month2     -2.811431   1.205006   1.000000
month3     -2.231573   1.253648   1.000000
month4     -1.411609   1.290860   1.000000
month5     -1.197453   1.558689   1.000000
month6     -2.169950   1.064674   1.000000
month7     -1.416491   1.119258   1.000000
month8     -0.913077   1.217883   1.000000
month9      0.262053   1.368391   1.000000
day2        0.387679   0.417665   1.000000
day3        0.207974   0.459192   1.000000
day4        0.453434   0.366952   1.000000
day5        0.356323   0.374394   1.000000
day6        0.540014   0.385519   1.000000
day7       -0.045688   0.366766   1.000000
```

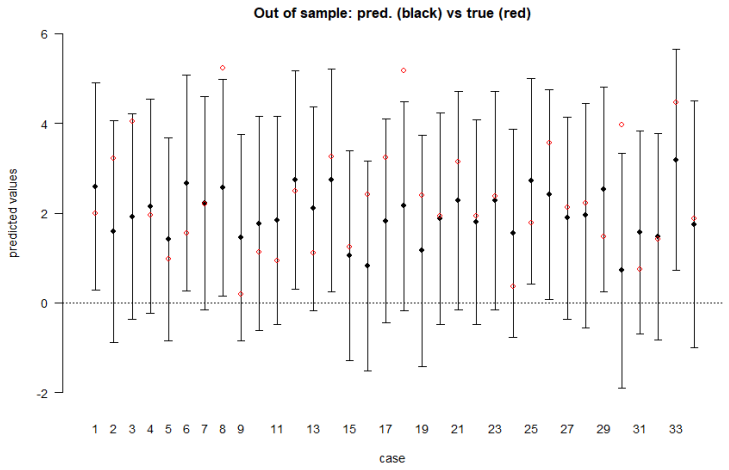


Figure 15. Examples of regression findings using ZS-prior.

Based on these results, we can say that unsignificant improvements were achieved. The prediction is rather approximate. There is no noticeable difference between the predictors of the LM-BAS multiple linear regression model based on G-prior or ZS-prior. The models were not likely to provide a robust solution given the uncorrelated nature of the data.

Concluding remarks

Because of a strong uncorrelated nature of the data, models explored with this study were not likely to provide a robust solution. The data presents many uncertainties. We have provided analyses for the multiple linear regression using the default reference prior. We have seen that, under this reference prior, the marginal posterior distribution of the coefficients is the Student's t-distribution. Therefore, the posterior mean and posterior standard deviation of any coefficients are numerically equivalent to the corresponding frequentist OLS estimate and the standard error. This has provided us a base line analysis following a Bayesian approach, which we can extend later with LASSO and Hierarchical Models introducing more different coefficient priors. The difference lies in the interpretation. Since we have obtained the distribution of each coefficient, we can construct the credible interval, which provides us the probability that a specific coefficient falls into this interval.

APPENDIX

Our study was based on different R scripts (Appendix). In the following table, we show the corrispondence between trated topics and related Script R used for the application.

Topics	Script
Data Overview and Exploration	DATA_INTERPRETATION.R GRAPHICS.R ANALYSIS_WITH_LM_AND_LM-BAS.R
Models LM and LM-BAS	ANALYSIS_WITH_LM_AND_LM-BAS.R
Models Hierarchical prior for β_j and LASSO	ANALYSIS_WITH_HIERARCHICAL_AND_LASSO V2.R
Dataset	forestfire.txt

Model 2 Description: Hierarchical prior for β_j

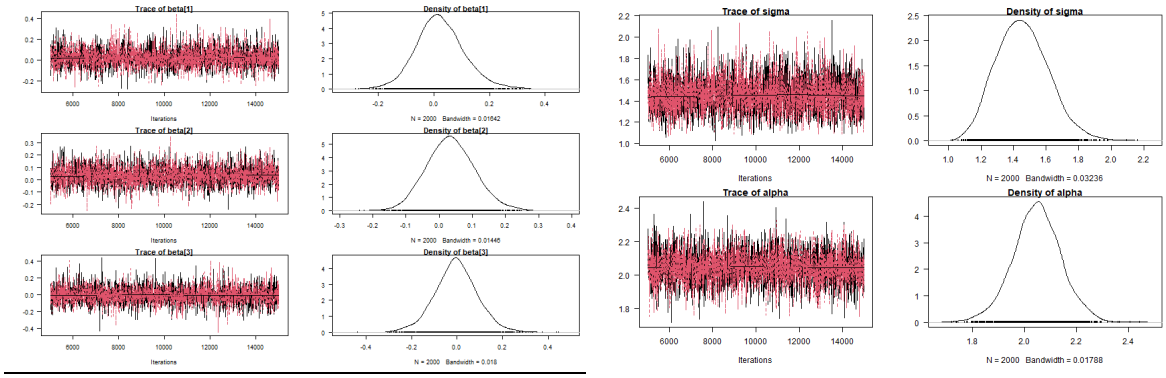
Hierarchical models try to formalize the idea that random variation operates at different levels and a statistical model should account for all of them. We usedthe following **hierarchial** model

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\alpha + X_i\beta, \sigma^2) \\
\alpha &\sim \mathcal{N}(0, 100) \\
\beta_j &\sim \mathcal{N}(0, \sigma_b^2) \\
\sigma^{-2} &\sim \mathcal{G}(0.01, 0.01) \\
\sigma_b^{-2} &\sim \mathcal{G}(0.01, 0.01)
\end{aligned}
$$

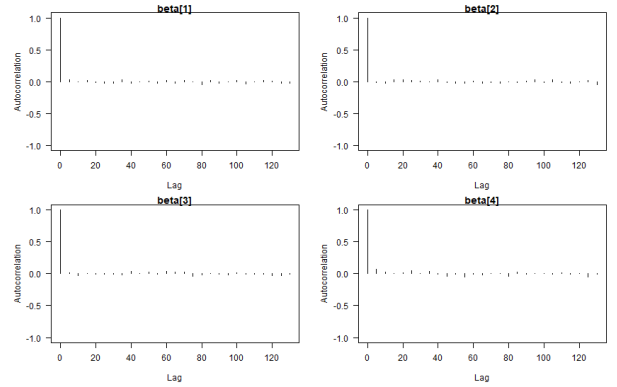
We use 10 regressors, because with this model we can manager directly all the data and are not provieded automatic transformations of the model like happened with LM and LM-BAS Models. In R-JAGS, we can config prior and likelhood, and use this models to obtain posterior evaluation. We can set also Hyperparameters. We choose little values to permit an optima convergence of method;
Results of following distributions a used like input parameters for Prior and Likelihood.

- inv.var ~ dgamma(0.01, 0.01)
- inv.var.b ~ dgamma(0.01, 0.01)
- alpha ~ dnorm(0.01, 0.01)

For dgamma we choose shape=0.01. Our parameters for dnorm (0, 0.001) permit to obtain the same as a Normal distribution with mean 0 and variance 1/0.001 = 1000



Model has a fastly convergence.




```
Iterations = 5005:15000
Thinning interval = 5
Number of chains = 2
Sample size per chain = 2000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean      SD Naive SE Time-series SE
alpha     2.0456735 0.09218 0.001457       0.001488
beta[1]   0.0242849 0.09119 0.001442       0.001597
beta[2]   0.0344919 0.07558 0.001195       0.001220
beta[3]  -0.0007547 0.09428 0.001491       0.001473
beta[4]   0.0889303 0.09818 0.001552       0.001705
beta[5]  -0.1026929 0.11722 0.001853       0.002268
beta[6]  -0.1179273 0.10125 0.001601       0.001883
beta[7]   0.0759264 0.10097 0.001597       0.001722
beta[8]  -0.0795347 0.08232 0.001302       0.001330
beta[9]   0.1063142 0.08423 0.001332       0.001452
beta[10]  0.0137221 0.07312 0.001156       0.001156
sigma     1.4554744 0.16373 0.002589       0.002589
```

Model Description: LASSO

The lasso is a shrinkage method so defined as:

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\alpha + X_i\beta, \sigma^2) \\
\alpha &\sim \mathcal{N}(0, 100) \\
\beta_i &\sim \mathcal{DE}(0, \sigma_b^2\sigma^2) \\
\sigma^2 &\sim \mathcal{IG}(0.01, 0.01) \\
\sigma_b^2 &\sim \mathcal{IG}(0.01, 0.01)
\end{aligned}
$$

that can be used to determine the best variables for predicting area burned. For thsi model is valid all we said and did from previous model.

Results of following distributions a used like input parameters for Prior and Likelihood. For Dgamma we choose shape=0.01. dnorm (0, 0.01) is the same as a Normal distribution with mean 0 and variance 1/0.01 = 100

- inv.var ~ dgamma(0.01, 0.01)
- inv.var.b ~ dgamma(0.01, 0.01)
- alpha ~ dnorm(0.01, 0.01)

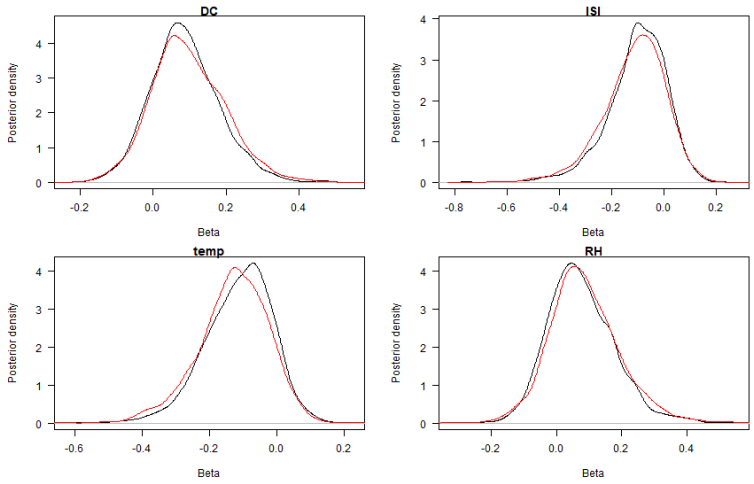
For Dgamma we choose shape=0.01. dnorm (0, 0.01) is the same as a Normal distribution with mean 0 and variance 1/0.01 = 100

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean      SD Naive SE Time-series SE
alpha     2.0466014 0.09139 0.001292       0.001292
beta[1]   0.0250136 0.09426 0.001333       0.002139
beta[2]   0.0401736 0.07637 0.001080       0.001151
beta[3]  -0.0003538 0.10022 0.001417       0.002184
beta[4]   0.0949567 0.09658 0.001366       0.002162
beta[5]  -0.1097916 0.11403 0.001613       0.003260
beta[6]  -0.1270874 0.10376 0.001467       0.002497
beta[7]   0.0805809 0.10372 0.001467       0.002539
beta[8]  -0.0856338 0.08551 0.001209       0.001522
beta[9]   0.1139760 0.08582 0.001214       0.001624
beta[10]  0.0168694 0.07761 0.001098       0.001156
```

Comparision of Models Hierarchical prior for βj and LASSO

The results obtained from the summary on the samples obtained with the two methods Hierarchical and LASSo are equivalent. Both models make a similar estimation of Regressors, even if they use different Likelihood



Unfortunately, this regressor estimation does not allow to obtain good results in the reconstruction of the real data. Also these models were not likely to provide a robust solution given the uncorrelated nature of the data.