*Methodology of econometric and statistical studies. AMSE – 18/01/2021*

# Fire management assistance

Predicting whether a fire will take on disastrous proportions.

Eva MANUKYAN, Elisa SEBASTIAN

# Table of contents

# Executive summary

Fires have harmful consequences for the planet on fauna, flora, housing areas and health.

As an example, mid-august, giant fires ravaged the western United States. Approximately two million hectares were burned, a disaster since it represents the size of Slovenia.

The issue is that those fires can be difficult to manage, especially during the summer, and when the resources are limited.

The purpose of this study is therefore to assist American authorities in charge of **fire management** and **firefighters** in their **decision making**. *When more than one fire breaks out, which one is likely to become a mega fire? How many firefighters to send? How many resources should be allocated to each?* The methods we are going to present can of course be extended to other countries, provided that sufficient data on geography, vegetation, and weather, among other things, are available.

In this context, **machine learning methods** (*logistic regression, K-nearest neighbors, Random Forest, XG boost*) will be used in order to **predict** whether a fire that has just started has a chance of becoming a ***"mega fire"*** (*greater than 1000 hectares according to the authorities*) or not.

Of course, we will also explain the factors that influence the occurrence of these mega-fires through *descriptive statistics* and a *literary review*.

To do so, we used data containing information on **55367 US fires** from **1992 to 2015** such as *weather, location*, and *vegetation*, including **5685 megafires**.

In order that predictions can be implementable in fire stations, the stakes of **big data**, including processing time, will be considered (*there are several thousand fires a year)*.

We concluded that the best model in terms of prediction is **XG-Boost** although it has the disadvantage of not being very interpretable.

Random forest also establishes a good compromise between processing time and accuracy.

We were also able to highlight, thanks to other models such as logistic regression, that meteorological conditions greatly influence the triggering of mega fires.

Given the performance of the models used, these algorithms can therefore be effectively used in fire stations.

However, artificial intelligence has its flaws and therefore cannot replace a real professional diagnosis of fires. Indeed, we must not lose sight of the fact that it is thanks to them that we can know, as an example, the cause of the fire.

# 1 Introduction

Fires, *especially large wildfires*, are a disaster not only from an ecological point of view since they destroy **fauna** and **flora**, but also for individuals because it impacts **health** and **housing areas**.

The latter are often on a large scale, and, according to the authorities, a **mega fire** is more than **1000 hectares**.

Unfortunately, we have been able to see in recent years that they are more and more present. We are thinking of the biggest bush fires of 2019-2020 that **Australia** has ever known, of no less than 10 136 fires for the first ten days of August 2020 in the **Amazon**, but also of giant fires in the West Coast of the **United States** that burned approximately **two million hectares,** which represents the size of Slovenia.

In most of these cases, the distress of the authorities and firefighters is significant. Indeed, these mega-fires are often **difficult to manage** because of limited **resources**. Furthermore, the summer months when **several fires** break out at the same time can be a source of **complex decisions**.

After this summer's repeated fires in the **United States** and the difficulty in putting them out, we asked ourselves how we could help the people in charge of fire management.

The purpose of this study is therefore to assist this country and its firefighters in their **decision making**, by **predicting whether a fire that has just started has a chance of becoming a *"mega fire"*** using **machine learning** algorithms, and of course to give some explanations on the factors that influence their occurrence.

Indeed, imagine that two fires are starting at the same time and there are not many firefighters available, knowing which of these fires is most likely to be of disastrous proportions can help the fire station to know which fire to focus on.

To answer those question, we used data containing information on **US wildfires**[1], historical **weather data** at a specific latitude/longitude from 1992 to 2015[2] , historical **vegetation data**[3] and a metric that represents the measure of the **remoteness** of a fire *(distance to the closest city)*[4].

---

[1] *Short and Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive* https://doi.org/10.2737/RDS-2013-0009.4

[2] *NOAA National Centers for Environmental Information (2001): Integrated Surface Hourly [1992-2015]* ftp://ftp.ncdc.noaa.gov/pub/data/noaa/

[3] *Meiyappan, Prasanth, and Atul K. Jain. "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years." Frontiers of Earth Science 6.2 (2012): 122-139.*

[4] *"World Cities Database." Simplemaps,* simplemaps.com/data/world-cities.

In more details, the data correspond to **55367 fires** in the United States from 1992 to 2015, including **5685 mega-fires.**

In the next section, we will present studies that, like ours, put the **machine learning and AI** at the service of **fire management**, along with **scientific facts** that highlight the factors that influence the magnitude of fires (***Section 2 literature review***).

Then, we will compare these relationships using descriptive statistics (***Section 4 descriptive statistics***) after presenting the database (***Section 3 data description***).

After doing pre-processing work, we use the **logistic regression**, **k-nearest neighbors**, **random forest**, and **XG-boost** to predict mega-fires on an untrained part of our database (***Section 5 Machine learning models***).

In our database, fires, and more mainly megafires are influenced by the **weather** (high temperature, high wind, low humidity and precipitation), their **cause** (e.g. lightning), the **seasons** (spring-summer) and the **distance to the nearest city** (low) as demonstrated in ***section 5.2- logistic regression.***

We conclude (***section 6***) that the best model in terms of classification is **XG-Boost** (*accuracy about 99.29%).*

If we are interested in the processing time, then **Random forest** is the best compromise, since it also obtains a good accuracy (*97.51%).*

It will be up to the authorities in charge of fire management to make its arbitrations.

# 2 Literature Review

This is not the first time that artificial intelligence has been used in fire management.

Although there do not seem to be any techniques in place in fire stations to know whether a fire that breaks out is going to take on disastrous proportions (*as in our study*), there are several techniques to **automatically detect forest fires** thanks to a system of automatic detection and geolocation by **processing images** of fire starts (*flame or smoke*).

For example, in Chile, sensors in trees use AI to better detect forest fires.
Also, the French company *Sogetrel* relies on a detection algorithm whose role is to warn the emergency services as quickly as possible before a fire gets too big.
The **Forest Fire Weather Index** *(FWI)*, developed in Canada in the late 1970s, estimates the meteorological danger of forest fires by considering the probability of their outbreak and their potential to spread.
This index is calculated from simple meteorological data: *temperature, air humidity, wind speed and precipitation*. Several studies have shown a clear correlation between the average FWI and the number of fire start, and this index is used today in most countries.

Nevertheless, as we have not found any studies like ours, we will focus in this literature review on the **factors that influence** the outbreak of a **fire**, and more particularly **mega-fires**.

**Global warming** may have its role to play. The *Global Forest Watch* has identified more than 4.5 million fires larger than one square kilometre worldwide in 2019. This is 400,000 more fires than in 2018 and two and a half times more than in 2001, *"Almost all (96%) of the most disastrous 500 megafires in the past decade have occurred during periods of unusual heat and/or drought."*

**Higher temperatures** encourage plant transpiration and reduce water content in the soil. As vegetation dries out, the risk of fire starts is greater. The amount of fuel available once a fire is started also increases.

**Decreased rainfall** during fire-prone seasons can aggravate the phenomenon.

These mega-fires are linked to hot weather conditions. In the **summer**, the **ground temperature** can rise to 80°C. The slightest badly extinguished cigarette butt can cause fires to start. Therefore, mega-fires are mainly found during the summer months.

**Wind** also contributes to the risk of fire. Very dry winds **lower the humidity** in the air. If it is extremely hot, the combination of these dry winds and heat aggravates the **dryness** of the soil.

During a fire, the wind coats the flames and increases the grip of the fire on the vegetation. The projections of **flaming particles** (*pieces of bark, pinecones, needles, leaves*) that are lifted and carried by the wind can then travel several kilometres from the place of origin of the fire, before igniting other secondary fires.

Remember that for a fire to start, it is essential to bring together three ingredients: a **fuel**, an **oxidizer** and an **activating energy**, *a source of heat*.
For example, in the case of forest fires, **vegetation** takes the place of fuel, the air and the oxygen it contains play the role of oxidizer and the slightest spark can then be enough to provide sufficient activation energy.

Opting for the least flammable plants and positioning them correctly can therefore **help reduce the risk of fire**. To this end, *Anne Ganteaume*, a researcher at the *IRSTEA* in Aix-en-Provence, has been offering a new guide since this spring, offering a set of practical advice (planting, plant care, etc.) for managers of green spaces and private individuals.

A fire can therefore take different forms, each of which is conditioned by the characteristics of the vegetation and the climatic conditions. Thus, we can distinguish (*These three types of fire can occur simultaneously in the same area*):

- **Soil fires** that burn the organic matter contained in the litter; their propagation speed is **low**.
- **Surface fires** that burn the lower strata of the vegetation, i.e., the upper part of the litter. They generally spread by radiation and affect **scrubland** or **heathland**.
- **Top fires** that burn the upper part of **trees** (tall woody trees) and form a crown of fire. They generally release large amounts of energy and their propagation **speed** is extremely high.

The **relief** can also play a role: *the slope conditions the inclination of the flames in relation to the ground and thus their speed of propagation*.

In addition to the characteristics presented above that favour the occurrence and magnitude of fires, *such as high temperatures*, there are places in the world where we do not necessarily imagine that mega-fires will break out. But this is not the first time that wildfires are ravaging parts of the **Arctic**, with areas of ***Siberia***, ***Alaska***, ***Greenland***, and ***Canada***.

For example, in Canada, a wildfire has burned at 45,500 hectares according to the *Northwest Territories Environment and Natural Resources agency*.
*Mr Parrington, Senior scientist Copernicus Atmosphere Monitoring Service*, informs us that *"It is unusual to see fires of this scale and duration at such high latitudes in June",*

*"But temperatures in the Arctic have been increasing at a much faster rate than the global average, and warmer conditions encourage fires to grow and persist once they have been ignited."*

Let us now look at the other characteristics of fires, particularly in the United States.

According to the *U.S. Forest Service Research Data Archive*, **85%** of fires result from campfires left unattended, burning of debris, equipment uses, negligently discarded cigarettes, intentional acts of arson, and are therefore caused by **humans**. Another cause can be lightning.

Another fact is that some states are more affected than others.

From **mid-August 2020**, the **West Coast** faced one of the most dreaded fire seasons in history.

In California, more than **1.2 million hectares have burned**, ten times more than at this time last year. More than **3 000 homes** have been destroyed according to the *California Forest and Fire Protection Agency (Cal Fire).*

*Why are there so many fires this year?* Experts say a combination of factors: **high temperatures, high winds, unusually dry forests and rarely seen thunderstorms**.

Indeed, the remnants of two tropical storms ravaged Northern California, causing nearly 15 000 lightning strikes on dry forests and brush.

We can therefore see that his country is not immune to the weather conditions we have been talking about so far.

We will now establish descriptive statistics on the database we have constructed to compare them with this literature.

# 3 Data description

Our study focuses on the United States.

Indeed, this country has made available **data on fires** on its soil from **1992 to 2015** through the Forest Research Data Archive.

In addition, **NOAA** (*the National Centers for Environmental Information*) made available meteorological variables that can be grouped into the first set of data using longitude and latitude. This centre also provides vegetation data.

Finally, a metric that represents the measure of the **remoteness** of a fire *(distance to the closest city)* have been added thanks to the "*world cities database*" coming from *simplemaps.*

With these combinations, the database includes **55367 fires** of which **5685 are mega fires**.

***We have chosen the following explanatory variables for this study, which are:***

- **Longitude** and **latitude** at the location of the fire
- **Cause** of the fire
- **State**
- **Dominant vegetation** in the area
- The **discovery month**
- **Temperature** in Deg C
- **Precipitations** in mm
- **Humidity** in %

*30, 15 and 7 days before the fire and the current day.*

- **Distance to the closest city** (non-dimensional)

More details on each of these variables can be found in ***section 4 Descriptive statistics***.

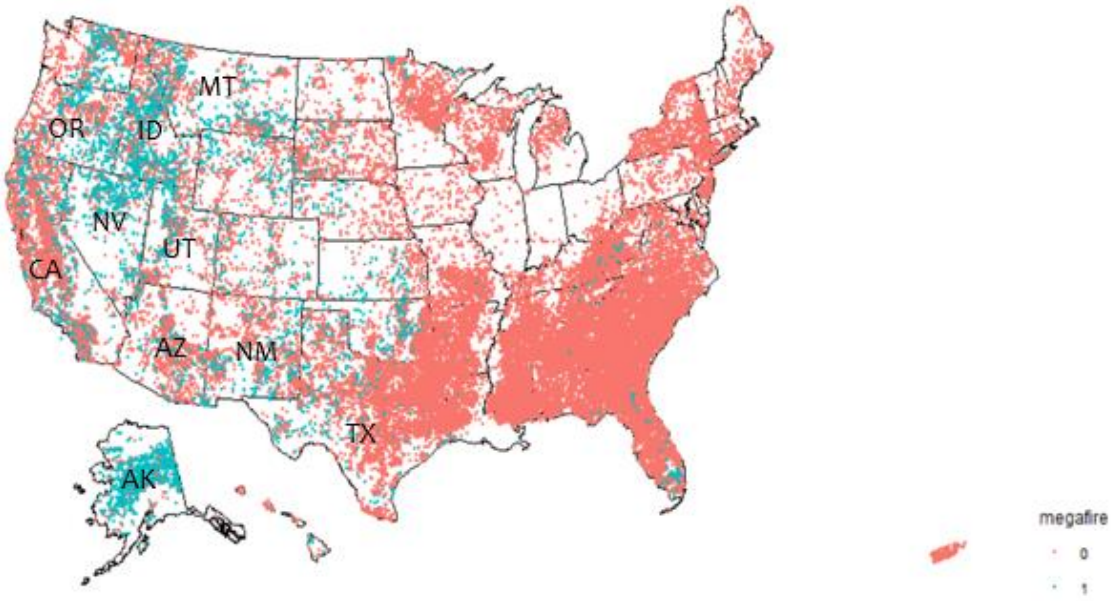
With this selection of variables, we do not have missing data.

# 4 Descriptive statistics

In this section, you will find descriptive statistics about the mega-fires and each explanatory variable we have chosen to keep.

## 4.1 Map of megafires

***Map of mega-fires in the United States for the period 1992-2015:***



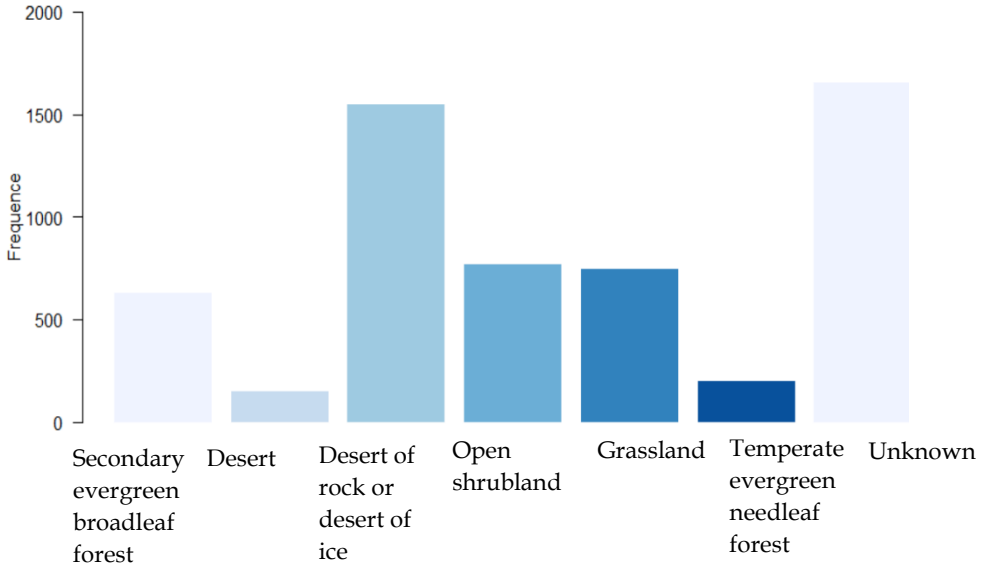

It is well known that megafires in the United States *(in green)* often occur on the West Coast because of its vegetation and weather.

For example, you will find below the 10 states with the most megafires for the period 1992-2015 in descending order.

- **Alaska (AK)** : 823
- **California (CA):** 590
- **Idaho (ID)** : 583
- **Nevada (NV)** : 408
- **Texas (TX)** : 389
- **New-Mexico(NM)** : 365
- **Oregon (OR)** : 361
- **Montana (MT)** : 342
- **Arizona (AZ)** : 283
- **Utah (UT) :** 243

## 4.2 Dominant vegetation in the area

***Figure 1.1: Dominant vegetation around megafires:***

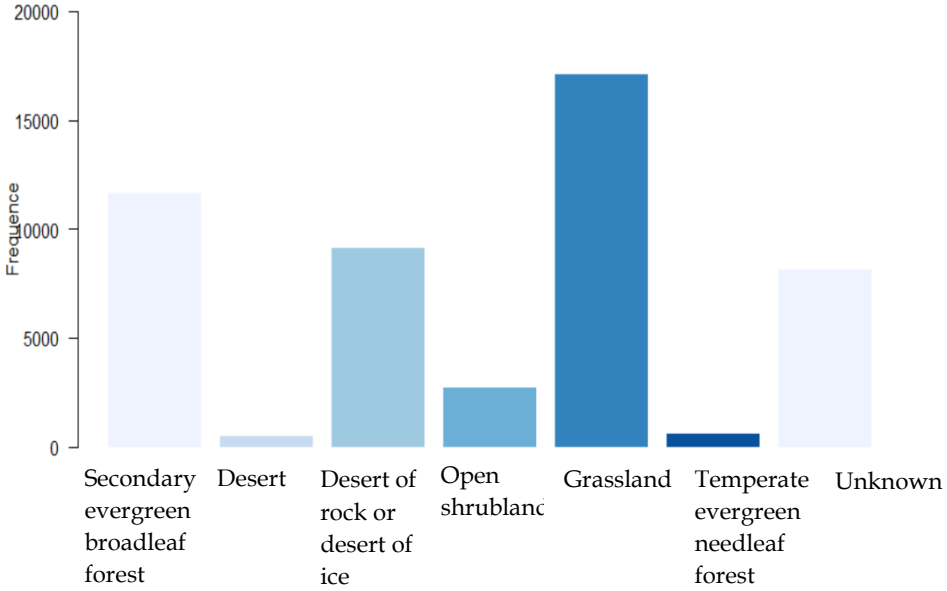



The vegetation that dominates around mega fires is often **rock or ice deserts**.
For example, we find this type of vegetation in Nevada and Montana.

We also have a majority where the vegetation is unknown (*this result is since fire-associated vegetation in Alaska is into this category as well as most of the vegetation in California and other states where megafire occurs*).

Nevertheless, this result can be explained by the fact that many **wildfires** contribute to deforestation and sometimes to **desertification**.

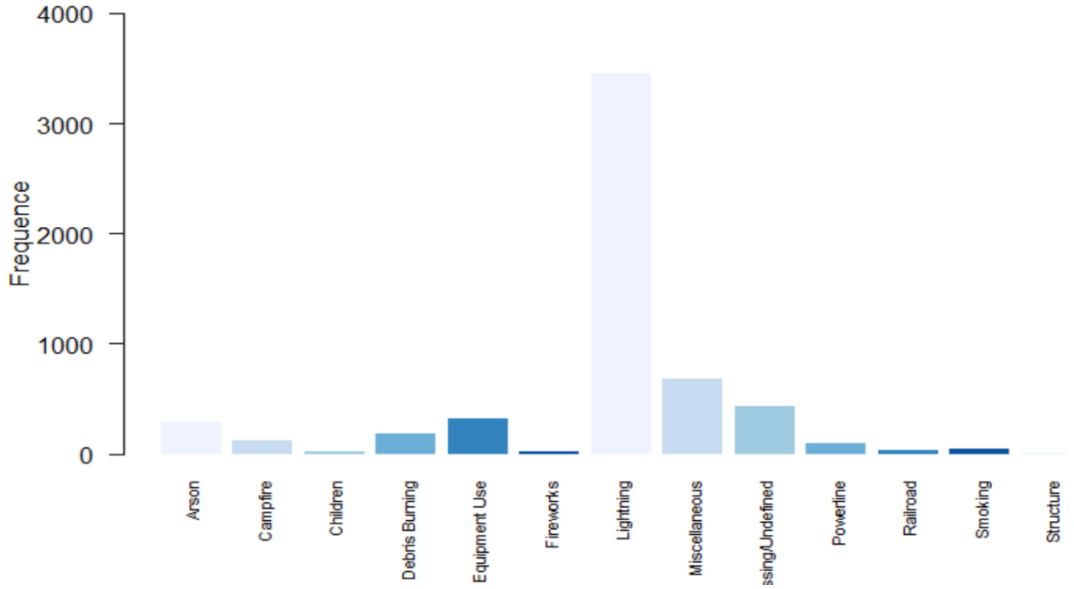
***Figure 1.2: dominant vegetation around smaller fires:***




The vegetation that dominates around smaller fires is often **grassland** *(mostly in New York and South Dakota)* and **secondary evergreen broadleaf forest** *(mostly in South Carolina and Georgia)*.

## 4.3 Cause of the fire

***Figure 2: Cause of megafires:***

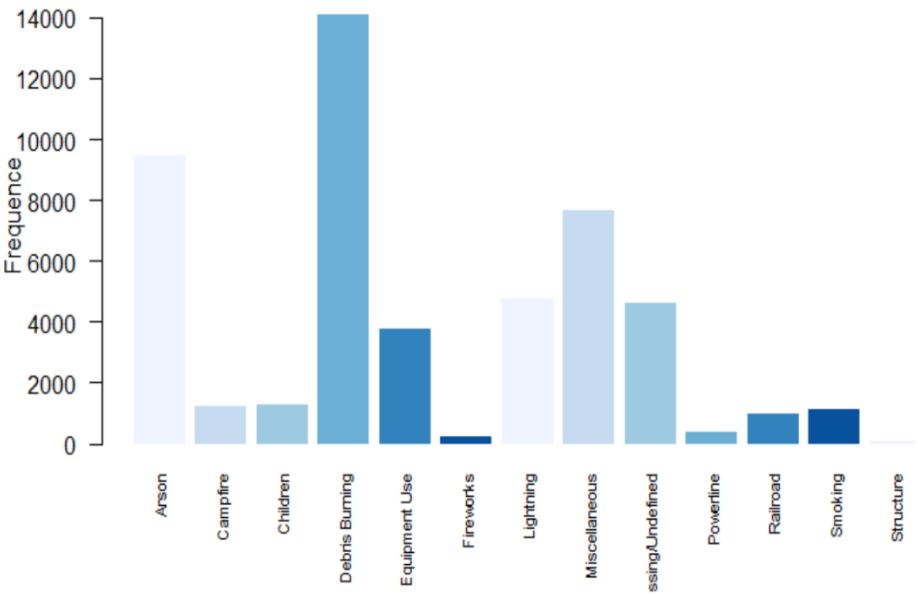



The principal cause of megafires is **lightning** (*it represents 60.76% of the causes of megafires*).

In detail, the return streaks of light are a series of strokes that produce the actual lightning flash that we see, and it is the hot lightning (*that occur for a longer period than cold lightning)* that starts fires by long-lasting bolts.

It is well known that **forest** fires are most likely to be mega-fires due to their **vegetation**. This result can therefore be linked to the fact that forests are often far away from **cities**. In fact, we see that the **lightning** cause have the ***highest mean of remoteness- distance to the closest city rescaled (0.306)***.

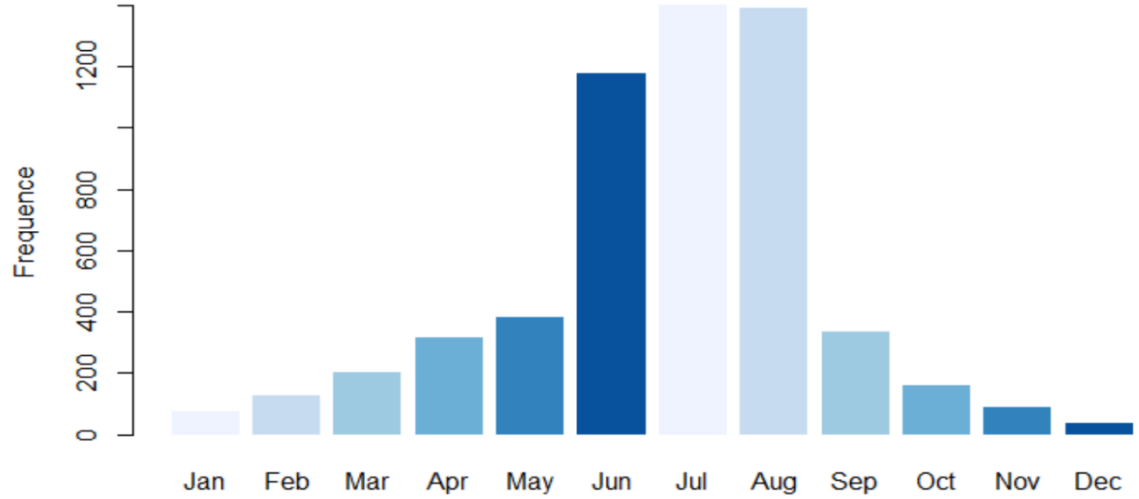
***Figure 2.2: Cause of smaller fires:***




Most of causes of other small fires are **debris burning**, **arson** and **miscellaneous** *(which represents respectively 28.36%, 19% and 15.43% of the causes of smaller fires)*. They are therefore often due to the human hand.

## 4.4 Month

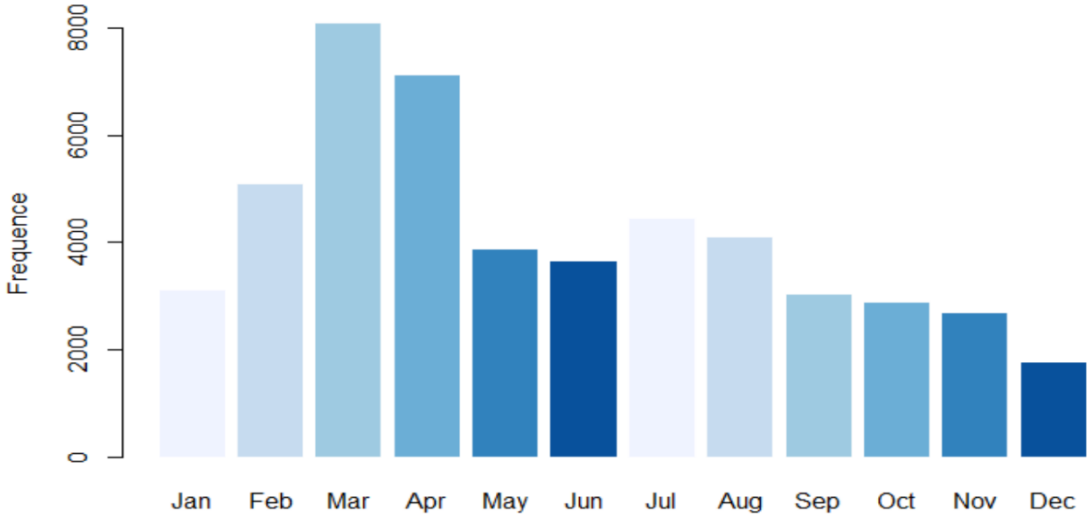
***Figure 3.1: Month where megafires occur:***




Most mega fires occur during the **summer** month: *June, July* and *August* (which represents respectively *20.69%, 24.64%* and *24.47% according to figure 3.1)*.

Furthermore, those three months have the **highest mean temperature** (*respectively 12.77, 13.14, and 12.86 Deg C*), the **highest mean wind** (*respectively 1.46,1.38 and 1.34 m/s*) and the **lowest mean precipitations** (*respectively 15.43, 10.5 and 11.86 in mm*).

***Figure 3.2: Months where smaller fires occur.***



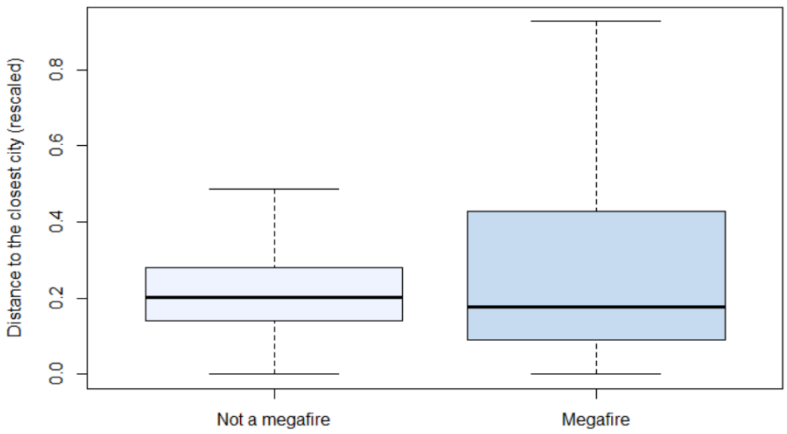

Unlike mega-fires, smaller fires typically occur in the month of **February, March**, and **April** (*which represent respectively 10.23%, 16.24% and 14.29% according to figure 3.2*).

Among the causes of **Arson** and **Debris Burning** fires, the following months are the most represented (*respectively 11.27%, 19.33% and 17.17% for Arson and 13.29%, 21.18£,16.8% for debris burning*). These causes were common in the smaller fires.

## 4.5 Distance to the closest city

This variable corresponds to the distance to the nearest city from the sight of origin of the fire (which has been rescaled from 0 to 1). Therefore, the larger the distance is, the larger this variable is.

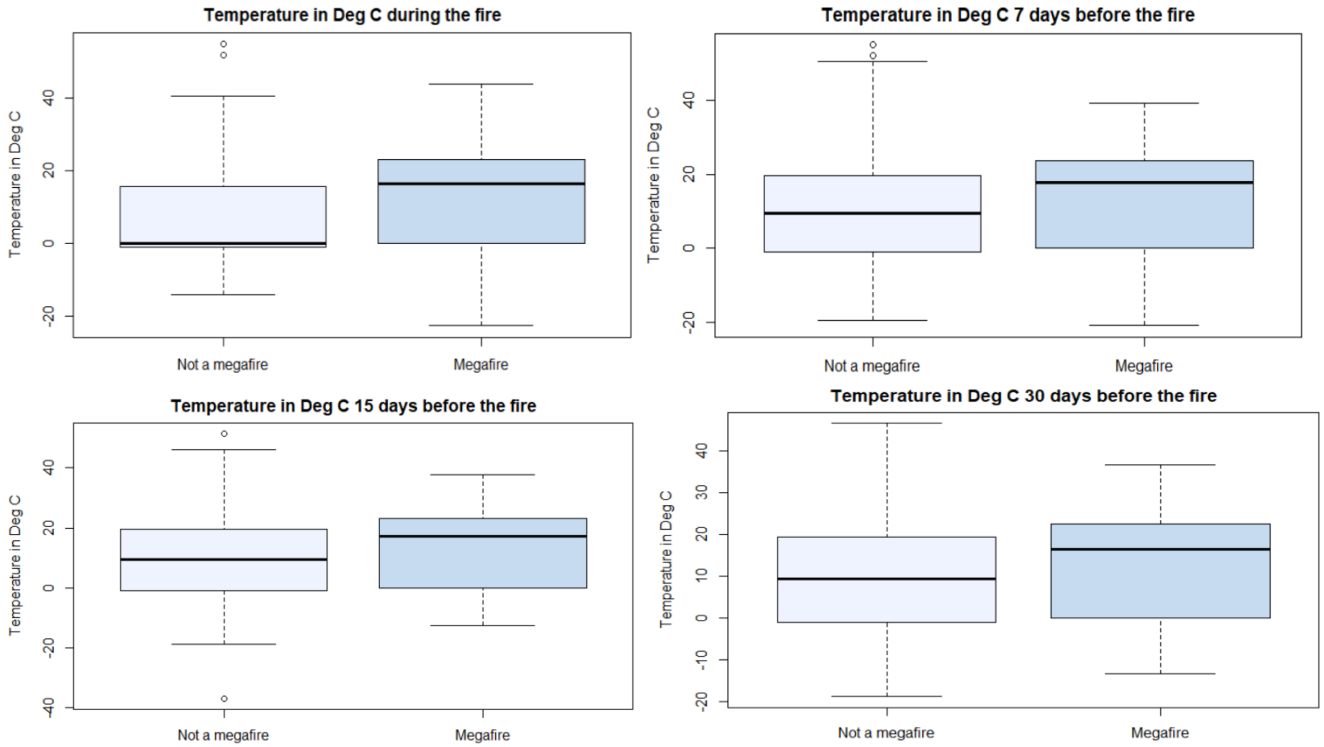
***Figure 4: Distance to the closest city (non-dimensional).***




Megafires have the highest mean distance to the closest city then others. This means that, on average, mega fires occur ***further away from cities*** than other fires. But looking at the median, we can see that it is the inverse.

## 4.6 Temperature in Deg C

***Figure 5: temperature in Deg C during the fire and 7,15 and 30 days before.***

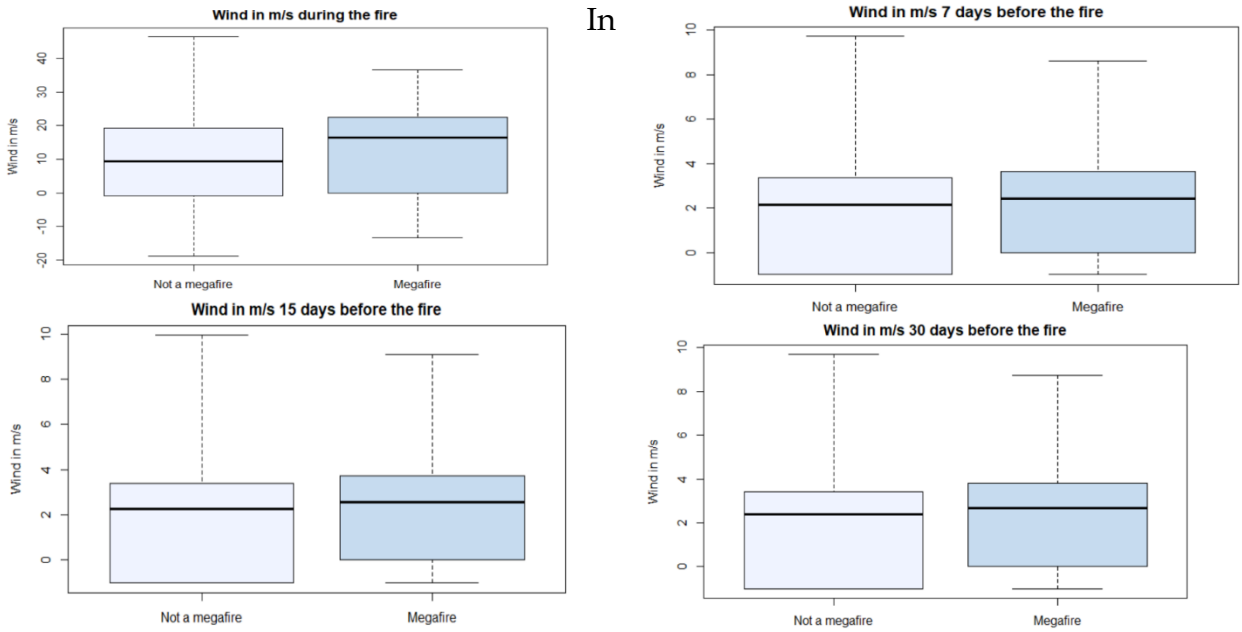



In any case, **mean and median** in terms of temperature in deg C are **higher for megafires** than for others.

Indeed, the average temperature in Deg C for megafires regarding the four periods is about **14.35°** against **9,75°** for smaller fires.

## 4.7 Wind in m/s

*Figure 6: Wind in m/s during the fire and 7,15 and 30 days before.*

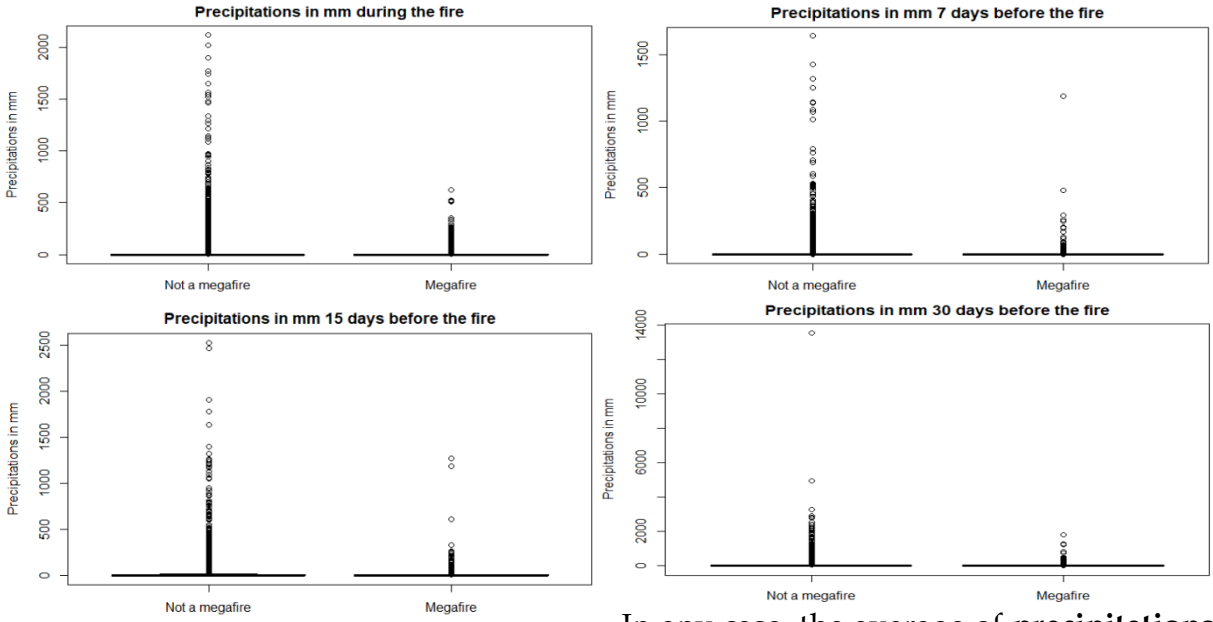



In any case, **mean and median are higher for megafires** than for other.

Indeed, the **average wind in m/s** for the four periods is about **2,2** for megafires against **1,61** for smaller fires.

## 4.8 Precipitations in mm

*Figure 7: Precipitations in mm during the fire and 7,15 and 30 days before*

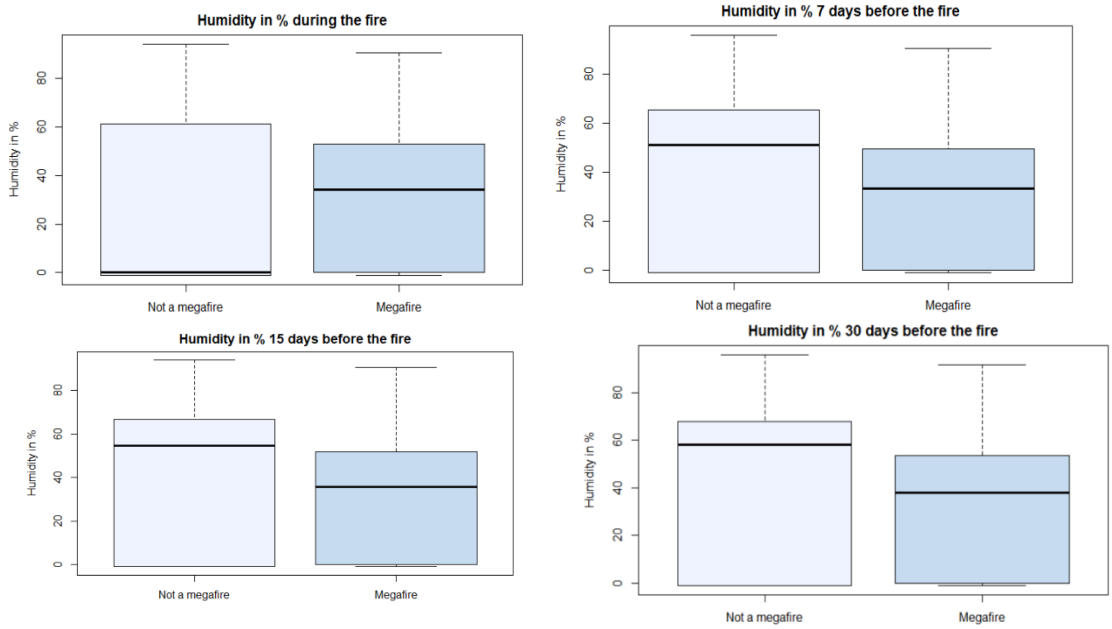



In any case, the average of **precipitations in mm is lower for megafires** than for other.

Indeed, if we look at the average precipitations for the four periods, we see that it is **5,68 mm** for megafires against **15,56mm** for smaller fires.

## 4.9 Humidity in %

***Figure 8: Humidity in % during the fire and 7,15 and 30 days before***




For the days **prior to the fire** (*7.15 and 30*), the **humidity was on average lower for the megafires** (*30.56%, 31.9% and 33.27% respectively*) than for the other smaller fires (*37.74%, 39.2% and 41.64% respectively*).

The same phenomenon is observed for the median.

Nevertheless, the **day of the fire**, the **average humidity is higher for the megafires** *(31.05%*) than for the others (*24.37%).*

The same phenomenon is observed for the median.

But this may be a consequence of the fire or could be explained by other external factors.

In the following section, we will therefore use predictive models and in particular logistic regression to capture causalities and not only correlations.

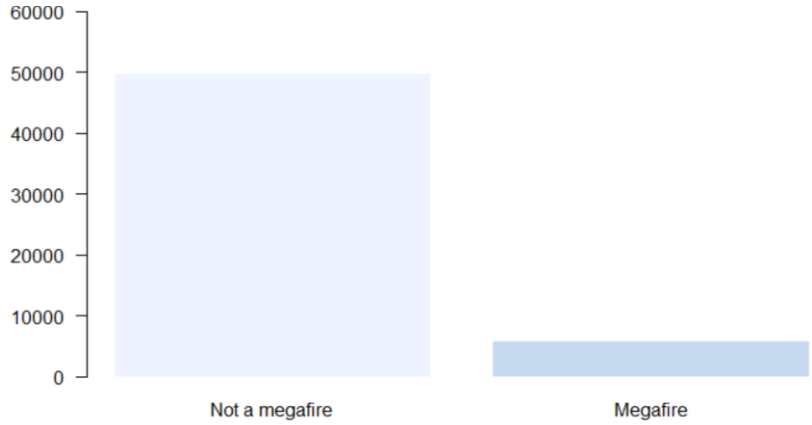
# 5 Machine learning methods

## 5.1 Data preparation

We randomly cut our base in half between a train set that we will use to **train** our models (***80%*** *of the data*) and a **test** set that we will use to evaluate the models *(**20%** of the data)*. Indeed, it is important to evaluate the performance of our models on a test sample since this study is intended to be generalized to other fire data.

Looking at Figure 9, we can see that the data is **unbalanced**.

***Figure 9: Frequence of megafire vs others***




This can cause problems of **misclassification** of the minority class (*mega-fires, being under-represented*) when using machine learning algorithms.

The **SMOTE** technique (*Synthetic Minority Over-Sampling Technic*) was therefore used on the train set in order to improve the performance of machine learning technics by **oversampling** the minority class.

We use **6 nearest neighbors** with **improved heterogenous distance function** (*which is appropriate for nominal attribute*): a randomly selected neighbor is chosen, and a synthetic example is created at a randomly selected point between the two examples in the feature space (***Appendix 1***)

The best solution for data not to be only based on synthetically created examples is to use in addition an **under-sampling technique**: *randomly decrease the examples in the majority class*.

We therefore end up with **44293** observations in the train set *(with half of the data containing megafires*).

## 5.2 Logistic regression

**Logistic regression** is a model used for **binary dependant variables** classification (*in our case, it is the variable **"Megafire"** which takes the value 1 if the fire size is greater than 1000 hectares and 0 otherwise*).

Indeed, it uses the **sigmoid function** to restrict **predicted probabilities from 0 to 1**, and, therefore a **threshold** can be assigned (*for example, if the predicted probability of being a megafire is greater than 0.8, we classify the observation as a Megafire*).

In our case, we assume a linear relationship between the **independent variables** and the **log-odds** of the event that **Megafire=1**.

The model has been written as below:

***Model 1: logistic regression***

$$Log\left[\frac{Megafire}{(1-Megafire}\right] = \beta 0 + \beta 1 cause + \beta 2 month + \beta 3\, Temperature + \beta 4\, Wind + \beta 5\, Humidity + \beta 6\, Precipitations + \beta 7\, Remoteness + \beta 8\, State.$$

We chose not to include the variables **longitude and latitude** since they are highly correlated with the distance to the nearest city (***figure 10 spearman correlation matrix***), and surely provide the same information as the states. In addition, their coefficient's interpretation has no meaning.

Also, we did not add in the model the *temperature, humidity, precipitation*, and *wind* 7, **15 and 30 days before the fire**. As we can see in ***Figure 10 (Spearman correlation matrix)***, these variables are highly correlated with each other and with those corresponding to the day of the fire.

But the main reason for not including these variables in the model is mainly summarized in ***Table 1 (GVIF rule).***

Indeed, after performing a logistic regression including these variables, computing the **Generalized Variance Inflation Factor** (***Table 1 GVIF***) and using the rule of *Hair and Al, 1995* ( $(\boldsymbol{GVIF}^{\frac{1}{2*Degree\_of\_freedom}})^2$**<**10 ), we realize that these variables pose a problem of ***multicollinearity***. It can increase the variance of estimated coefficients and make them very sensitive to minor changes in the model. The result is that those are unstable and difficult to interpret.

*Note: You will find explanations on this subject in the appendix below table 1.*

Find below the results we found with ***model 1 (logistic regression):***

- *Only for parameters that are significantly different from 0 at a level of 5%.*
- *We did not show the result for all states (only for highest and the lowest coefficient) because they are control variables.*

***Table 2: results of Model 1- Logistic Regression***

| Category | Variable | Exp(B) | References |
|---|---|---|---|
| **Cause of the fire** | *Campfire* | 1.25 | Arson |
| | *Childen* | 0.21 | |
| | *Debris Burning* | 0.31 | |
| | *Lightning* | 3.55 | |
| | *Miscellaneous* | 1.25 | |
| | *Missing/undefined* | 1.43 | |
| | *Powerline* | 2.23 | |
| | *Railroad* | 0.47 | |
| | *Smoking* | 0.54 | |
| **Month** | *January* | 0.76 | April |
| | *February* | 0.72 | |
| | *March* | 0.59 | |
| | *May* | 0.81 | |
| | *July* | 0.61 | |
| | *August* | 0.79 | |
| | *September* | 0.68 | |
| | *October* | 0.75 | |
| | *December* | 0.47 | |
| **Temperature** | *Temperature the day of the fire in Deg C* | 1.02 | |
| **Wind** | *Wind inm/s the day of the fire* | 1.33 | |
| **Humidity** | *Humidity in % the day of the fire* | 0.99 | |
| **Precipitations** | *Precipitations in mm the day of the fire* | 0.99 | |
| **State** | *Idaho* | 0.08 | Alaska |
| | *Rhode Island* | 3.08e-10 | |
| **Remoteness** | *Distance to the closest city* | 0.29 | |
| **Vegetation** | *Unknown* | 0.69 | Desert |

***Controlling for other variables, the odds of being a megafire (versus smaller fire) increase by a factor of:***

- 1.02 for a **one degree increase** the day of the fire (***positive effect on megafires***).
- 1.33 for a **one m/s concerning wind** the day of the fire (***positive effect on megafires***).
- 0.99 for a **1% increase in humidity** the day of the fire (***negative effect on megafires***).
- 0.99 for a **1mm increase in precipitations** the day of the fire (***negative effect on megafires***).
- 0.29 for an **additional unit of remoteness** (***negative effect on megafires***).

These meteorological results are in line with scientific facts that we presented in ***section 3 - literature review***. Indeed, fires develop mainly when the **temperature is high,** and the **soil is dry** and spread widely by **wind**.

Nevertheless, we can see those coefficients are close to 1, which means that it does not necessarily play a big role on whether a fire is a mega fire or not.

Regarding the remoteness variable that accounts for the distance to the nearest city, *it seems that a mega fire is more likely to occur than a smaller fire if it occurs near a city, rather than far away*. We had seen in ***figure 4 (distance to the closest city (non-dimensional))*** that the average distance to the closest city was higher for the mega-city, at least for our data.

***Regarding qualitative variables, controlling for all other variables, the odds of being a megafire:***

- When the cause is **lightning**, is **3.55 times** that of being a megafire when the cause is **Arson** (***highest positive effect of causes on megafires***).
- When the cause is **children**, is **0.21 times** that of being a megafire when the cause is **Arson** (***highest negative effect of causes on megafire***)
- In **may** is **0.81 times** that of being a megafire in **April** (***lowest negative effect of the month on megafires***).
- In **December** is **0.47 times** that of being a megafire in **April** (***highest negative effect of the month on megafires***).

In view of these results, we find once again that mega fires are more likely to be triggered by **lightning** (*and less likely to be caused by children*), *other things being equal*. This is what we saw in ***section 4 - descriptive statistics***.

Moreover, we had seen that the mega-fires occurred mostly in the **summer**. Nevertheless, this is because of the **weather conditions**, but these are included in the model, so we control their effects. Moreover, the fact that, *all things being equal*, there is more chance of having a mega fire in **April** can be explained by other factors not included in the model.

Now that we have shown the factors that influence megafires, we are focusing on the first goal of this project: to ***help fire management authorities by developing models that classify a fire that has just occurred as a fire smaller than 1,000 hectares or larger than 1,000 hectares***.

We use various evaluations of the model, all of which are based on a test sample external to the model as presented in ***Section 5.1 - data preparation***.

The **accuracy** of the model (*i.e well-classified observations in relation to all observations in the database*) is about **91,58%.**

Nevertheless, accuracy score is a good indicator of the quality of the model if there are equal number of samples belonging to each class.

But our **test set is not balanced**, only train set were balanced in order to build a better model. This is the reason why other important statistical metrics (***Appendix 2***) have to be seen (***Table 3: metrics for model 1- logistic regression***).

***Table 3: metrics for model 1- logistic regression***

| | Specificity | Sensitivity - Recall | Precision | F1 score |
|---|---|---|---|---|
| *Logistic regression* | 94,82% | 63,23% | 58,31% | 60,67% |

The ***ratio of smaller fires correctly predicted to all observations of smaller fires*** is about **94,82%**. So, our model seems to predict small fires well.

Nevertheless, the ***ratio of megafires correctly predicted to all megafires observations*** is lower: about **63,23%.** It is therefore more difficult for our model to predict megafires.

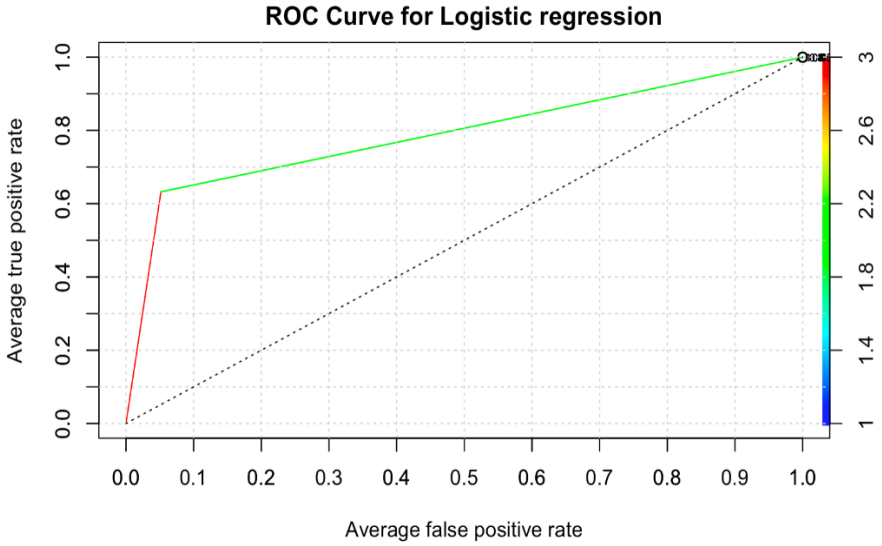
This is even more remarkable because the ***ratio of megafires correctly predicted to all observations classified as megafire*** is only about **58,31%.**

To conclude, the F1-score which is the ***weighted average of precision and recall*** is about **60,67%.**

Another parameter of the performance of the model can be **ROC curve** *(Receiver Operating Characteristics*). Graphically, the OCR measure is often represented as a curve that gives the **rate of true positives** (*the fraction of positives that are actually detected*) versus the **rate of false positives** (*the fraction of negatives that are incorrectly detected*).

Bigger is the area under the ROC curve (AUC) for the specific model, better the predictive power of the classifier is.

***Figure 11: ROC curve for model 1-logistic regression***




The **AUC** (*area under the red and green curve*) is about **0.79** (*close to 1*) which means that our model is far from being random (*diagonal where the true positive rate equals the false positive rate*).

*You will find in the next sections machine learning models, which have the disadvantage of being less interpretable, but have the advantage of delivering in most cases better performance.*

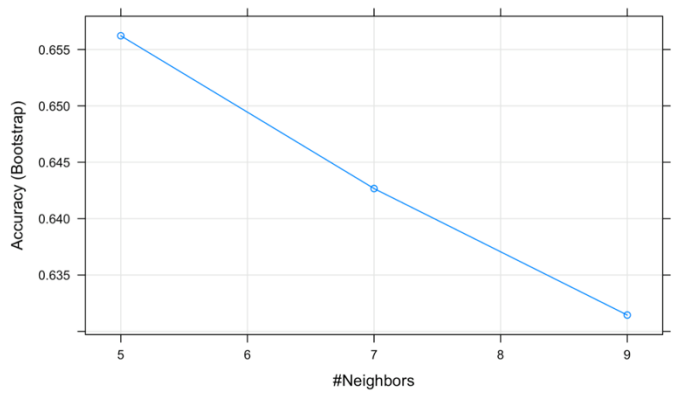
## 5.3 K Nearest Neighbours

K Nearest Neighbours (***appendix 4***) is an algorithm which receives a set of data that is already labelled on which it will be able to train and define a prediction model. This algorithm can then be used on new data to predict their corresponding output values.

It is important to compute an optimal **value of neighbours** that the model will consider.

In fact, for different number of neighbours the accuracy on **bootstraps** (*small samples of the same size repeatedly drawn, with replacement*) are computed, averaged. We choose the number of neighbours with the maximum one.

***Figure 12: Detection of number of neighbours for the KNN classifier.***




Looking at *figure 12* we can conclude that taking 5 closest neighbours to the original data point is the best choice.

After training KNN classifier we end up with an accuracy of **88.64%** *(fire cases in the test set have been correctly predicted by the classifier).*

***Figure 13: Statistical metrics for KNN Classifier.***

| | Specificity | Sensitivity | Precision | F1 score |
|---|---|---|---|---|
| *Random Forest* | 88.01% | 94.10% | 47.32% | 62.97% |

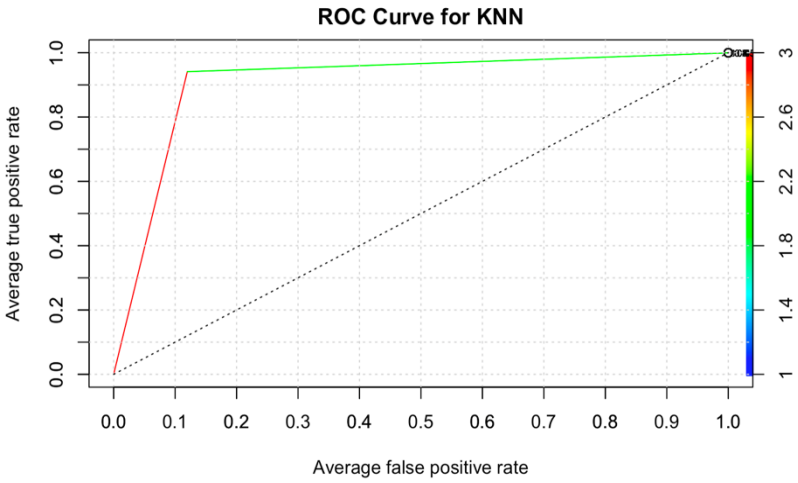
**The sensitivity** of the KNN classifier indicates that 94% of megafires were correctly predicted among actual megafires in our test set.

Looking at **specificity** of the KNN, we can conclude that 88% of test set's fires smaller than 1000 hectares were correctly predicted to all small fires.

According to the **precision**, 47% of predicted megafires were correctly predicted (among all observations predicted as megafire).

We can see that the ablity of the classifier to predict megafires is not as good as we supposed it to be. The **F1** score is about **0.63**.

***Figure 14: ROC curve for KNN Classifier.***




Accoding to *figure 14*, AUC for KNN classifier is equal to 0.91 (close to 1).

Considering all parameters seen above in this section, KNN can be described as a classifier with a good performance, nevertheless, F1 score is not so important as we could suppose.

## 5.4 Random forest

**Random Forest** is a *supervised learning technique* which comes from a set of choices in the form of a tree. The main interest of such a structure is its decision hierarchy.

This set of choices will make it possible to predict which class the fire belongs to (*mega-fires or not*).

The "forest" it builds, is an ensemble of decision trees, usually trained with the **"bagging"** method (*Boostrap aggregating: the general idea of the bagging method is that a combination of learning models increases the overall result).*

**Put simply: random forest builds multiple decision trees** (*it searches for the best feature among **a random** subset of features*) **and merges them together to get a more accurate and stable prediction.**

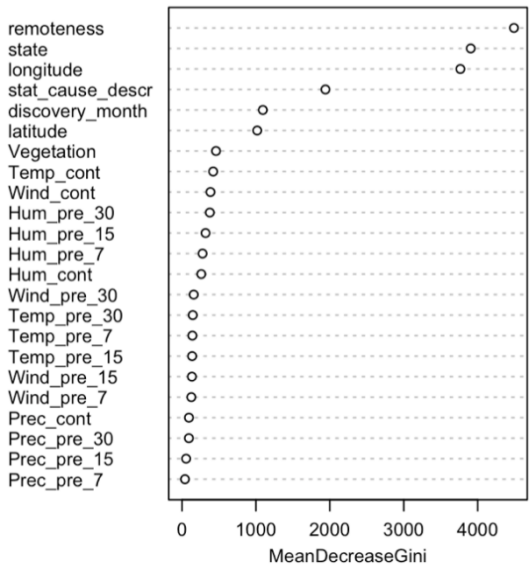
In our classification task this technique can be very useful, thanks to **the importance of variables** that can be determined. We can therefore continue and observe the importance of the variables in Random Forest model using **Mean Decrease GINI** that account for importance of variables **(Appendix 3).**

Higher Mean Decrease Gini is, more important the variable in the model is.

According to the ***figure 15*** the most important variables in the Random Forest model with 300 trees are (*in the descending order of importance*):

- **Distance to the closest city**
- **Longitude**
- **State**
- **Cause of the fire**
- **Month**
- **Latitude**

***Figure 15: Gini importance of variables***




Implementing the Random Forest which is composed of **300 trees** (*more trees did not bring much improvement in accuracy*), we obtain an accuracy of **97,55%.** *In other words, 97,55% cases of fires in test set were correctly classified.*

As it was already mentioned, accuracy is not always a reliable indicator of the performance of the model. The reason why, statistical metrics computed from the confusion matrix will be presented in *figure 16.*

***Figure 16: Statistical metrics for the Random Forest.***

| | Specificity | Sensitivity | Precision | F1 score |
|---|---|---|---|---|
| *Random Forest* | 97.51% | 97.89% | 81.83% | 89.15% |

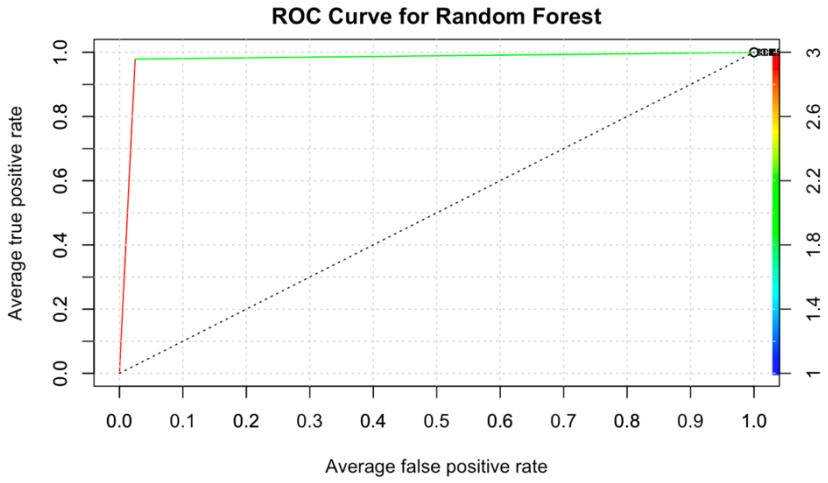
Looking at **specificity** of the Random Forest, we can conclude that **97.51**% of test set's fires smaller than 1000 hectares were correctly predicted to all small fires.

**The sensitivity** of the test is the proportion of megafires which were correctly classified among actual megafires. This proportion is about **0.97**. We can say that the classification of actual megafires is good.

According to the **precision** given by the Random Forest, **81.83**% of predicted megafires were correctly predicted (among all observations predicted as megafire).

The Random Forest has a good predictive power for megafires (*F1 score is about 0.89 close to 1*).

***Figure 17: ROC curve for the Random Forest Classifier.***




For Random Forest, AUC is equal to **0.977** (*very close to 1*)

Considering all the parameters, we can say that Random Forest has an important predictive power, and it manages to predict well megafires.

The question we can ask ourselves: *is it possible to use Radom Forest in Big Data?*

The answer is **Yes**. Nevertheless, considering the amount of information that is generated, the algorithm may encounter difficulties. These difficulties are mostly related to the **processing time**.

For this reason, a **parallel computation** can be implemented (***Appendix 5***).

*To demonstrate this, we have used other technics based on principles of the Random Forest:*

- **Sequential Random Forest in parallel** (the method is based on bootstrapping and on the independent construction of many trees, it is naturally adapted to parallel computation)
- **Sub-sampling Random Forest** (use of subsampling to reduce the size of the sample bootstraps. We use subsamples without replacement that represent 20% of the initial sample (10% for class 1 and 10% for class 0)
- **The m-out-of-n bootstrap Random Forest** (we use here only 20% of observations in the bootstraps (hence the name m out of n) but here the bootstraps are created without replacement so that the samples use all the observations in our database)
- **Divide and conquer Random Forest** (the original data set is divided into K subsets. So contrary to other methods, here we use several random forests and not just one)

In the figure below we present training time of each technique and the error on the test set we obtained.

***Figure 18 : Comparison between test errors and treatment times for different Random Forests.***

| | Sequential Random Forest | Sequential Random Forest in parallel | Sub-sampling Random Forest | The m-out-of-n bootstrap Random Forest | Divide and conquer Random Forest |
|---|---|---|---|---|---|
| **Test error %** | 2.03% | 2.03% | 2.69% | 2.47% | 2.56 % |
| **Training time** | 46.98seconds | 24.72seconds | 1.03seconds | 7.69seconds | 1.6seconds |

Comparing numbers, **Sequential Random Forest** (*basic*) provides a classifier with less error (*2.03%*).
Thanks to treatment in parallel, the time of treatment for Random Forest can be reduced (*from 46.98 seconds to 24.72 seconds*).

Nevertheless, considering both **treatment time and test error**, other technics proposed above can be interesting.
***Subsampling method, the m-out-of-n bootstrap method*** and ***divide and conquer*** allow to considerably **reduce training time by 45, 6, 29 respectively**.
Even if test errors are a lit bit greater for these technics, the differences are not especially important: 0.66 points of percentage, 0.44 point of percentage, 0.53 points of percentage, respectively.

There is a **trade-off between treatment time and the test error** when we speak about the implementation of the Random Forest in Big Data.

## 5.5 XG – Boost

**XGBoost** (*eXtreme Gradient Boosting*) ***(Appendix 6)*** is an optimized open-source implementation of Gradient Boosting.

XGBoost is part of the family of ensemble methods. The difference from traditional methods is that instead of training the best possible model on the data, ***we will train thousands of models*** on various subparts of the learning dataset and then have them vote on our decision.

Model built with XG -Boost predicts **99,29**% of fire cases correctly.

*Figure 19: Statistical metrics for XGBoost model.*

| | Specificity | Sensitivity | Precision | F1 score |
|---|---|---|---|---|
| *XGBoost Model* | 99.54% | 99.03% | 96.15% | 97.57% |

Looking at **specificity** of the XGBoost model, 99.54% of test set's fires smaller than 1000 hectares were correctly predicted among all actual small fires.

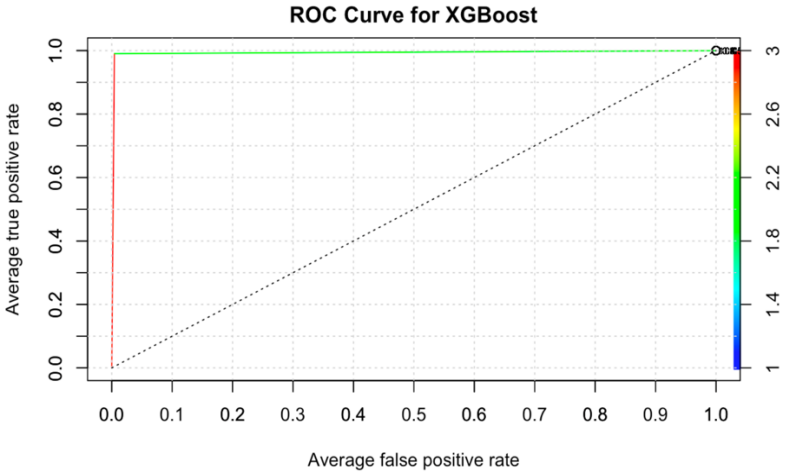
**The sensitivity** of the test is the proportion of megafires which were correctly classified among actual megafires. This proportion is about **0.99**. We can say that the classification of actual megafires is almost perfect.

According to the **precision**, 96.15% of predicted megafires were correctly classified (among all observations predicted as megafire).

We can affirm that XGBoost model has almost a perfect predictive power for megafires (F1 score is about 0.98, very close to 1).

In order to comapare this model with other ones, looking at the ROC curve and the value of the area under the curve.

*Figure 20: ROC curve for XGBoost Classifier.*




A simple look at this curve allows to conclude that the classifier is very **powerful**. The area under the curve is **0.99, almost one.**

It is worth mentioning that the model took almost one hour to train (*even if parallel computation was used*). It is true that in perspective of Big Data this algorithm can take a lot of computational time, but one more time a trade-off between treatment time and the quality of the classifier.

# 6 Conclusion

Based on the descriptive statistics (***Section 4***) and logistic regression (***Section 5.2***), we can conclude that the factors that influence the **probability of a fire becoming a mega fire** are **meteorological** (*higher temperature and wind, and lower humidity and precipitation*).

We were also able to show through the random forest (***section 5.4***) and its most important variable indicator that geographical variables such as *distance to the nearest city, states* play a role, as well as the *month of fire discovery* and its *cause*.

The latter results are less interpretable in terms of the **sign of the effect** (*positive or negative*) on mega fires.

Indeed, we can see that there is a trade-off between the **interpretability** of machine learning models and their classification **performance** as shown in figure below.

***Figure 21: Models' metrics***

| | Accuracy | Specificity | Sensitivity | Precision | F1 score | AUC |
|---|---|---|---|---|---|---|
| **Logistic Regression** | 91.58% | 94.82% | 63.23% | 58.31% | 60.67% | 0.79 |
| **KNN** | 88.64% | 88.01% | 94.10% | 47.32% | 62.97% | 0.91 |
| **Random Forest** | 97.51% | 97.89% | 81.83% | 89.15% | 97.51% | 0.97 |
| **XGBoost** | 99,29% | 99.54% | 99.03% | 96.15% | 97.57% | 0.99 |

Looking at table, comparing all implemented models, we can clearly see that **XG Boost** model is the best, in terms of **predictive performance**.

**KNN** and **logistic regression** also perform well, but only in terms of accuracy. Indeed, if we look at their **F1 score** we realize that both models classify mega fires less well than smaller fires.

Finally, **Random forest** is a good compromise in terms of performance/treatment time, especially if this project is to be generalized to fire stations.

It is therefore possible to classify the mega-fires of other fires to help the authorities in charge of fire management.

Nevertheless, the field experience of the fire departments remains especially important in order to carry out these projects.

# 7 Appendix

## ***Appendix 1: SMOTE and K nearest neighbors***

Several popular classification algorithms assume that classes are balanced and therefore construct the corresponding error function to maximize overall accuracy. In the case of an unrepresentative dataset, the result would lead to biased predictions towards the minority class (overlearning).

**SMOTE (Synthetic Minority Over-Sampling Technique)** is a technique used to process unbalanced datasets. First introduced by *Nitesh V. Chawla*, SMOTE is a technique based on nearest neighbours with a distance between data points in feature space.

The principle of SMOTE is to generate new samples by combining the data of the minority class with those of their close neighbors. Technically, SMOTE can be decomposed into 5 steps:

1) Choice of a characteristic vector of our minority class.

2) Selection of the nearest k-neighbors (k=5 by default) and choice of one of them at random.

3) Calculating the difference for each feature value i, and multiplying it by a random number between [0,1].

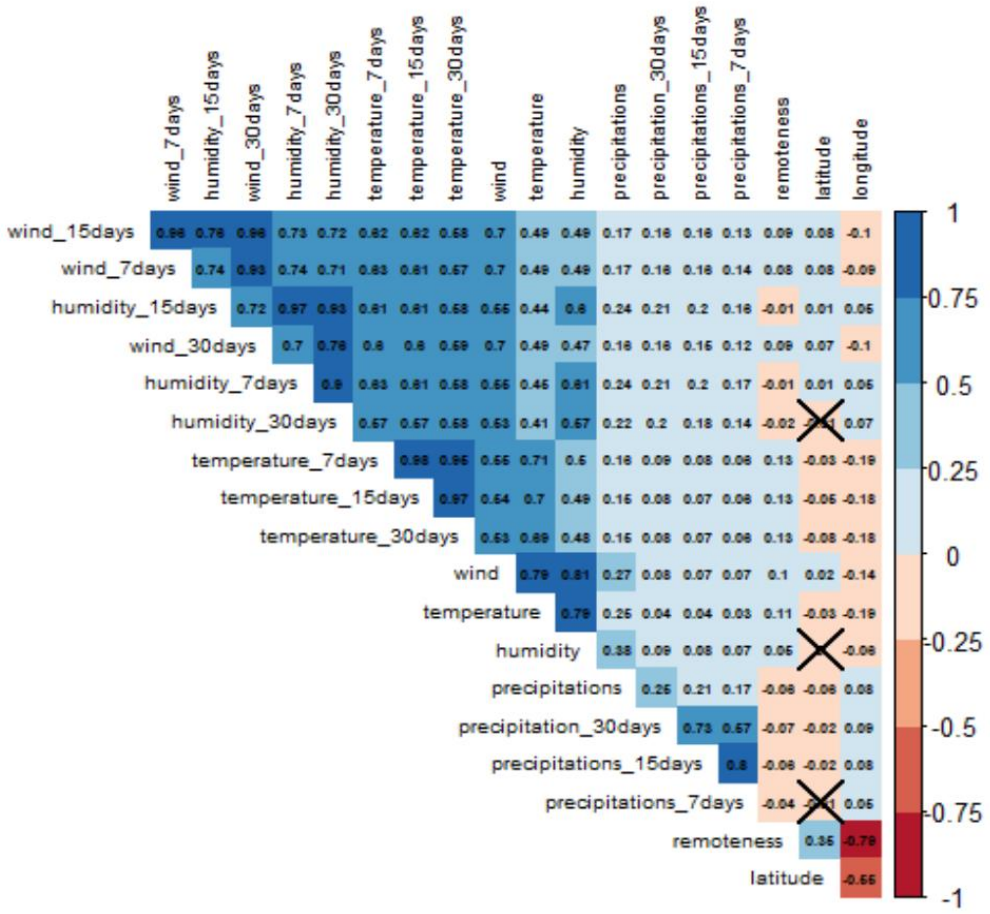
4) Adding the previous result to the feature value i of the vector in order to obtain a new point (new data) in the feature space.

5) Repeating these operations for each data point of the minority class.

We used the HVDM (Heterogenous Value Distance Metric) that returns the distance between 2 inputs vector x and y and which is defined as follows (with m the number of attributes:

$$HVDM(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)}$$

***Figure 10: Spearman correlation matrix (for quantitative variables).***



## *Table 1 : GVIF rule*

| Variables | GVIF | Df | GVIF^(1/(2*Df)) | Rule (GVIF^(1/(2*Df)))² <10 |
|---|---|---|---|---|
| **state** | 97,54 | 49 | 1,05 | 1,10 |
| **Vegetation** | 16,73 | 6 | 1,26 | 1,60 |
| **stat_cause_descr** | 2,43 | 12 | 1,04 | 1,08 |
| **discovery_month** | 6,60 | 11 | 1,09 | 1,19 |
| **Temp_pre_30** | 33,70 | 1 | 5,80 | 33,70 |
| **Temp_pre_15** | 68,07 | 1 | 8,25 | 68,07 |
| **Temp_pre_7** | 34,69 | 1 | 5,89 | 34,69 |
| **Temp_cont** | 8,05 | 1 | 2,84 | 8,05 |
| **Wind_pre_30** | 29,70 | 1 | 5,45 | 29,70 |
| **Wind_pre_15** | 51,64 | 1 | 7,19 | 51,64 |
| **Wind_pre_7** | 25,39 | 1 | 5,04 | 25,39 |
| **Wind_cont** | 7,50 | 1 | 2,74 | 7,50 |
| **Hum_pre_30** | 19,52 | 1 | 4,42 | 19,52 |
| **Hum_pre_15** | 39,34 | 1 | 6,27 | 39,34 |
| **Hum_pre_7** | 24,16 | 1 | 4,92 | 24,16 |
| **Hum_cont** | 7,32 | 1 | 2,71 | 7,32 |
| **Prec_pre_30** | 4,59 | 1 | 2,14 | 4,59 |
| **Prec_pre_15** | 5,60 | 1 | 2,37 | 5,60 |
| **Prec_pre_7** | 2,40 | 1 | 1,55 | 2,40 |
| **Prec_cont** | 1,30 | 1 | 1,14 | 1,30 |
| **remoteness** | 1,47 | 1 | 1,21 | 1,47 |

*Some information about columns in this table:*

- **GVIF** : inflation in **size of the confidence ellipse** or ellipsoid for the coefficients of the predictor variable **in comparison** with what would be obtained for orthogonal, **uncorrelated data**. *(In the case of a single coefficient, this specializes to the usual VIF).*
- **Df**: number of coefficients in the subset.

- ***GVIF^(1/(2*Df))****:* to make **GVIFs comparable** across dimensions. (*In the case of a single coefficient, it represents the proportional change of the standard error and confidence interval of their coefficients due to the level of collinearity)*.
- ***Rule****: 10 as the maximum level of VIF (Hair et al., 1995)*

## *Appendix 2: metrics*

- The **specificity** is the ratio of **correctly predicted negative observations to all negative observation**. **TN/TN+FP**. It is therefore designed to find the ability of the model to classify negative observations.
- The **recall** also called **sensitivity** is the ratio of **correctly predicted positive observations to all positive observations. TP/TP+FN**. It is therefore designed to find the ability of the model to classify positive observations.
- The **precison** is the ratio of **correctly predicted positive observations to all observations classified as positive**. **TP/TP+FP** is therefore designed to find the ability of the model to classify positive observations.
- The **F1 score** is the **weighted average of precision and recall *2*(Recall*Precision)/(Recall+Precision)***

## *Appendix 3: Mean Decrease Gini*

**Mean Decrease in Gini** is the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. This is effectively a measure of how important a variable is for estimating the value of the target variable across all of the trees that make up the forest.

- For each branch in split:
  - Calculate percent branch represents *(used for weighting)*
  - for each class in branch:
    - Calculate probability of class in the given branch.
    - Square the class probability.
  - Sum the squared class probabilities.
  - Subtract the sum from 1. *(This is the Gini Index for branch)*
- Weight each branch based on the baseline probability.
- Sum the weighted Gini index for each split.

## *Appendix 4: KNN*

*Start Algorithm*

- Input data: *D*
- A distance definition function: *d* (We use *Euclidian distance* which is defined as: $\sum_{i=1}^{n} |x_i - y_i|$ )
- An integer number: *K*

*For a new observation X for which we want to predict its output variable y:*

1. Calculate all the distances of this observation *X* with the other observations in the dataset *D*
2. Retain *K* observations from the dataset close to *X* using the distance calculation function *d*
3. Take the values of y for selected K observations:
    a. *If a regression is performed*, calculate the mean (or median) of retained *y*
    b. *If a classification is made,* calculate the mode of retained *y*
4. Return the value calculated in step 3 as the value that was predicted by K-NN for the observation *X*.

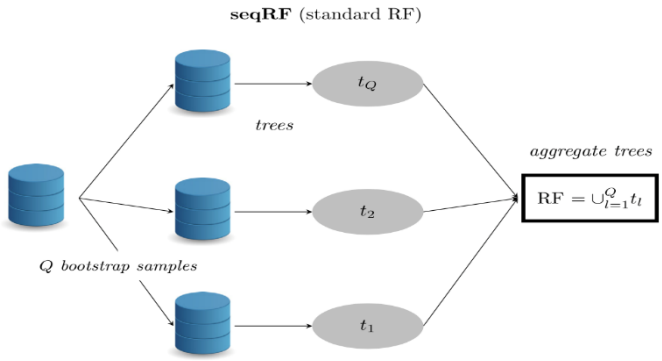
## *Appendix 5: Random Forest and big data*

**Sequential Random Forest:** A sequential calculation means that we use only one calculation process.

In more detail, this algorithm uses parallel learning. In fact, the goal of this algorithm is to reduce the variance of predictions through a bagging approach applied to several decision trees.

**Bootstrap Aggregating (bagging)** is a technique used to improve trees that produce a classification that is only slightly better than a random classification. It creates new samples by randomly drawing from the old sample with a discount (i.e. for a bootstrap sample about 63% of truly different observations are expected), with a random subset of variables (based on the principle of "random projections"), and the trees are trained with these subsets. Each tree will output a result. In the case of a classification, the class that appears most often is then chosen (majority vote). Moreover, by using independent models (several decision trees), it reduces the forecast error and there is less chance of "overfitting".

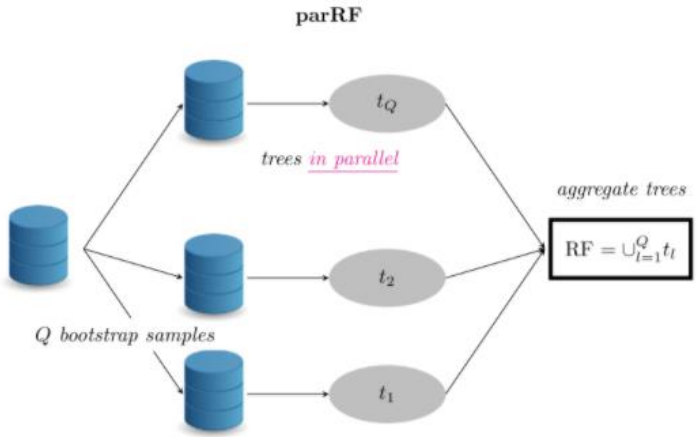
So basically, for each tree we select a bootstrap sample of individuals and at each step, the construction of a node of the tree is done on a subset of randomly drawn variables.




**Sequential Random Forest in Parallel:** As the method is based on **bootstrapping** and the independent construction of many trees, it is naturally adapted to parallel computation.

It is the same principle as for sequential random forest but instead of building all bootstraps and Q trees sequentially, they can be built in parallel.

Using a computer with 4 hurts, we were supposed to find 4 times shorter treatment time. But as soon as, the algorithm has to combine all results founded and aggregate them, the time of treatment is higher.

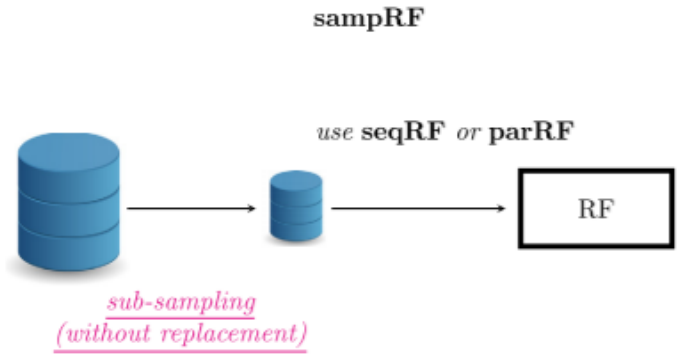



**Sub-sampling Random Forest:**

A simple technique used for very large databases, *use subsampling to reduce the size of the sample bootstraps*. Here subsamples without replacement are used, they represent 20% of the initial sample (10% for class 1 and 10% for class 0).

This technique can be used sequentially or in parallel (we personally do it in parallel).

Indeed, not all data are likely to be necessary to obtain accurate estimates in learning methods and it is found in the literature that sampling methods are an important and reliable way to deal with Big Data.

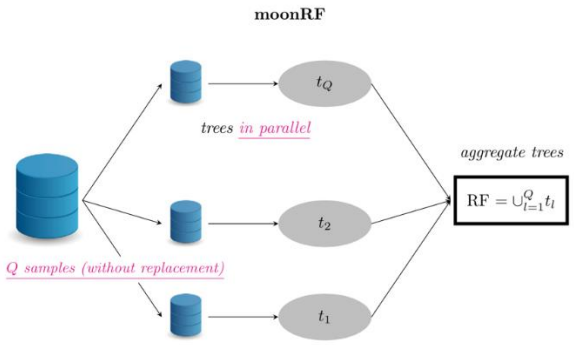



**The m-out-of-n bootstrap Random Forest:** proceeds by building bootstrap samples with only m observations taken without replacement in {1,...,n} (for m≪n). This method is illustrated in Figure

Initially designed to address the computational burden of standard bootstrapping, the method performance is strongly dependent on a convenient choice of m.

Only a percentage (in our case 20%) of observations in the bootstraps (hence the name m out of n) but here the bootstraps are created *without replacement* so that the samples use all the observations in our database. Parallel computation is also possible.

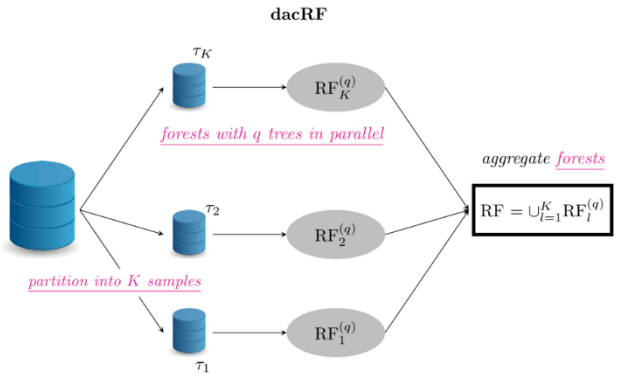
In order to implement this technique "without replacement", the calculation is here done tree by tree and not thanks to several trees that we put in the Random Forest function created by R.




**Divide and conquer Random Forest:** A standard alternative to deal with massive datasets while not using subsampling is to rely on a "divide-and-conquer" strategy. The large problem is divided into simpler subproblems and the solutions are aggregated together to solve the original problem. The data are split into small subsamples, or chunks, of data.

A random forest with q trees is constructed from each of the subsets and all the forests are finally grouped into a final random forest. All this is done using parallel computation.

So unlike other methods, here we use several random forests and not just one.




## *Appendix 6 : XG-Boost*

**XGBoost** has a lot of advantages:

1. *Parallelization:* by default, the algorithm uses all the cores of the machine's microprocessor, which saves a lot of time.

2. *Regularization:* XGBoost includes regularization, which avoids over-fitting. Thus, one has more chances to obtain a model that performs well on training samples, test samples, and also on new data. It is said to be generalizable.
3. *Non-linearity:* Being based on decision trees, XGBoost captures all types of data linkages, including non-linear ones.
4. *Cross validation:* integrated in the algorithm, there is no need to program it otherwise.
5. *Missing data:* managed natively by the algorithm. It is able to capture and understand their structure, in case they are not due to pure chance.
6. *Flexibility:* possibility to do regression, classification and ranking. Moreover, it is possible to define an objective function to be optimized during the training of the model. Very useful in classification when you want to give more weight to false negatives than to false positives or the reverse.
7. *Availability & scalability:* XGBoost is usable with most platforms (Windows, Linux, macOS). It can also run in a distributed way.
8. *Tree pruning (tree-pruning):* Consists of removing branches (terminal parts) of decision trees that are not very useful for prediction. This simplifies the final model and improves predictive performance (generalization).

**Gradient boosting** combines weak "learners" into a single strong learner in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model to predict values  by minimizing the mean squared error , where i indexes over some training set of size n of actual values of the output variable y.  The principle of boosting is to improve the prediction quality of a weak learner model by giving more and more weight to values that are difficult to predict during learning. In this way, the model is forced to improve.

# 8 Bibliography


https://www.kaggle.com/capcloudcoder/us-wildfire-data-plus-other-attributes

Atul K. J, Meiyappan, Prasanth, (2012) "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years.", Frontiers of Earth Science 6.2 122-139

Cartault J.L., Clair B., Kapp D., « Incendie »

Cartlon J., (2020) « Pourquoi les incendies de forêt sont si graves cette année dans l'Ouest américain »

Charty C., Creuchet B., Grelu J., Lafitte J.J., Laurens D., Le Gallou J.Y., Le Quentrec M., (2010) « Changement climatique et extension des zones sensibles aux feux de forêts »

Coudur B., (2015) « Influence de la végétation et du relief dans les feux de forêt extrêmes : étude de la dégradation, de l'accumulation et des propriétés de combustion des composés organiques volatils issus des feux de forêt »

Genuer R., Poggi J.M., Tuleau-Malot C., Villa-Vialaneix N., (2017) "Random Forests for Big Data " , Big Data Research 9 28–46

National Parc Service, (2018) « Wildfire Causes and Evaluations »

NOAA National Centers for Environmental Information, (2001) "Integrated Surface Hourly 1992-2015"
ftp://ftp.ncdc.noaa.gov/pub/data/noaa/

Karen C., Short (2017) "Spatial occurrence data for the United States 1992- 2015", 4th Edition. Fort Collins, CO: Forest Service Research Data Archive
https://doi.org/10.2737/RDS-2013-0009.4