Master's Thesis

Master in Telecommunication Engineering

Big Data against forest fires in Galicia

Pablo Pérez Brenlla

Supervisor: José López Vicario

Department of Telecommunications and Systems Engineering

Escola Tècnica Superior d'Enginyeria (ETSE)
Universitat Autònoma de Barcelona (UAB)

January 2016

El sotasignant, *José López Vicario*, Professor de l'Escola Tècnica Superior d'Enginyeria (ETSE) de la Universitat Autònoma de Barcelona (UAB),

CERTIFICA:

Que el projecte presentat en aquesta memòria de Treball Final de Master ha estat realitzat sota la seva direcció per l'alumne *Pablo Pérez Brenlla*.

I, perquè consti a tots els efectes, signa el present certificat.

Bellaterra, *14 de gener de 2016*.

Signatura: *José López Vicario*

“To my family, friends and teachers that,
unconditionally, have supported me along the way.”

Barcelona, January 2016

Resum:

Galícia concentra una increïble quantitat d'incendis forestals any rere any. Tot i que s'està fent un enorme esforç en matèria de prevenció i extinció, sembla que no n'hi ha prou. Amb la idea d'ajudar en aquest sentit, aquest projecte utilitza tècniques de Machine Learning amb l'objectiu d'aconseguir una millor distribució dels mitjans existents. Les dades analitzades corresponen a més de 99.000 incendis declarats a Galícia entre el 2000 i el 2014 i a les condicions meteorològiques de cadascun dels dies d'aquest període. El sistema proposat està dividit principalment en tres seccions. En la part descriptiva, es fa un estudi temporal, geogràfic i causal general i també a nivell de municipi. Aquest anàlisis a baix nivell ens ofereix un informe detallat per cada municipi, amb informació valuosa per bombers i autoritats locals, desconeguda fins ara per ells. La part predictiva consisteix en un algorisme que prediu si va a produir-se un incendi en un municipi en un dia determinat, amb una taxa d'encert acceptable. Per últim, la part prescriptiva indica com utilitzar les dues seccions anteriors per tal d'establir un nivell d'alerta i mesures específiques en diferents zones.

Resumen:

Galicia concentra una increíble cantidad de incendios forestales año tras año. Aunque se está haciendo un enorme esfuerzo en materia de prevención y extinción, parece no ser suficiente. Con la idea de ayudar en este sentido, este proyecto usa técnicas de Machine Learning con el objetivo de lograr una mejor distribución de los medios existentes. Los datos analizados corresponden a más de 99.000 incendios declarados en Galicia entre 2000 y 2014 y las condiciones meteorológicas para cada día de este periodo. El sistema propuesto está dividido principalmente en tres secciones. En la parte descriptiva, se hace un estudio temporal, geográfico y causal general y también a nivel de municipio. Este análisis a bajo nivel nos ofrece un informe detallado para cada municipio, con información valiosa para bomberos y autoridades locales, desconocida hasta ahora para ellos. La parte predictiva consiste en un algoritmo que predice si va a ocurrir un incendio en un municipio y día dado, con una tasa de acierto aceptable. Por último, la parte prescriptiva indica cómo usar las dos secciones anteriores para establecer un nivel de alerta y medidas específicas en distintas

Summary:

Galicia focus an incredibly amount of forest fires year after year. Even if an enormous effort is being made in terms of prevention and extinction, it does not seem to be enough. With the idea of helping on this issue, this project uses Machine Learning techniques in order to achieve a better distribution of the existing resources. The data analyzed corresponds to more than 99.000 fires declared in Galicia since 2000 to 2014 and weather conditions for every day in this period. The proposed system is mainly divided into three sections. At the descriptive part, a general temporal, geographical and causal study is made and also at a municipal level. This low level analysis delivers us a detailed report for every municipality with many valuable information for local authorities and firefighters, unknown to them so far. The predictive section consists on an algorithm that predicts whether a fire will take place on a given day and municipality with an acceptable success rate. Lastly, the prescriptive part shows how to use these two previous parts in order to establish an alert level and specific measures for some different areas.

Index of Contents

1. INTRODUCTION

This Introduction chapter has the aim to show to the lector some key information to understand the rest of the text. Firstly, in this initial chapter, the motivation behind this project will be introduced, in which the severity of the situation concerning forest fires in Galicia is shown. After this, the geographic and demographic context of the autonomous community is presented in the Context section, as well as some data about forest mass and firefighting resources distribution. Next, the main goals of this thesis are set. Last, the organization of the rest of the paper is shown, so the reader gets a first idea about the structure of the document.

1.1. Motivation

It is very well known in Spain that most of the forest fires (also called wildfires) occurred in this country, are concentrated in the north-western part and more specifically in Galicia. With the naked eye, it might seem paradoxical that this region is the hardest hit by forest fires as long as it has a very humid climate characterized for a lot of rains along the whole year (the four provincial capitals in Galicia are in the top 8 of the 53 Spanish provincial capitals in terms of rain) [1], contrary to the drier and warmer Mediterranean area. Even if this fact is known, what people are usually unaware of is how extreme this situation is. For understanding how critical the case is, two pictures are shown below. The first one (Figure 1) indicates the distribution of forest fires from 1996-2005 among the autonomous communities in Spain. It can be seen that more than **half of the fires in Spain are produced in Galicia, even if it represents less than 6% of the Spanish total territory**.

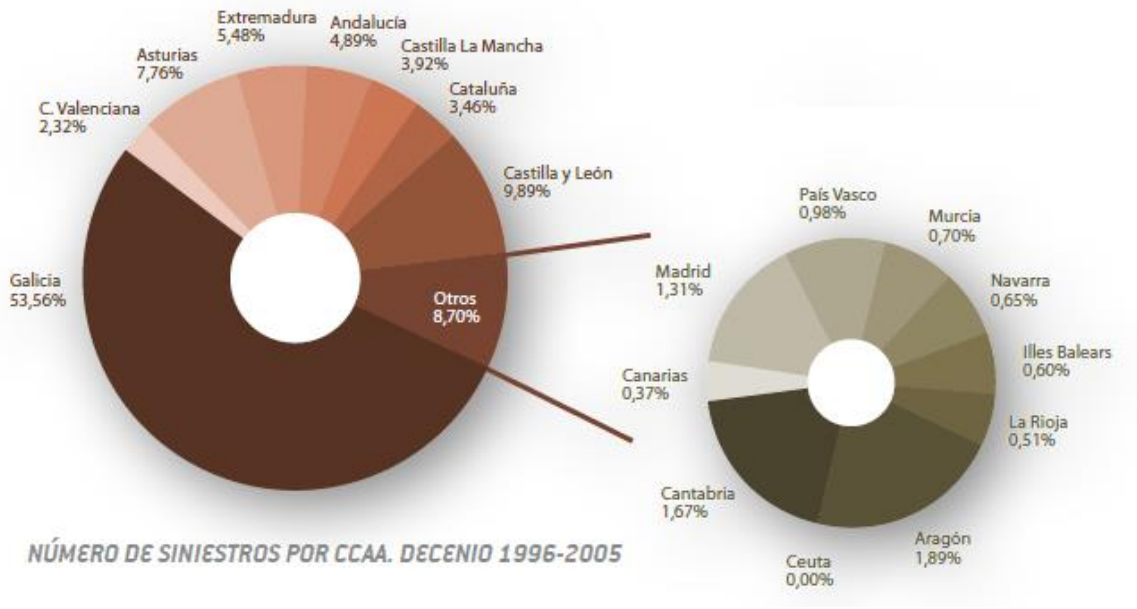


Figure 1: Number of forest fires over autonomous communities in Spain from 1996 to 2005.
(Source: *Ministerio de Medio Ambiente y Medio Rural y Marino, Evita el fuego… la diversidad es vida.)*

By taking into account this data, it can be deduced that the climate is not the most important factor affecting this problem. Main cause in forest fires is man-made, and also in Galicia's case, the numbers are striking. Around 35% of fires ignited in Spain are caused due to negligence or accident and 45% are intentional [2]. **In Galicia**'s case, just 5% were caused by negligence or accident and **81% were intentionally** caused over the last 15 years.

Figure 2, also astonishing, shows the location of every single fire taking place in Galicia since 2000 to 2014. By placing a dot for every ignited fire, we can easily distinguish Galicia's map. In last 32 years, 1.5 million hectares were burnt in the region, equivalent to more than half of the total surface of it (2.95 million hectares) [3].

By looking at these results, it is obvious that there are three important problems concerning forest fires in Galicia. A big one concerning extinction, and two more even bigger; sensitization and prevention.

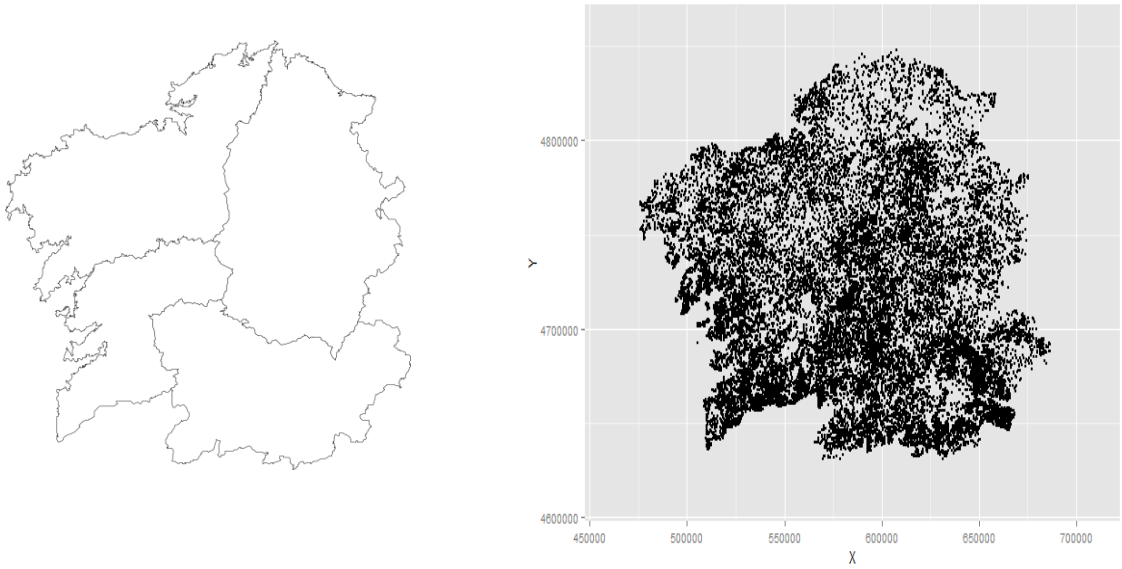


Figure 2: Galicia's map painted by the fires produced since 2000 to 2014.

Mankind behind almost every fire adds a very random component to the fires distribution. In the region, it seems to be really bizarre and unpredictable, on both geographical and temporal cases. Fires start in very different locations from year to year and also differing a lot temporally. Taking a look to the following plot (Figure 3), it can be seen that not in every year, summer months are the hardest in this sense.

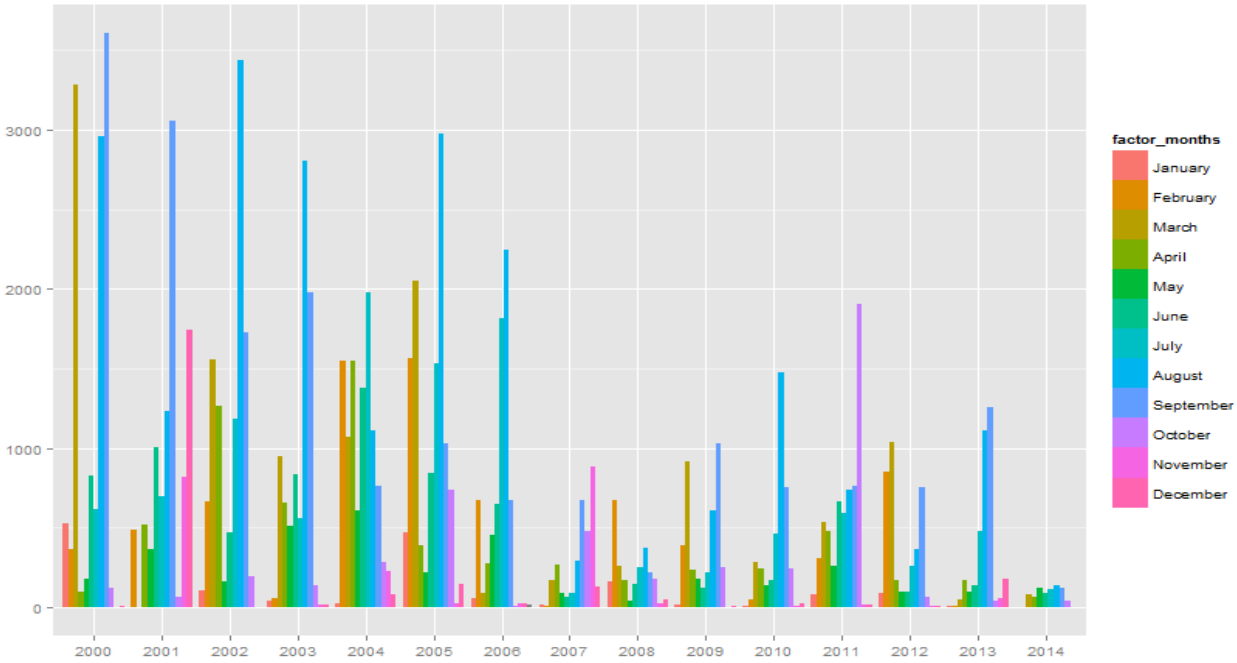


Figure 3: Number of fires in Galicia distributed over months and years.

This extremely high amount of fires supposes every year a brutal economic and ecological damage in Galicia, while endangering human lives. Drastically reducing these figures in the next years should be an urge for the Galician government, but a good solution seem that have not come yet.

1.2. Context

Geographically talking, Galicia is one of the 17 regions of Spain, more specifically the most northwestern one (Figure 4). With 29.574 km2, it occupies the 5.8% of the Spanish territory and has a population of around 2.73 million people. Although depending on the Spanish government, it also has its own government with important competences on different ambits such education, health system and the important at this case, on firefighting.

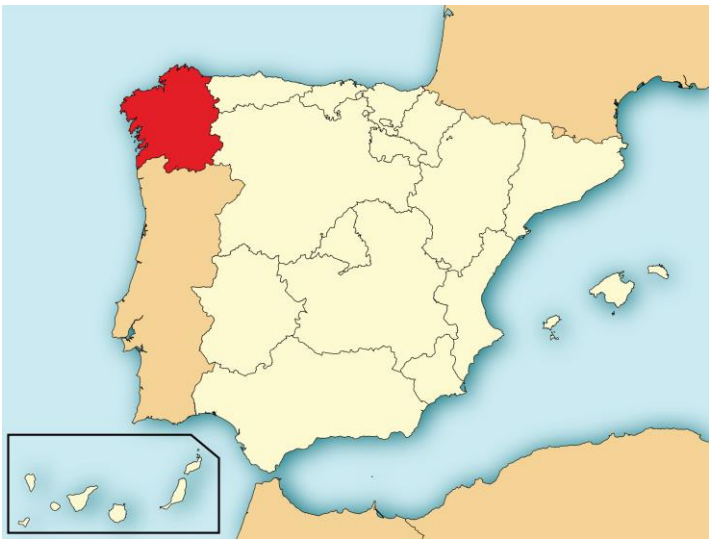


Figure 4: Map of Spain where Galicia is highlighted in red color.
(Source: https://es.wikipedia.org/wiki/Galicia)

At national level, Galicia is well known for owning a really important wooded area. Around half of the territory is occupied by trees, while the Spanish mean is at the 29% [4]. Field uses distribution in Galicia is shown in Table 1:

Field Use	Surface (ha)	%
Forest	2.030.681	68.66%
Agricultural	822.626	27.82%
Artificial elements	81.520	2.76%
Wetland	2.311	0.08%
Water	20.307	0.69%
Total	**2.957447**	**100.00%**

Table 1: Field uses distribution in Galicia.

From the total forest surface, 70% (1.424.094 ha.) corresponds to wooded area. 31% of this wooded area is covered by conifers, the 52% for leafy areas (natives and not natives) and 17% for a mixture of both of them [5].

Politically talking, the region is divided into 4 provinces (Figure 5a): A Coruña, Lugo, Ourense and Pontevedra (thick black lines on the map below left) with unequal sizes (7.950, 9.856, 7.272 and 4494 km2 respectively). Most of population is distributed along the occidental provinces (Figure 5b), around 1 million in A Coruña and other million in Pontevedra and just around 350.000 people in Lugo and the same in Ourense. In such a way, population density is also bigger in Pontevedra (211 inhabitants per km2), followed by A Coruña (141) and far from them Ourense (43) and Lugo (34).

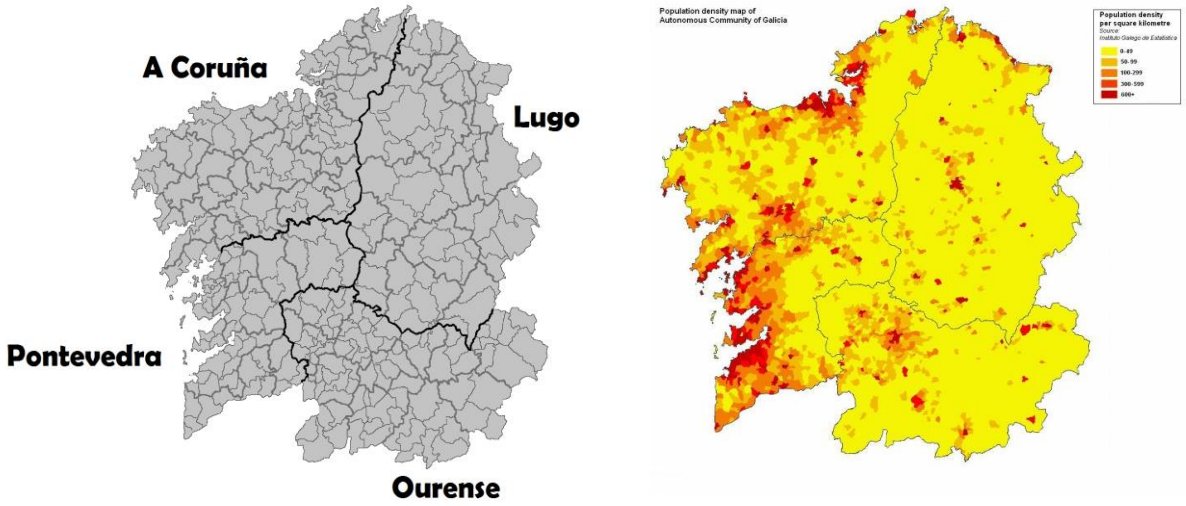


Figure 5: Geographic (a) and demographic (b) distribution of Galicia.
(Source: https://es.wikipedia.org/wiki/Galicia)

In a lower level, the region is divided into county levels (thick grey lines on the map) and these counties group different municipalities which are the lowest level entity with government. These different municipal terms also have a very different distribution of territory and population from one another.

The geographical distribution of firefighting resources is divided into 19 different districts, with district chiefs who coordinate the different local brigades (see picture below). These local brigades can move from one municipality to another, or even between districts in extreme cases.

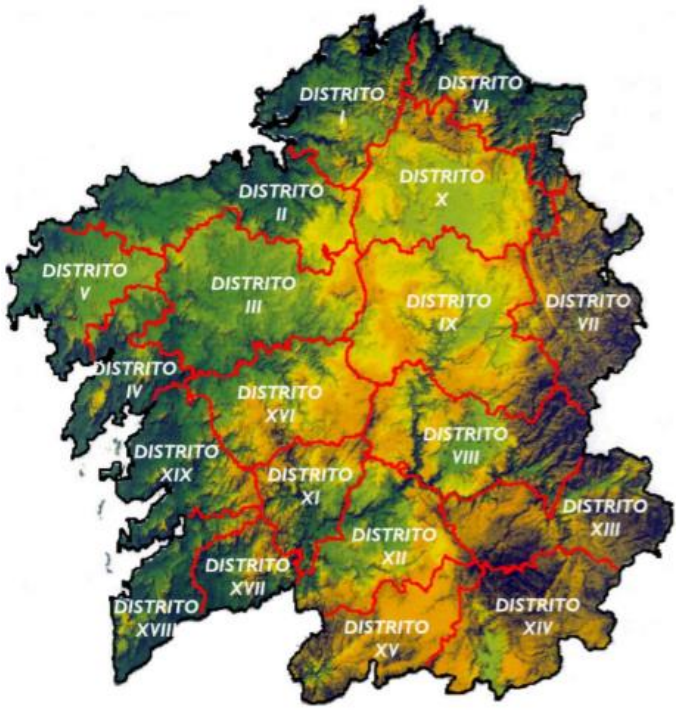


Figure 6: Geographic distribution of districts concerning firefighting.
(Source: *Pladiga 2015)*

1.3. Goals

The main goal of this thesis is to make prevention labors easier and **achieve a more efficient distribution of the existing resources** through obtaining a better model on the distribution of fires in Galicia, at a general and at a municipality level, by using Machine Learning techniques. This modelling will be created by analyzing the fires at a geographical, temporal

and causal way. As it has been told in the Motivation section, Galicia's fires seem to have a very important random component, as long as they do not seem to follow any logical distribution. Another important goal is to model this apparent randomness into valuable information that could help at distributing resources.

In order to get this big goal, we will set three different sub-goals (Figure 7). The first one is to give a descriptive understanding to competent authorities about what has happened, where, when and why in the past, at an overall and local level. Therefore, this descriptive analysis will be made firstly for the whole region and after that for every municipality.

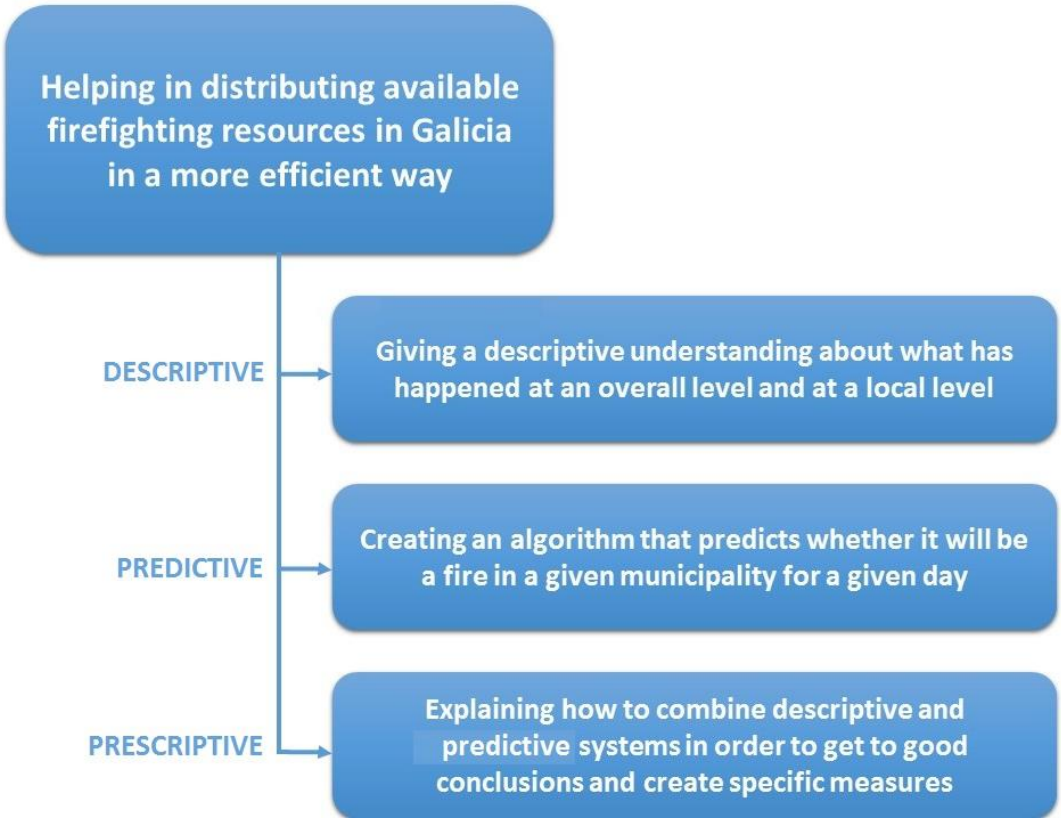


Figure 7: Goals.

So far, the firefighting forces distribution was based on the Forest Fires Daily Risk Rate or IRDI (Índice de Risco Diario de Incendio forestal) by its Galician initials [6], a five levels heat-map, updated every day, showing higher or lower levels of fires risk depending on weather conditions for every zone. This may not be a very bad mechanism to distribute forces

from a high level perspective, but it has been showed that this heat-map does not provide much information for a local level forces distribution.

By interviewing some squad local leaders (1st level chiefs, just above field firefighters), we have been told that, while doing surveillance labors, they are moving most of times by guessing or intuition, but without having a real knowledge about what are the hottest regions inside their action area or the most dangerous weeks, historically talking. From these conversations, it was thought that **providing this information to local entities is really important**, as long as it has been observed that the fires distribution varies quite a lot from village to village.

The second sub-goal consists on **providing an algorithm that predicts whether it will be a fire in a given municipality for a given day**. Having an algorithm such as this, based not only in weather conditions, but also in historical fires, can predict with a reasonable effectiveness fires, and this tool can help to focus efforts in certain parts. It is not thought as a substitute to IRDI, but a complement, as long as IRDI may have more weather and terrain factors into account, but this algorithm works also with historical fires data.

Last sub-goal is to explain to the final users how to proceed while **combining the descriptive and predictive system with the aim of reaching better decisions**. It is considered that both mechanisms taken together in a smart way, considering also the local knowledge that firefighters already had of their region, may reduce the number of fire incidents in a significant way.

1.4. Organization

Once the introduction was made in Chapter 1, where a first approach to the existing problem was commented and the most important goals were set, Chapter 2 will serve for knowing the state of the art concerning fires prevention in Galicia and some technological projects that were developed for helping in this issue in Galicia and abroad.

Next, Chapter 3, is the Tools section, where an introduction to the technology and software used for developing this project will be made. More particularly, we will take a look on R

programming language, RStudio as the environment to work with this language and some Machine Learning principles that will help in solving the problem. Moreover, the Logistic Regression Model will be commented, since it was used for building the algorithm used in the predictive system. Finally, the required previous studies done are explained in this section.

With this clear, a system that can lead to a good solution of the problem is proposed. In Chapter 4, the Proposal System is explained, formed by three main components: the descriptive, predictive and prescriptive subsystems. At Chapter 5, in the Methodology section, the process from the data acquisition to getting the main outputs of every subsystem is explained. Thereupon, the Results section, Chapter 6, will be structured in a similar manner. To start with, the main outcomes on the descriptive analysis will be shown. Later we will show the results of the predictive algorithm by taking some different metrics and for finishing, some examples of combining these two parts will be made, as the prescriptive system results' section.

Afterwards, some Future Lines will be set in order to continue with different research lines, Chapter 7, and finally the Conclusions will be at Chapter 8.

2. STATE OF ART

This section has the aim of showing the main instruments currently used to distribute the available resources to fight forest fires along Galician geography and showing some technological solutions given to fight fires in Galicia and internationally. First, an in-depth explanation about how IRDI works, how it is calculated and its effectivity will be shown and, subsequently, some technological proposals against fires are presented, some of them related with Big Data and Machine Learning.

As mentioned before, nowadays, the main tool for coordinating and managing all personal and material means is the **IRDI**. This map shows the risk level of fire ignition for very small areas (smaller than municipalities) by means of a five levels classification. This rate is based on the Forest Fire Weather Index (FWI), also known as Canadian rate.

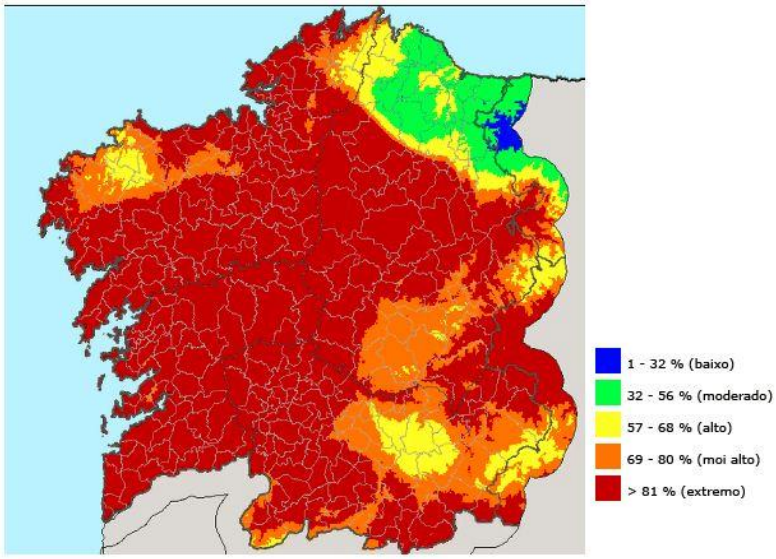


Figure 8: IRDI map example.
(Source: http://mediorural.xunta.es/areas/forestal/incendios_forestais/irdi/)

An example of the IRDI map decreed for one day is shown in Figure 8. On it, the risk level for every part of Galicia can be appreciated, where the red color means an extreme danger (>81%), the orange means a very high risk (69-80%), the yellow means high risk (57-68%), the green means a moderate risk (32-56%) while the blue stands for indicating a low risk (1-32%).

For getting these different levels, the system is based on the conjunction of the following inputs: daily temperature observations, relative humidity and wind. From this inputs, and some other correction factors such as the month and hour of the moment when the measures were taken, or the pending, sun exposition and the type of combustible in that area, they calculate the rate level. The meteorological indicators needed are extracted from fixed meteorological stations or from portable weather stations.

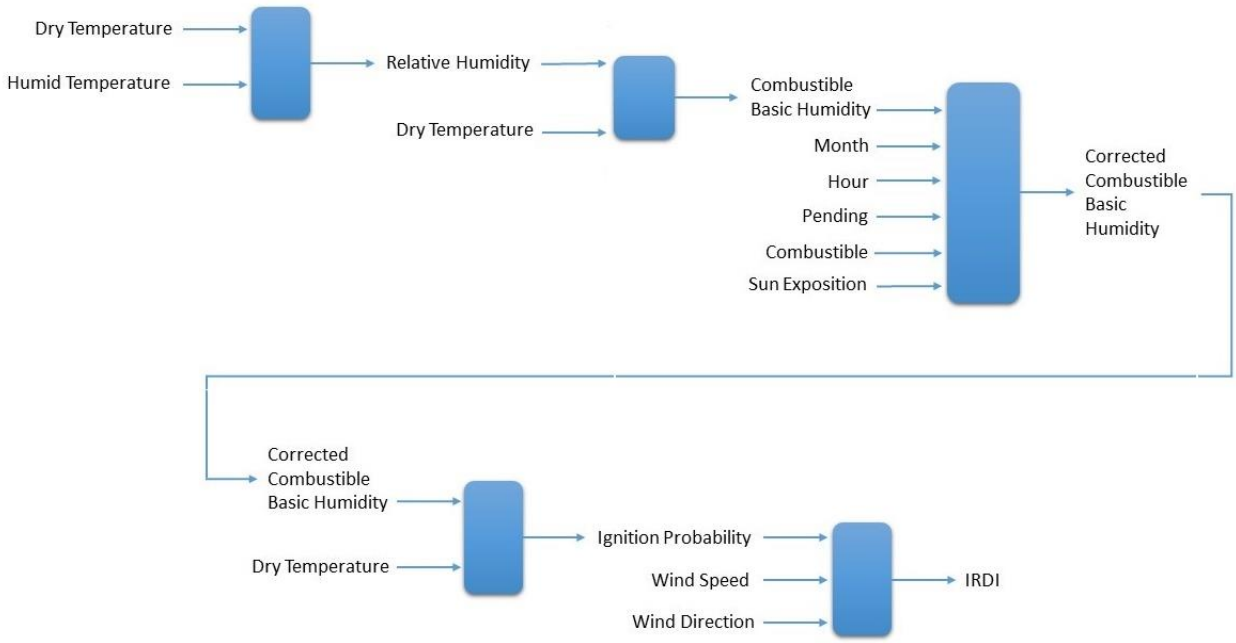


Figure 9: IRDI process calculation.

The process followed for calculating this risk level is the following: first of all, the temperature is measured with both the thermometer dry and humid and, from the difference of these temperatures, the relative humidity can be gotten at that exact point. Once this parameter is calculated and in combination with the temperature gotten by the dry

thermometer, the dead fine combustible basic humidity value is obtained. This value has to be modified by a correction factor that is gotten for this result when combining it whit the month, hour, type of combustible, sun exposition and pending of the moment and exact situation where the measures were taken. From this new indicator, if combined again with the dry thermometer temperature, the ignition probability is given. Finally, by combining this probability with the wind and speed direction, the risk level for that day and place is gotten. This whole process is graphically explained in Figure 9.

Independently to IRDI, some High Risk Areas or ZAR (Zona de Alto Risco) are set. These small areas are set depending on the mean of fires taking place in previous 10 years in those zones. For these extreme cases, more intense surveillance actions will be conducted and some harder restrictions and preventive measures will be carried through. It should be remarked that these mechanisms are working independently one from each other.

IRDI and ZAR are just one small part of the firefighting program proposed by the Galician government every year. The whole firefighting plan is detailed in a document called *Pladiga*, that acts like a road map in order to achieve the goals set by themselves. In this document, main objectives are set, an explanation on how to calculate the risk rate is made, they organize the firefighters structure, designate functions and missions for every entity and create a prevention, detection, dissuasion and extinction plan.

Even if an enormous effort is being made due to reduce the number of fires, it seems that something is not working properly, as long as the number of fires produced is still very high. Some studies such as one exposed in the 6th Spanish Forest Congress show that the IRDI is a bad ignition rate indicator because most of the fires are started when low risk levels are present [7]. This affirmation can be contrasted when paying attention to Figure 10.

On it, the total number of fires started in last years are classified depending on the existing level alert when the fire took place, going the different levels from 1 to 5 and being 5 the highest. In the figure, it can be seen that the majority of the fires are starting most of the times with a level alert of 1 or 2. In that study, it is also shown that this indicator is the worst from the seven analyzed from different countries. This said, it has to be mentioned again that Galician's fires are more unpredictable because of mankind responsibility, so the indicator

may not be as bad as this study shows, since the conditions may make any other indicator worse than it really is, if used in this region.

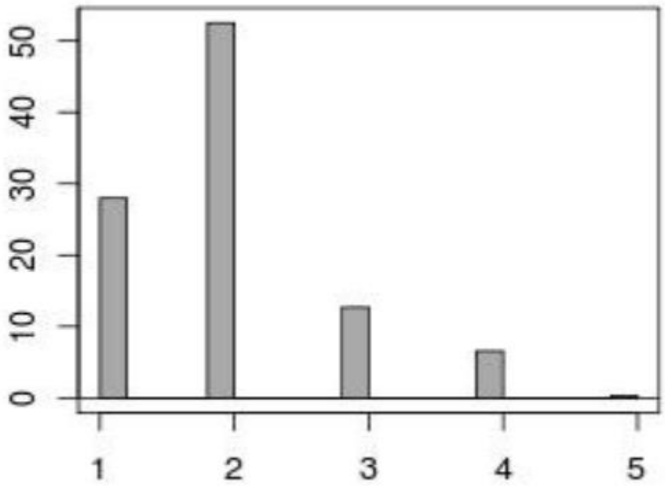


Figure 10: percentage of fires depending on the IRDI level.
(Source: Marey-Pérez, M.F., Rios-Pena, L., Franco-Vázquez, L. *Metodología para la validación de los diferentes índices meteorológicos de riesgo de incendio para Galicia*.)

Apart from these traditional techniques, some other modern methods are starting to be looking for reducing fires in Galicia. More and more people are aware of this big problem and a lot of researches are taking place in some different ambits.

Concerning Machine Learning in Galicia, a study was made by the Superior Council of Scientific Research (CSIC) in Santiago, where they have made an algorithm to predict the starting of the fires season by relating fires records and weather conditions in Galicia [8]. At this project, developed in 2010, they were able to predict the day when the intensive periods of fires would start (one in winter, and another in summer), 3 months in advance, and just with a medium error of about 3 days. For this, they had to use a British meteorological prediction system (there is not a prediction system in Galicia for such a long period) and fires records from the previous 25 years. This algorithm provided really important information while planning the prevention systems, and could save a lot of money, since, for example, having helicopters available is very expensive, so adjusting the day could be critical in this sense.

At national and international levels, it also has been an important advance in using these tools for this purpose. A lot of projects involving Machine Learning and firefighting were

developed all over the world. To mention some of them, Prometheus is a spatially explicit fire growth simulation model that provides operational and strategic assessments of potential fire behaviors over time and space (Figure 11). It was implemented and validated by the Canadian government of Alberta with good results [9]. Another example of using Big Data for this purpose is a project developed by the collaboration of many universities from the United States in which they simulated fire patterns in different landscapes, so they could stablish different patterns differing on the type of field the fire was taking place [10]. Nowadays there is an important field in computational science called wildfire modeling that has the aim of predicting fires behaviors.

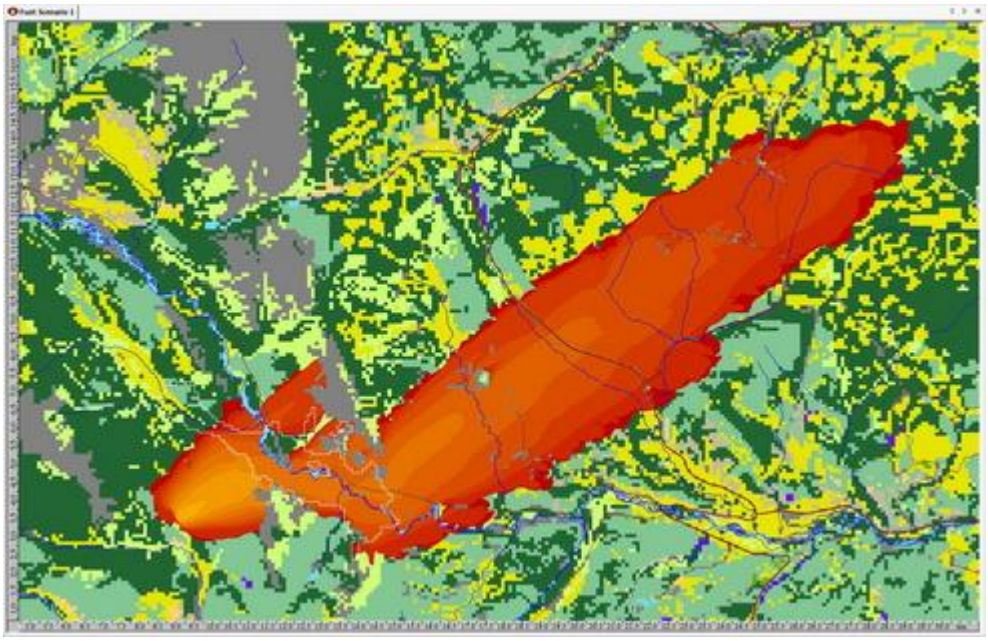


Figure 11: Prometheus example simulation.
(Source: http://www.firegrowthmodel.ca/)

A more similar project to the one exposed in this document is one developed by the University of Minho (in north Portugal) where they tried to predict the size of the fires taking place for a given day, given the weather conditions on that day. For that purpose, they used Data Mining techniques achieving acceptable results [11]. Another similar project is the one realized in the Sydney area were experts tried to predict the probability of large-fire (>1000 hectares) ignition days, by examining historical records and combining them with relative influences of the ambient and drought components [12].

Apart from Big Data and Machine Learning techniques, researches are looking to fight wildfires with other different technologies. A remarkable example, created in Galicia, is Integra WildFire [13], an advanced system to detect early fires. Its functional principle is as simple as emitting a light beam, and waiting for a response coming from a smoke reflection. This device is sensible to the dispersion produced by this kind of reflection, so it can detect very early fires up to distances around 5km. The prototype is shown in Figure 12.

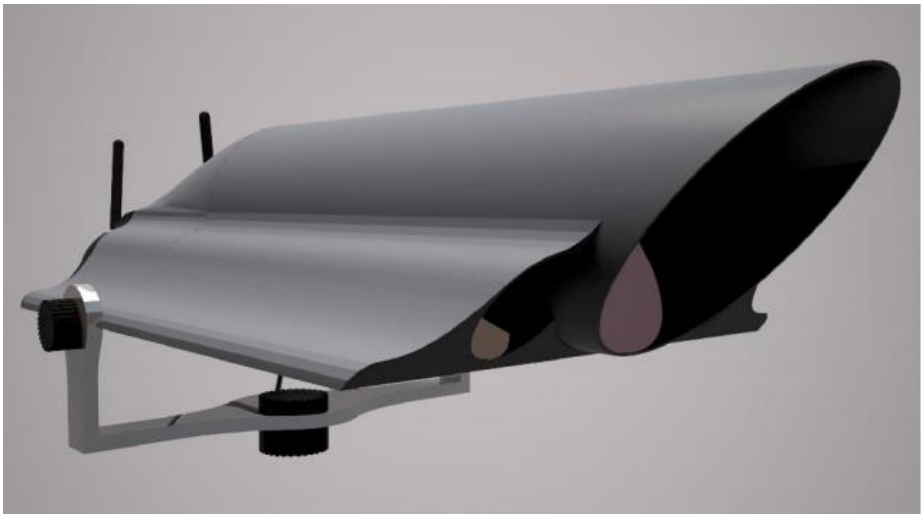


Figure 12: Integra WildFire prototype.
(Source: http://www.integraciones.com//)

Another important example is the FUEGO system developed at Berkeley University, consisting of a high resolution satellite being able to take pictures of the Western U.S. every few seconds in search of hot spots that could be newly ignited fires [14]. By detecting the fire in its initial phase, they would reduce in an enormous percentage the burning area of every fire, but for the moment, the United States government has not implemented the system. Other examples are using drones to collect information about an existing fire and providing Internet connection on the fire place [15], or even some experiments are being conducted where a device blasts fires with compressed air and water [16].

The list goes on and on, but most of these projects look for detecting or extinguishing, but just few of them to prevention. Plus, **public administration does not seem to invert on these kind of new technologies**, wasting all the money as they traditionally have done.

As a conclusion to this section, it should be said that IRDI is not an optimal solution to fight fires in Galicia. Nevertheless, a lot of projects are being developed, in Galicia and outside, regarding Machine Learning and other technologies, and some of them could serve as a very good complement that lead to an important decrease on the number of fires in Galicia. Besides, it would be crucial that the government supports projects like this, since they are the ones that coordinate and manage the firefighting plan.

3. TOOLS

In this section all the basic concepts that were important to develop the project are explained, as well as the main tools used. This is, R programming language, Machine Learning, Logistic Regression, and all the previous study that has been made in order to learn all the necessary knowledge before getting down to business.

3.1. R Programming Language

R language is a programming language focused in graphical and statistical analysis. It is an evolution of the S language, created in 1976. By this year, all the statistical computation was made by Fortran subroutines, which John Chambers, Rick Becker and Allan Wilks, belonging to Bell Labs, found really tedious. So, they decided to build their own Fortran macro libraries that would lead to the creation of their intern language: "Statistical", that would end to be called by its initial "S", before even reaching out from Bell laboratories as a distributable product. In 1988, "S" was rewritten completely into "C" code, becoming a similar version than the one we have nowadays. In 1998, the 4th version "S4" was liberated with a more object-oriented programming. In this same year, "S" won the "Association for Computing Machinery's Software System Award", one of the most precious award in the Computer Science field.

Along those years, "S" license changed hands a lot of times and suffered a lot of company fusions and divorces. Even with this, the basis of the language was strong and it did not suffer too many modifications, maintaining the initial fundamentals. While "S" was changing owners and denominations, Ross Ihaka y Robert Gentleman decided, in 1999, to implement their own dialect that was called "R". After two years it was released under a GPL (General

Public License) and this decision is probably the one that made R language so important nowadays. The current logo of this program language is shown in Figure 13.

Figure 13: R logo.
(Source: https://developer.r-project.org/Logo/)

To this day, the human being is producing and collecting billions of data thanks to big advances in technology. Biology, forecasting, medicine… the fields where data science can be applied to is really broad. Also it is being really important the boom we are living concerning smart cities and social networks were a lot of information need to be analyzed.

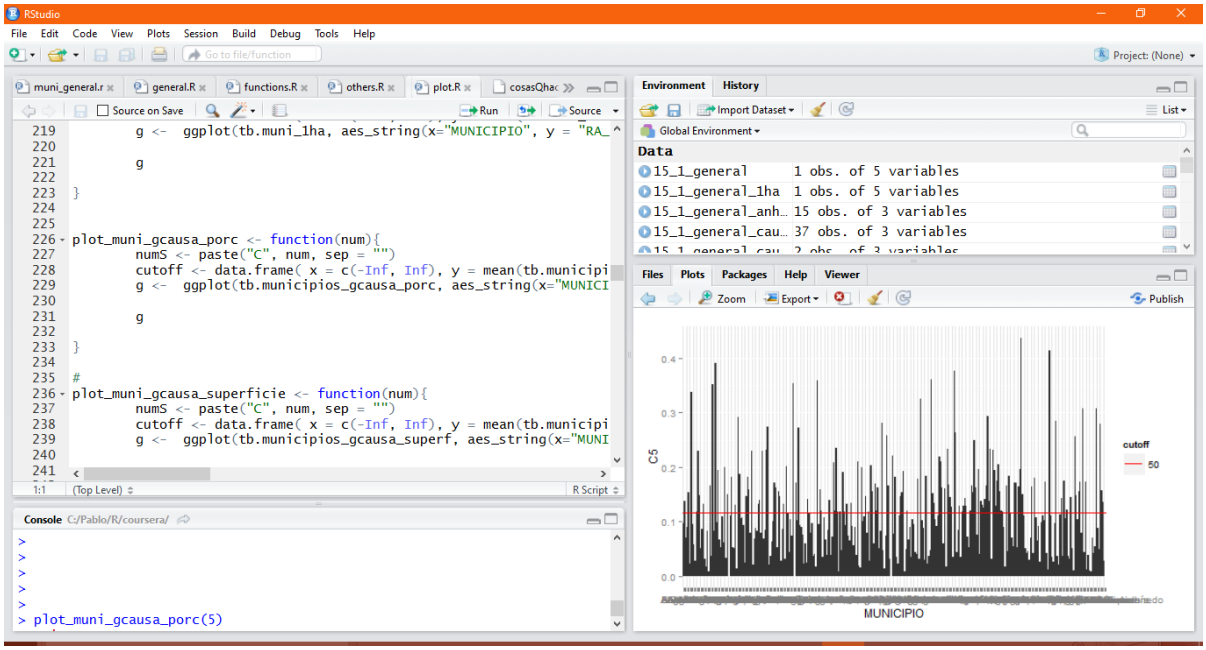


Figure 14: RStudio interface.

R language is one of the preferred tools used in order to analyze all this information due to its statistical and calculus orientation. Also, R is an interpreted language, this is, the language is interpreted by a virtual machine that is able to understand and execute the code, making it faster than non-interpreted languages. Moreover, R can be integrated with different databases

and other programming languages like Perl, Python, Ruby or Java and it provides a very strong graphics creator. Plus, it can be used through a free, powerful and intuitive interface called RStudio (Figure 14).

As mentioned before R is open source and it can run in most OS such as UNIX, Windows or MacOS. Just for finishing, mentioning that what is making R bigger and bigger nowadays is the enormous community holding it by adding new libraries (CRAN repository), documentation and solving doubts.

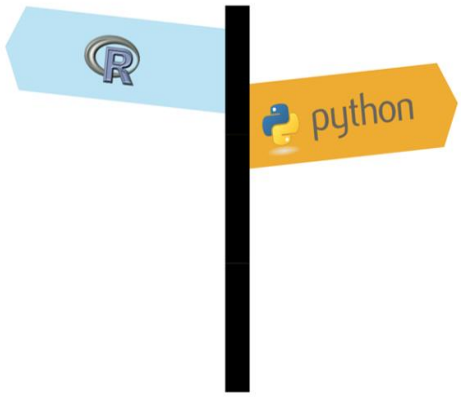


Figure 15: Two paths for Machine Learning programming.
(Source: http://analyticstraining.com/wp-content/uploads/2015/07/R-vs-Python1.png)

The most important rival to R language is Python (Figure 15). This is a general purpose programming language that is well known by its flexibility, readability and simplicity. Even if the last language has also a huge community behind it, it is more scattered for being a general purpose language. R shows a more powerful visualization engine (Figure 16) and it may be better for such a project with a really important statistical component. For these reasons, R language was chosen for the aim of this project even if it is not as easy to learn as Python.

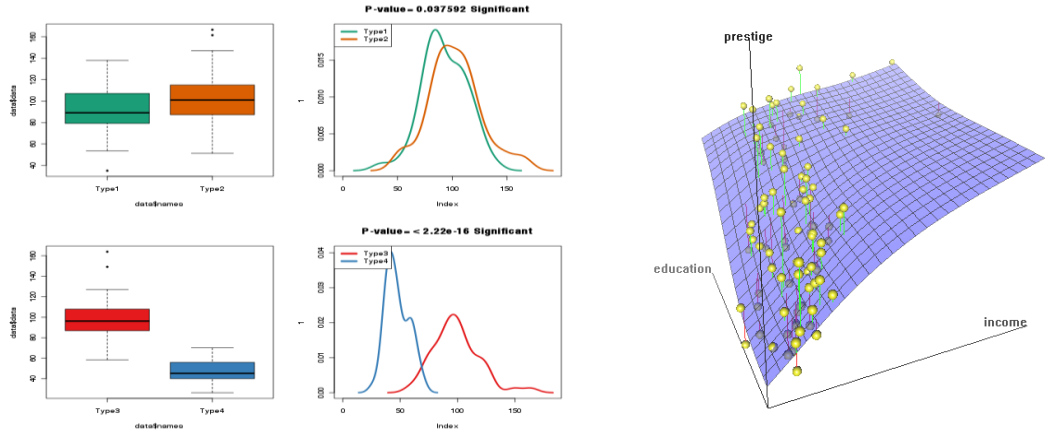


Figure 16: Plot examples produced by R language.
(Source: http://socserv.socsci.mcmaster.ca/jfox/Courses/R/ICPSR/)

3.2. Machine Learning

Machine Learning is an Artificial Intelligence's branch that has the aim of developing techniques that can help computers to learn by their own. More specifically, by creating algorithms capable of generalizing behaviors and recognizing patterns from some information provided as examples. So, Machine Learning can be seen as a process of inductive learning, since it leads to obtain generalizations from particular cases. A figure symbolizing the Machine Learning concept is shown below.

At a basic level, it could be said that Machine Learning tasks try to extract knowledge from some properties not observed in an object, based on the properties that have been observed from the same or similar objects. This science tries to predict future behaviors from what have happened in the past. For example, Machine Learning algorithms try to predict if a client will like some new products based on what the client used to like or buy.

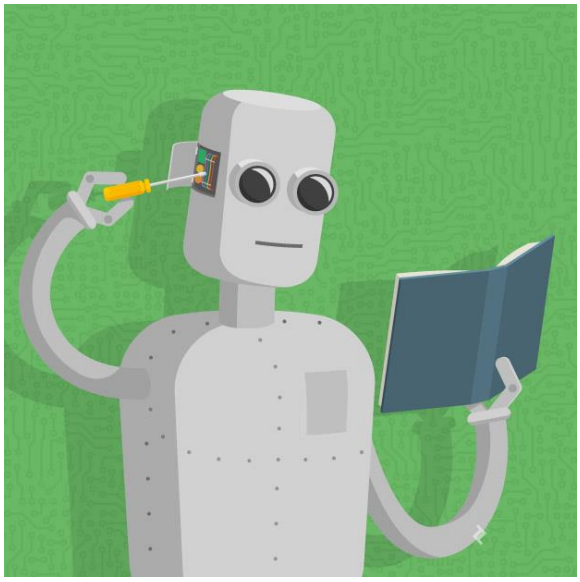


Figure 17: Machine Learning concept.
(Source: http://assets.toptal.io/uploads/blog/image/443/toptal-blog-image-1407508081138.png)

There are a lot of different types of problems that can be solved through inductive learning, where the main difference from one another is the kind of object that is being tried to predict:

- Regression: it tries to predict real values. For example, the price of a jewel given the weight, based on the prices than other jewels of different weights were taken.

- Classification: it tries to predict the classification of an object among a set of prefixed classes. If the set of classes consists on just two classes, we can talk about binary classification. If three or more, multiclass classification. An example of a classification problem would be to organize some photos by image recognition into animal groups: dogs, cats, elephants and cows, for example. This case would be a multiclass classification.

- Ranking: it tries to predict the optimal sorting in a set of objects following a predefined relevance order. For instance, giving the order in which an Internet search engine returns some resources.

On the other hand, and depending on how the examples are provided to the algorithm, there is another Machine Learning classification into two main groups: supervised and unsupervised learning:

- Supervised Learning: a function is generated stablishing a relation between inputs and outputs, where examples are labeled a priori, this is, the algorithm knows the classification of the examples given for acquiring the knowledge. The example showed in the regression explanation would be also supervised.

- Unsupervised Learning: in this case, the modeling process is made from a set of examples formed only by inputs to the system, without knowing the correct classification. So now, the algorithm needs to be able to recognize patterns by itself in order to label new inputs. The example about animals mentioned as a classification problem could be also unsupervised.

The problem presented in the prescriptive system of this project will correspond to a **binary classifier**, since it will try to classify all the examples given into two groups: fire or not fire, and it belongs to the **supervised learning group**, since it will learn from examples where the output of the system is previously known. For solving this kind of problems, one of the best approach is using a **Logistic Regression Model**, that will be explained in the next section.

We talked in this section about Machine Learning and not about Big Data, since this last concept only applies when working with data stored in clusters of servers, and in this project, they were not used. However, the methodology could be extrapolated if working with Big Data.

3.3. Logistic Regression Model

In statistical modeling, regression analysis is a statistical process for estimating the relationship among variables. It helps to understand how the typical value of a dependent variable changes when any of the independent variables is varied, while the other independent variables are fixed. In Figure 18, an example of a regression analysis with just a dependent and an independent variable is given.

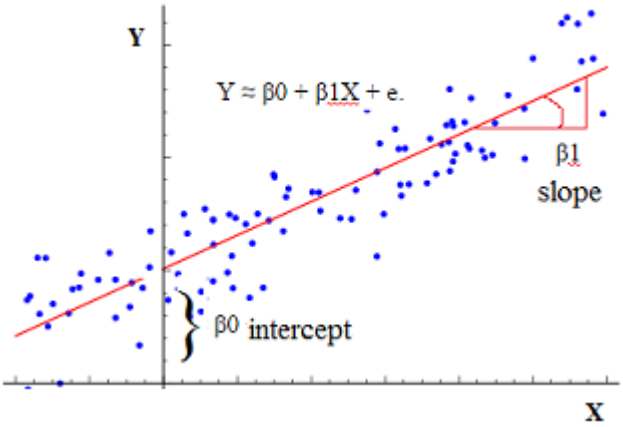


Figure 18: Regression Analysis example.
(Source: www.analyticbridge.com)

Logistic Regression, is a regression model where the dependent variable is categorical. The binary logistic model is used to estimate the probability of a binary response based on one or more predictors (or independent) variables. The logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function (Figure 19), which is the cumulative logistic distribution. Besides, it uses a standard normal distribution for errors.

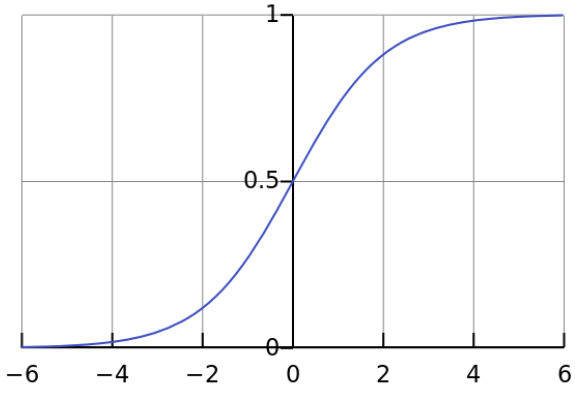


Figure 19: Standard Logistic Curve.
(Source: https://en.wikipedia.org/wiki/File:Logistic-curve.svg)

This model can be framed at the Generalized Linear Models (GLM) group. However, the model of logistic regression, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular, the key differences of these two models can be seen in the following two features of logistic

regression. First, the conditional distribution y | x is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes [17].

3.4. Previous Studies

Previously and simultaneously to this project development, some information about R programming, Machine Learning and forest fires in general and in Galicia was acquired. All of these concepts were unknown to me before starting, but as long as they seemed really interesting and innovative, this project served as a great opportunity to learn about these areas of study.

For starting, some researches were made in the **forest fires area**. Main factors behind forest fires generation and propagation should be learnt for fires in general, but also only for Galicia's case, since it is a very particular region in this sense. At that same time, some interviews were conducted with some expert firefighters in order to get the necessary knowledge about the current situation and understand their needs. Also the analysis of the state of the art concerning traditional and technological firefighting techniques in Galicia and internationally should be mentioned as previous study.

While doing this state of the art analysis, even before getting to this Machine Learning solution, some other ideas at different fields were explored in order to get a good solution for the forest problem in Galicia. One of those first ideas was a low power sensor network implantation distributed along the Galician forest in order to monitor temperature or chemical changes that may be indicative of a starting fire. This possibility was ruled out because the number of sensors required to cover all the Galician surface was excessive. Even so, it could lead to a good solution if combined with this project, since with the data extracted from this thesis, these sensor networks could be placed in specific areas, in forests where fires are too common.

In the end, the **Big Data** option gathered momentum after some interviews with both local brigade chiefs and firefighters, and when it was discovered the whole amount of data that could be analyzed (Figure 20).

At that moment some deeper concepts about fire factor risks, fire propagation, distribution and causes were learnt in order to study the feasibility of the project and finally it was decided that Machine Learning techniques could offer a good solution to this problem.

Figure 20: Big Data idea coming after interviews and discovering all the available data.

In order to learn about R programming and Machine Learning, a set of courses was taken in Coursera: the "**Data Science Specialization**" course (Figure 21). This program, conducted by leader teachers in Johns Hopkins University (Baltimore, Maryland), is in fact a set of nine courses that offers a great introduction to the Data Science world. Just seven from the nine courses were needed in order to achieve some good bases to cover the objectives of this project.

First, some basic ideas about the field were given, as well as some conceptual knowledge but without coding a single line of code. With this clear, in the 2nd course, some fundamentals on R language were given. The 3rd course was about getting data from different types of sources, like the web, APIs or databases, cleaning this data in order to extract the valuable information from raw data and share it. 4th course covered the essential exploratory techniques for

summarizing data and constructing graphics for making information clearer. 5th course presented the fundamentals of statistical inference in a practical approach and some of its theories, including Bayesian statistics, likelihood ratios, variances study, and asymptotic statistics. In 6th course, linear models, predictors and regression models were presented, including the Logistic Regression Model. Also residual data and variability would be analyzed. Last but not least, the 7th course was about prediction and machine learning, providing some basic concepts on training and test sets, overfitting and error rates. Also some inductive learning methods such as Classification Trees, Naive Bayes, Random Forest or Forecasting were presented in this last course.

Figure 21: Data Science Specialization course.
(Source: Jennifer Brendle)

It should be mentioned that the learning process achieved by taken these courses was carried out concurrently to this project, during the first three months of it, in such a way that the accomplishment of the courses was providing the necessary knowledge for developing the project at the correct order. That is the reason why it is considered as previous learning.

4. PROPOSED SYSTEM

In this section, the conceived system in order to achieve the goals set in the Introduction chapter is shown. The system described in this project is a complex one formed by three main pieces. The first one is the descriptive system. It will give a general overview about the forest fires problem in Galicia and show some differences at the temporal, causal and geographic distribution of fires, as well as a local vision for every single municipality. The second part is the predictive system, that has the aim of predicting whether it is going to be a fire in a given municipality and day. By merging these two subsystems, the prescriptive will consist on a deeper analysis that will lead to specific measures and to set an alert level.

4.1. Descriptive System

The objective of this subsystem is to **provide an understanding on the distribution of Galician forest fires at a general and at a municipality level** that can lead to a more efficient distribution of resources. So, the descriptive section holds two different parts: the general and the local, where the first is necessary to get the second. In both of them, geographical, temporal and causal studies are provided. In Figure 22, a scheme of the different analysis proposed for these two part is presented.

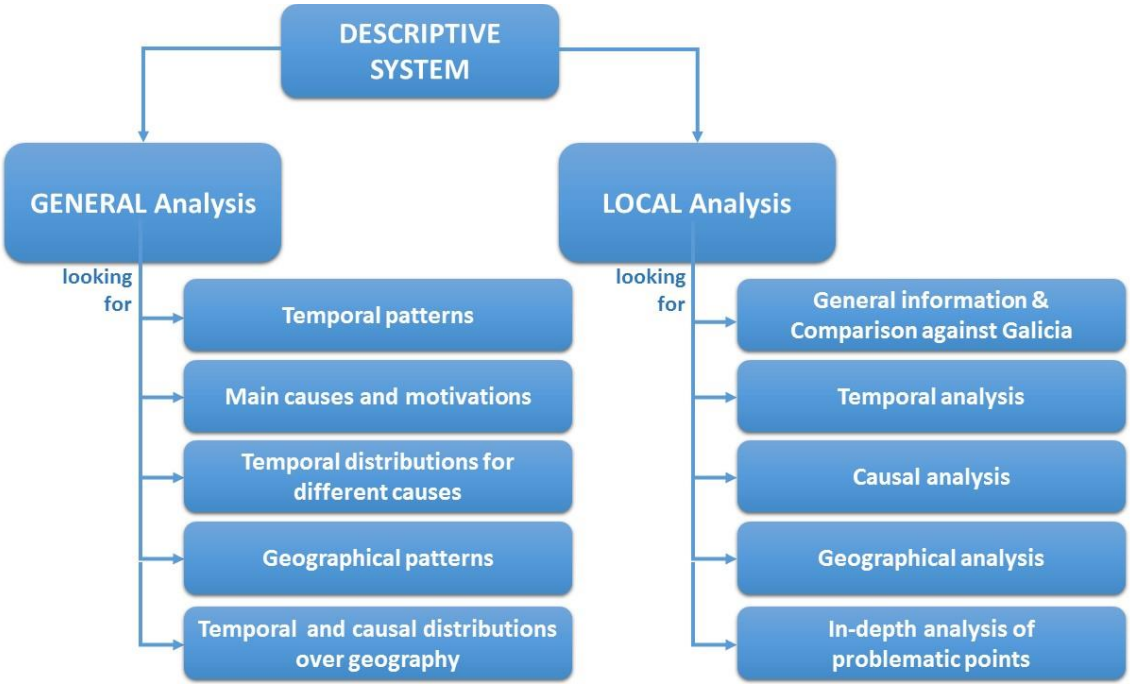


Figure 22: Descriptive System Diagram.

At the **general analysis**, some studies will be performed in order to model this apparently randomness that forest fires in Galicia seem to provide. For this, at the beginning, some patterns along time will be searched. This temporal study seeks to find the most dangerous months of the year, or the hours of the day that focus more fires. For instance, with this analysis it could be appreciated the distribution for all fires along the months. After that, the same will be made but differing by causes, so the main causes, and motivations behind arsons can be found at a general level (Figure 23).

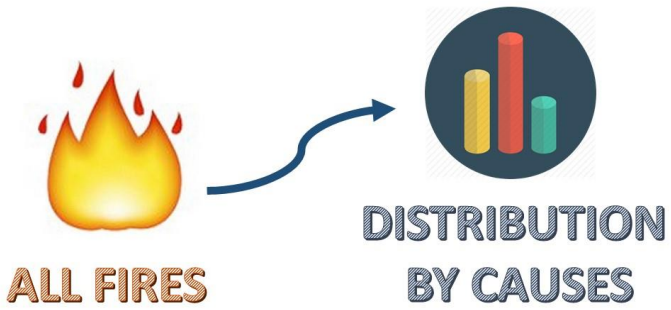


Figure 23: Descriptive System dividing all the fires by causes.

This process will continue by looking for differences at the temporal distribution of different causes, and also by comparing causes and weather conditions. As an example of this case, this analysis could serve to appreciate how just the natural fires (instead of all) were distributed along the different months of the year (Figure 24).

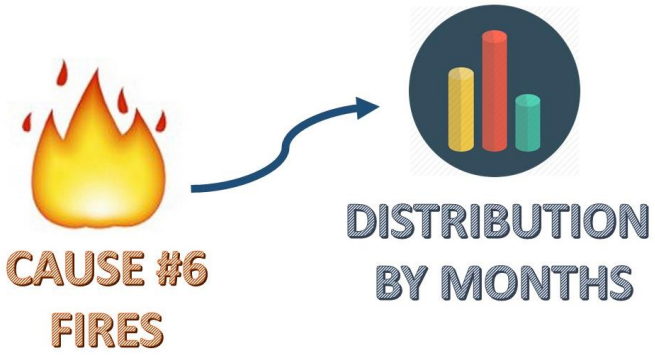


Figure 24: Descriptive System dividing just the natural fires by months.

The descriptive system also tries to find differences between weather conditions such as maximum temperature, relative humidity and velocity and wind speed from cause to cause (Figure 25).

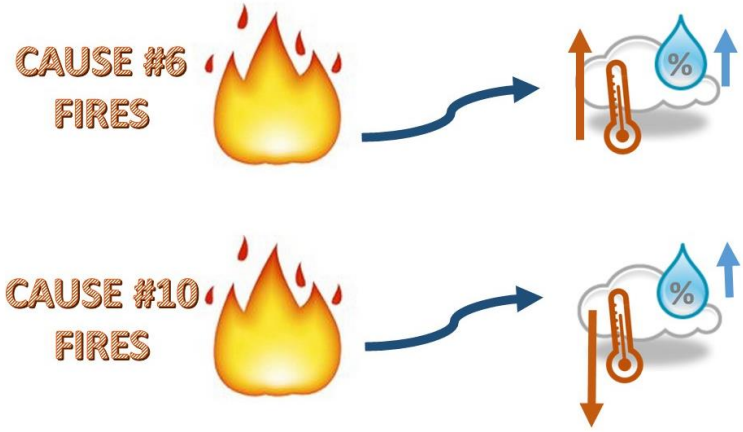


Figure 25: Descriptive System comparing general weather conditions for different causes.

At a geographic dimension, the descriptive analysis can also show the general distribution of fires differing by location. With this analysis, it could be find which places were hotter than the others in terms of forest fires activity in general, and also discriminating by the initial parameters set: hour (Figure 26), weekday, month, motivation, cause… By doing so, some

geographical patterns will appear, distinguishing, for example, in which areas of Galicia occur more fires due to stubble-burning or in which, fires were more abundant at certain hours or in certain months. This process is also able to distinguish small fires from big fires and place them on the map depending on the different parameters.

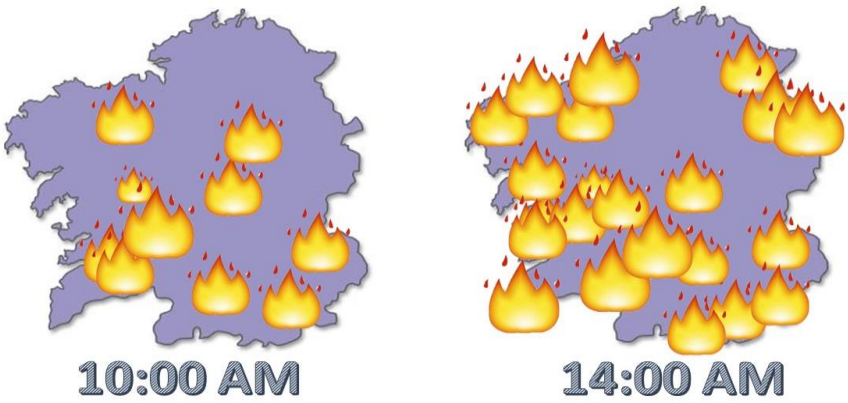


Figure 26: Descriptive System is able to show the geographic distribution of fires attending to one criteria.

As the surface of every municipality is known, the descriptive system also performs a study by calculating two important parameters: the relationship between the number of fires in a municipality and its surface and also between the number of burned hectares and the surface. With this analysis, it could be find which are the most problematic municipalities in general terms, and also by studying every parameter mentioned before (Figure 27). For example, it could be known which are the hottest municipalities in November.

Figure 27: Descriptive System is able to find which municipalities are more problematic at one motivation.

By doing this, **very important variances into the temporal, causal or geographic distribution of forest fires in Galicia will be discovered**, and would open the possibility of reaching a better distribution of resources from a high perspective level.

Last formal analysis in this section will consist on making an analysis like the general one made to analyze the main problems in Galicia but applied to every single municipality, one by one. By doing this, the most important part on the descriptive phase will be gotten; a report showing forest fires trends for every municipality.

At a lower level, creating this **report for every municipality** would be also very beneficial for them, since this **information** is, until now, **totally unknown to local authorities and** what is even more important, **to firefighters** patrolling, surveilling or battling fires in the area. Also, the big dispersion on the distribution of fires from one municipal term to another, would justify the need of creating this report.

Figure 28: A report for every municipality with brand new information.
(Source: http://www.ospox.com/image/attendance%20report.png)

This detailed report (Figure 28) will describe the main problems affecting forest fires in every municipal term. It will show a comparison of the situation on the municipality with the rest of Galicia, so they are able to know if the situation is very worrying or not, and also a temporal, causal and a geographical analysis of the fires distribution in their term, so they know where their strengths and weak points are. Also, a more detailed analysis is offered by performing a temporal analysis of the most problematic causes on the area.

4.2. Predictive System

The predictive part's aim, as it was said in the introduction of this section, will consist on building a Machine Learning algorithm based on a Logistic Regression Model, which tries to predict whether a wildfire will occur in a municipal term in a given day or not.

This system will be composed by a set of **weather and temporal information inputs** and a **binary output** indicating a fire occurrence or not. The inputs for the algorithm will be, for every day, the maximum temperature, relative humidity, wind speed and direction and rains expected. Also the month and the number of week are taken into account as parameters (Figure 29). It will work individually for every municipality, so it will run with examples with data from the closest meteorological station and the occurrences of fires in the past in that municipal term.

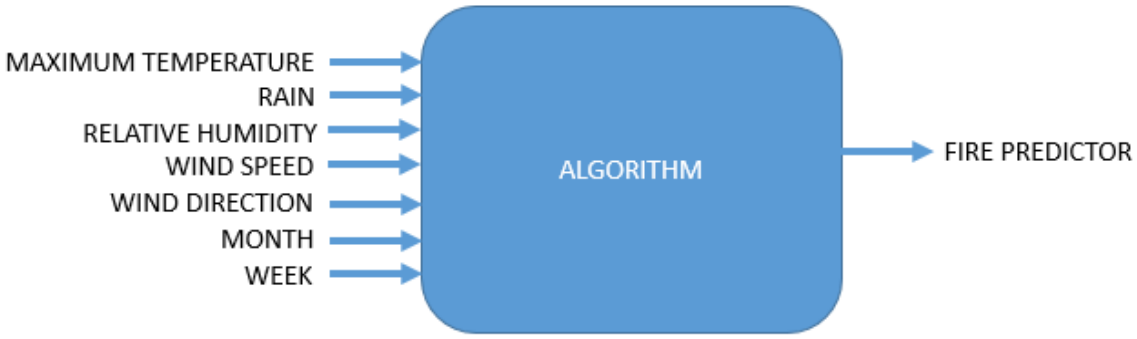


Figure 29: Predictive Algorithm Diagram.

First of all, it is important to define the problem that is wanted to solve into the groups mentioned in the Tools Section. In this case, it corresponds to the binary classification (it has to decide and classify the different inputs into two different groups, fire or not fire), and supervised group (the examples given for learning will indicate, if for those inputs, whether a fire occurred or not). The algorithm responsible for doing this, will learn on the basis of historical fires records occurred on that area using a **Logistic Regression Model** (ideal for binary classification problems).

The algorithm works following a **2-phase model**. In the first one, the algorithm needs to be trained with different input variables and the logical output taken for those inputs in the past

(0 or 1, or Fire or Not Fire in this case). The algorithm must **learn** from the resulted output of these examples given. By its own experience, in the second phase, the algorithm must be able to **predict** a logical output for further new values of the same input variables. So, it is important to notice that the input values in the training phase must be the same that afterwards are going to be used for predicting. For this reason, causal references cannot be introduced as inputs to the system, since it is impossible to know in advance which one will be the cause of a fire that have not taken place yet.

For testing purposes, every row in the table is split into two different groups: **training and testing** (Figure 30). This is an intelligent way in which the algorithm can learn from some examples; the randomly selected training set (around 70% of the rows), and apply the algorithm gotten to the remaining samples; the testing set. Since these samples also show an output, the algorithm can deduce it success rate by comparing its prediction to what had actually happened. This comparison technique produces a table called **Confusion Matrix**, which is an important metric in supervised Machine Learning algorithms. Another metric called **AUC** (Area Under the Roc Curve) will be used for giving more generic results.

Figure 30: Data set divided into Test and Training Set.

Since the data set was really unbalanced (many more examples for non-fire cases than for fire cases in a single municipality), a technique called **Random Over-Sampling** (ROS) had to be apply in order to balance the data set, and getting to better results.

4.3. Prescriptive System

The main idea behind the prescriptive system is to give some recommendations to local authorities and firefighters chiefs about how to use the descriptive and predictive systems in order to get to good measures. As we saw in the previous sections, in the descriptive system's part, it was showed that some important changes were produced if distinguishing by temporal, geographical and causal groups. After getting this knowledge, an in-depth report was given as the main output of the descriptive system for every municipality that offered key information to local authorities and firefighters.

The point here is that the report is offering just information, but this information should be analyzed in order to reach some conclusions that finally will lead to take some actions that may be beneficial for fighting forest fires in the municipality. In other words, one main objective of this section is to explain how to **convert the information from the report into real measures** that fight against fires (Figure 31). In order to show how this analysis should be done, in this section some examples showing how to do it, will be given, but it is important to notice that **what really makes this prescriptive system strong is the combination of this given information added to the local knowledge provided by local people**.

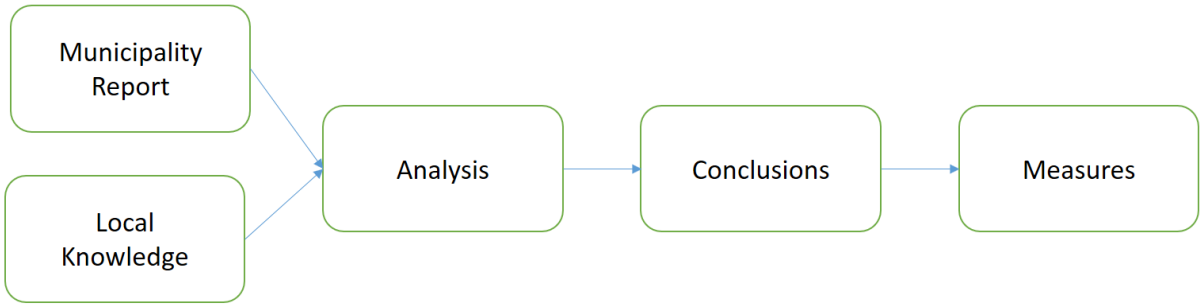


Figure 31: Prescriptive subsystem scheme: Measures.

On the other hand, the prescriptive system's main output is an algorithm that predicts with an acceptable success rate whether or not a fire will take place on a municipality on a given day. Even if acceptable, the algorithm is not 100% reliable, and some analysis should be made by firefighters by combining the result given by the algorithm also with the data obtained from the municipal report and with their own experience and knowledge of the area under their

control (Figure 32). By doing this exercise, and also paying attention to IRDI, they can **stablish a much more precise level alert** than if only guided by the algorithm.

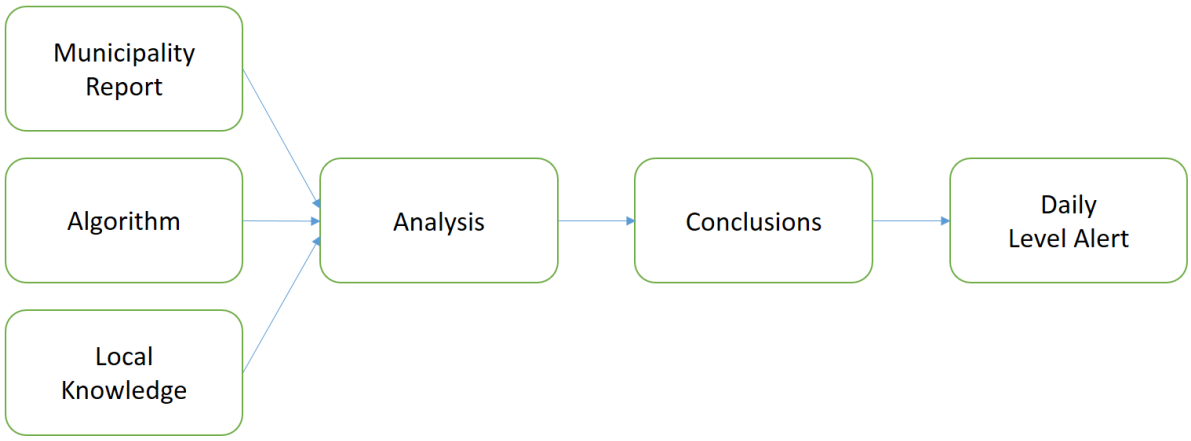


Figure 32: Prescriptive subsystem scheme: Daily Level Alert.

To sum up, the prescriptive system consists on a set of recommendations that authorities should follow in order to get from the municipality report to specific measures to fight against fires, and also to distribute their resources in a much more efficient way, since the most dangerous periods are now known at a local level. Besides, when helping with the algorithm, they could set a Daily Level Alert on its area.

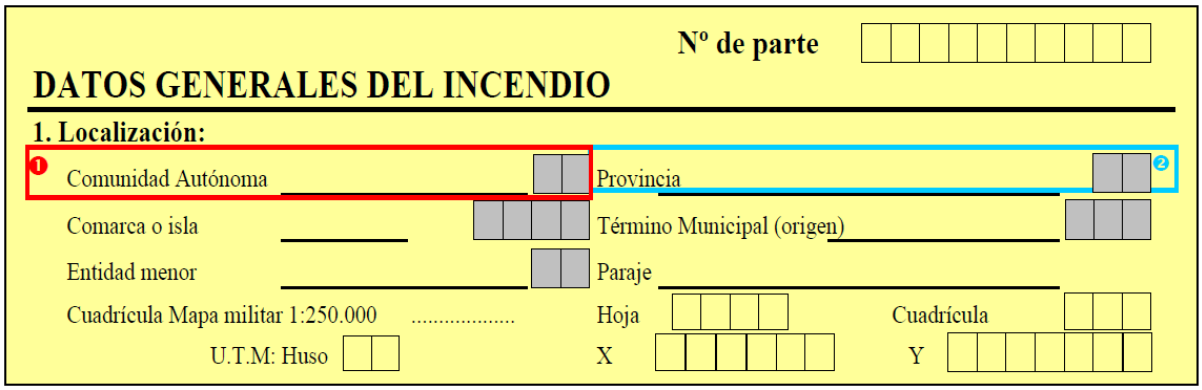
5. METHODOLOGY

In this chapter, the process followed in order to implement this project will be discussed, passing by the three main components of it: the descriptive, predictive and prescriptive systems. In addition, the data acquisition problem will be commented.

5.1. Data Acquisition

A really important part in a Machine Learning exercise like this, is data acquisition. This project uses a mix of data gotten by different means related to fires, weather conditions, and municipalities.

Every time a fire is extinguished in Spain, the firefight chief must fill a **form** with a lot of valuable information about the fire (Figure 33). The form gives information about the province, municipality and coordinates where it took place, hour, day, month and year of detection, arrival, control and extinction times, cause (natural, negligence, accidental, intentional, unknown or reproduced), and it even shows information about subgroups of causes inside these big groups, such as motivations in the case of intentional fires. Plus, it shows data about the total area burned, and more information that was not used for this project because it was considered with no value for this purpose or because it was not complete.



Nº de parte

DATOS GENERALES DEL INCENDIO

1. Localización:

Comunidad Autónoma

Provincia

Comarca o isla

Término Municipal (origen)

Entidad menor

Paraje

Cuadrícula Mapa militar 1:250.000 .....

Hoja

Cuadrícula

U.T.M: Huso

X

Y

Figure 33: Fire form extract.
(Source: Comité de Lucha contra Incendios Forestales, *Apuntes para la codificación en oficina del Parte de Incendio Forestal*)

Taking together all this valuable data from every single fire taking place in Spain, the Spanish government created an enormous database. This information is not public but it can be accessible by asking to the Ministry of Agriculture, Nature and Food. After demanding it, they provided us a Microsoft Access Database with all the fires started in Galicia from 2000 to 2014, both years included, listing information from **more than 99.000 forest fires** in the region (Figure 34).

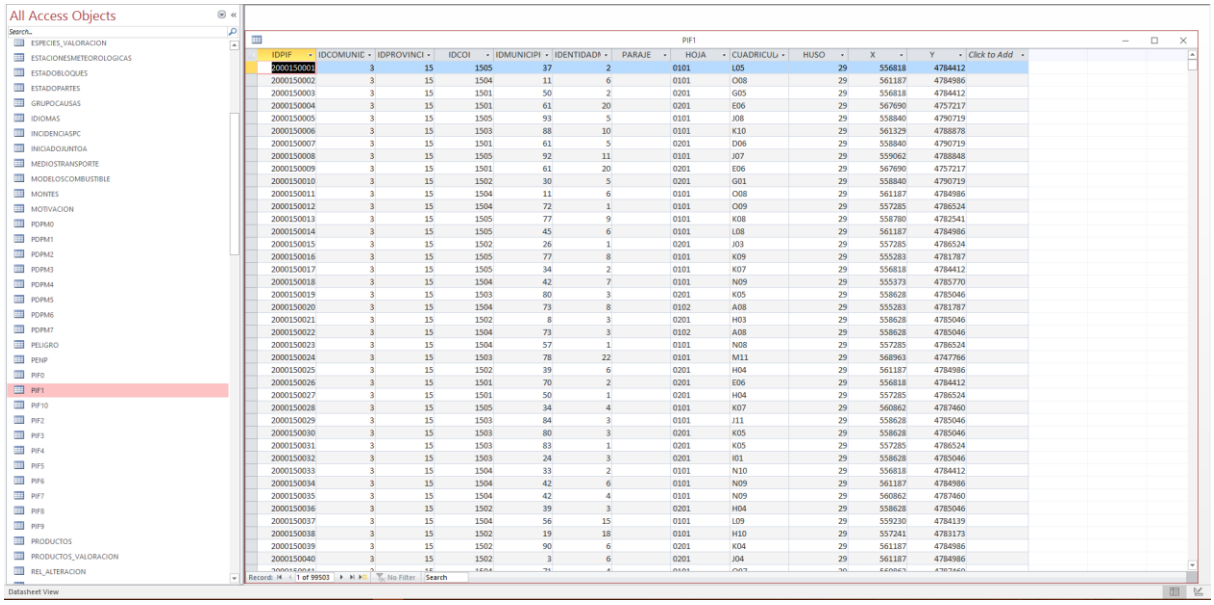


Figure 34: Access Database extract.

This raw data, had to be cleaned and processed, in order to extract just the valuable information. After a lot of sub-setting, sorting and transformations, **this raw data was converted into useful tables** thanks to R language.

In addition to those fires information, meteorological data was collected from a Galician Regional Ministry of Sustainable Development's website called www.meteogalicia.es (Figure 35). This is a **weather** observation and prediction site based on satellite images and meteorological stations displaced strategically along Galicia. They provide historical information from all meteorological stations they are using. For the purpose of this project, data has been collected from **ten stations**, selecting those that have been collecting information for a longer time, since 2000, and looking for a good geographical distribution.

From every station, daily values for maximum temperature, rain, relative humidity and speed and wind direction were collected. Even if the areas represented by each of these stations are really large, by assigning every municipality to one of them, we will get, at least, in a coarse sense, a representation of what was the weather like for every point in the map and every day.

Cod. Validación	Data	Parámetro (Unidades)	Valor
1	01/01/2014	Temperatura máxima (ºC)	12,2
1	02/01/2014	Temperatura máxima (ºC)	13,5
1	03/01/2014	Temperatura máxima (ºC)	11,8
1	04/01/2014	Temperatura máxima (ºC)	10,4
1	05/01/2014	Temperatura máxima (ºC)	12,5
1	06/01/2014	Temperatura máxima (ºC)	12,8
1	07/01/2014	Temperatura máxima (ºC)	11,6
1	08/01/2014	Temperatura máxima (ºC)	11,9
1	09/01/2014	Temperatura máxima (ºC)	9,4
1	10/01/2014	Temperatura máxima (ºC)	10,1

Figure 35: Meteogalicia webpage.

By working with this meteorological data, and combining it with the fires database, we get some important information as long as, up to this point, we can have an idea on the weather

conditions of a specifically area in a day where a fire may have started, how big it was and why it was caused. Last but not least, a database adding the size of every **province and municipality** was joined, so a relationship between the number of fires or burned hectares in a region and its area could be stablished. The combination of all this data is symbolized in Figure 36.

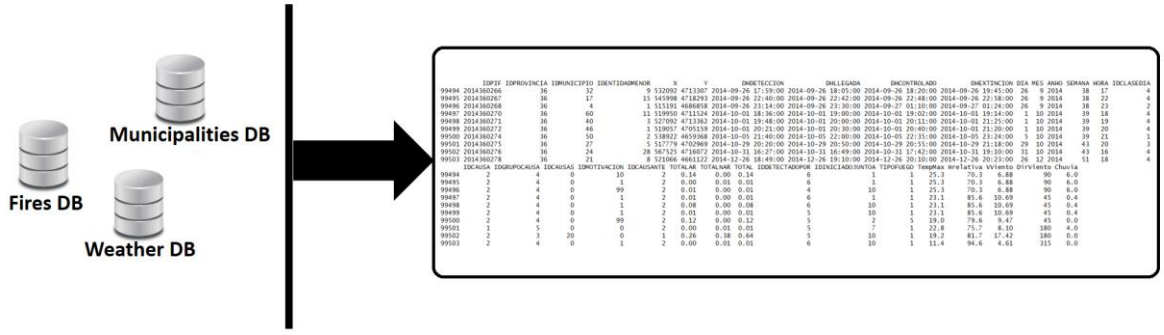


Figure 36: Databases integration process.

It has to be remarked that this data acquisition process has been one of the longest and hardest parts of this project as long as the data collected in different formats had to be handled (acquisition, subset and transformation) in a different way and combined with all the other data to create a full meaning from all the separate pieces.

5.2. Descriptive System

Once all the pieces were shaped as a unified data set, it was time to start to extract valuable information. First of all, a general exploratory analysis was made in order to obtain a first approximation of what is going on. Analyzing data taken from all the fires in Galicia would give us a general point of view and the first ideas about which are the most important problems. This process was done by drawing plots dividing all the fires into years, months, weeks, weekdays, hours, causes, motivations, combustible burned, firefighters action time, or geographically splitting them into provinces and municipalities. By doing this, some broad patterns or trends were discovered.

After this, a second approximation was made in order to get more interesting results. Now that the main problems were discovered, it would be easier to go in a deeper examination.

This second approach was made in a similar manner to the first one, but instead of taking all the fires, this time fires were split by causes and motivations and an individual analysis was made for every group of causes, specific causes and motivations. By doing this for every case, a lot of new patterns and trends were discovered. This process was also followed in order to relate individual causes or motivations with weather conditions such as maximum temperature, relative humidity and velocity and wind speed.

Next step made was to study the geographic distribution for all the different parameters of interest: years, causes, months, motivations…This phase of the descriptive analysis consisted on plotting a single Galician map for every of these different factors and study the differences from one to another. A similar analysis was made but splitting the fires into municipalities.

Last formal analysis in this section will consist on making an analysis like the general one made to analyze the main problems in Galicia but applied to every single municipality, one by one. By doing this, the most important part on the descriptive phase will be gotten; a report showing forest fires trends for every municipality.

5.3. Predictive System

The main objective of this section was to build an algorithm that predicts whether it is going to be a forest fire given a day and a municipality with an acceptable percentage of success. In order to reach this ambitious objective, the process followed is described below.

First step should be to choose some input variables for the system. This is not a simple decision and a hard study was made at the final phase of the project in order to get the most important variables, erase the ones not adding utile information and find the best possible model. But for starting, some variables were chosen and a table was created for every municipality. In it, every row shows temporal and meteorological information for a single day from 2000 to 2014, and of course, the output, indicating whether that day a fire occurred or not.

At the beginning, the results originated were really unsatisfactory, more or less the same as the produced by a random classifier. The first proposed algorithm was a multivariate linear

regression taking into account as variables, the maximum temperature, relative humidity, wind speed and direction. The linear model shows some limitations when used in classification problems (as this one is), so first step taken in order to get a better system was to replace this linear model with a model from the Generalized Linear Models family called Logistic Regression Model. This kind of analysis is designed for predicting the result of a categorical variable (that only can take a few values) in function of other predictor variables.

This milestone meant an increase in the success rate, but not such a big one as it was expected. At this point, the problem seemed to be another and after a careful search, it was discovered. The drawback was not given by the analysis used but by the fires distribution. In every municipality, the number of days with at least one fire was much lower than the number of days without fires and this is a problem to most learning algorithms. Most classifiers in supervised machine learning are designed to maximize the accuracy of their models. Thus, when learning from an unbalanced data sets (much more examples from a dominant output class), they are usually overwhelmed by the majority class examples.

Firstly, in order to compensate this effect, it was decided to use a technique known as Random Under-Sampling (RUS) consisting on deleting most of the rows where no fires were registered so the samples of both classes would be balanced. This process was not showing good results, since it was throwing away a lot of important data (almost everything in areas where a small number of fires were taking place). The solution that came with magnificent results was the one knowing as oversampling. Instead of throwing away some data, this technique consists on repeating the values of the minority group, with a final result of getting a balance set of samples between the two groups without ruling precious information out. This technique is known in the Machine Learning ambit as Random Over-Sampling (ROS).

Lastly, some more predictor variables were added as columns to pretty the final result up. In the end the predictors taking into account were the maximum temperature, relative humidity, wind speed and direction, rain, month and week of the given day.

5.4. Prescriptive System

The prescriptive system was realized by making an in-depth analysis of the statistics shown by the descriptive report of some municipalities, reaching to conclusions and finding possible solutions to the problem based on the temporal, causal and geographical distribution of the fires in the municipality.

6. RESULTS

The Results chapter offers the main outcomes and achievements gotten after the implementation of the system descripted in the Chapter 4, by following the methods explained in the previous section.

This chapter is also structured in three different categories: one showing the main results for the descriptive system, indicating intermediate and final results, another showing the results for the predictive algorithm created in the predictive section and, for the prescriptive case, some examples will be given in order to show how the procedure for taking good decisions should be done.

All the outcomes shown in this chapter are result of the analysis made with all the data mentioned in the Data Acquisition section at the Methodology chapter (all the fires registered in Galicia since 2000 to 2014) by performing an analysis in RStudio. It is impossible to embody all the plots in this document, but in order to justify these results, the most important ones will be shown in the document.

6.1. Descriptive System

In the descriptive system results section, the outcomes will be shown following the chronological order in which they were found, starting from the general findings needed to get a first overview the problem (temporal, geographical and causal) and finishing by the municipal terms analysis.

The number of fires, burned hectares, and the ratios between hectares burned for municipality surface, number of fires for municipality surface and hectares burned for every fire will be mentioned a lot in this section. From now on, they will be named by their initials, number of fires (**NF**), burned hectares (**BH**), hectares-surface ratio (**HSR**), number of fires-surface ratio (**FSR**) and hectares-fire ratio (**HFR**).

6.1.1. General Analysis

The general analysis section shows the main results obtained by making a deep exploratory data analysis over the data collected. A temporal, causal and geographic analysis of the fires distribution was made, from which a lot of differences were found by splitting the fires among these parameters.

6.1.1.1. Temporal General Analysis

In this section, a temporal analysis from all the fires produced in Galicia in last 15 years is shown, by studying how they are distributed by years, months, weeks, weekdays and hours.

The annual analysis showed that the trend in these last 8 years is that the NF has been reduced almost to the half with regard to the 8 previous years. However, the HFR has dangerously raised from 2010 to 2013 indicating more big fires than usual at previous years (2014 has been a very unusual year since a small number of fires has been produced on it). There is an enormous peak for BH in 2006 due to the enormous crisis suffered in Galicia in August that year. In the figure below, the BH for every month (columns from January to December) and year (rows) can be seen.

```
> mesesVSanhosHAQ
      V1   V2   V3   V4  V5   V6   V7    V8    V9   V10  V11  V12
2000 412  445 8918   90 212 1167 1412 12788 20390   111    0    8
2001   0  579    4  525 431 1079  945  1523  5756   146 1138 6228
2002 418 1570 4142 3380 255  625 1410  9274  4871   179    0    1
2003  67   46 2465 1110 559 2296  283  7449  5371   157    7   11
2004  13 3041 2560 3770 986 5177 9869  4787   986   612  198   97
2005 681 2367 7258  422 186  932 2064 35576  3394  3811   16  745
2006  87 1194  196  492 694  982 5956 83274  1918     3   14   53
2007  17    9  246  468 118   11   58   398  1600   569 3263  294
2008 276 3827  370  381  60   68  422   455   180   211   39   47
2009  24  729 5234  451 437   88   92   736  2464   455    3   26
2010  28   83 1403  603 278  183  505 10039   936   727    4   18
2011 315 1264 1984 2442 490 2174  799  1818  2068 28962   59   15
2012 188 3618 6295  238 115   60  432  2562  1994    88    5    3
2013  80    3   63  278 106  347  426  7747  9347    25  145 1211
2014   0    0  379   93 173   99   37    49    30    32    0    0
```

Figure 37: Burned hectares' distribution by months and years.

The analysis by weeks (Figure 38) shows two main periods of fires, the first one between 10th and 12th weeks of the year (middle of March), and a second one wider from 28th to 39th with a maximum between the 31st and 37th week (corresponding to August and the beginning of September). If talking about BH the critical period is from 31st to 33rd weeks and the periods with a higher HFR are located between 31st to 36th, 41st to 42nd (October) and 49th to 51st (December). It is curious that these weeks have a high HFR because the conditions are usually really hard for fire propagation at this time of year, but the fact is that the prevention measures at this time are weaker and this circumstance makes the produced fires bigger than usual. From now on, the red horizontal line will represent the mean for all the values represented in the plots.

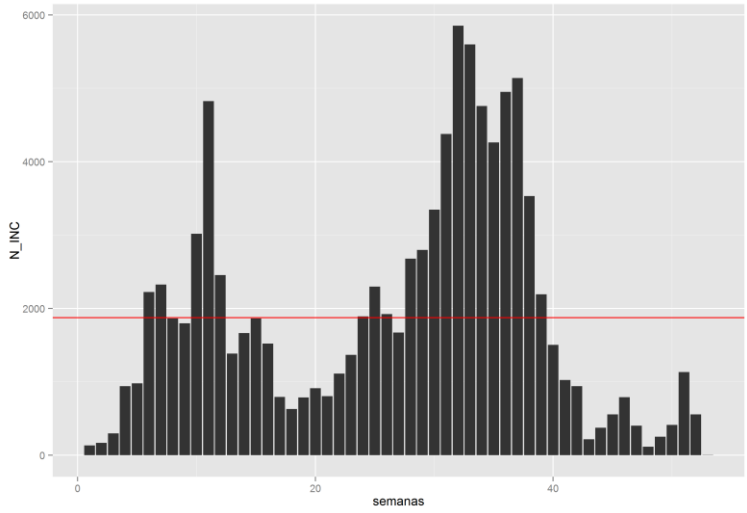


Figure 38: Number of fires from all the years divided into weeks.

According to weekdays, statistics evidence a slightly higher activity at weekends both in NF and BH. According to the hourly distribution, as it can be seen in Figure 39, there is a wide upper between 15 and 23 hours in NF, but the hours of the day with higher BH are from 14 to 19h. So it is critical to detect fires at an early stage within those hours.

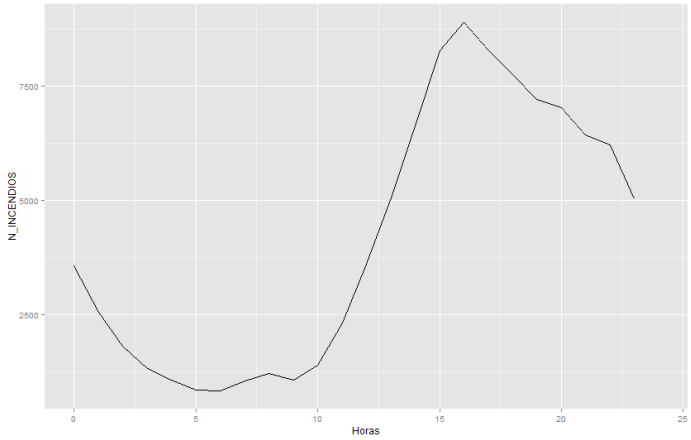


Figure 39: Distribution of fires according to the hour of the day they were started.

6.1.1.2. Causal General Analysis

By changing from a temporal to a causal analysis, a lot of information can be extracted also from a general perspective. As explained in the Introduction chapter, the large majority fires started in Galicia are intentionally provoked. In order to offer a better understanding on the situation, we will start by explaining the causal classification given by the database. There are **six main groups or fire causes**:

1- Natural cause
2- Negligence
3- Accident
4- Arson
5- Unknown
6- Reproduced

Natural causes fires are referring to those ones provoked by lightening, the unknown cases are the fires that cannot be classified in any of these groups due to lack of data and the reproduced fires are those new fires created from an existing one. Causes 2, 3 and 4 need a deeper description, since they enclose a lot of different subcategories. Negligent and accidental fires are divided into sub-causes and arsons are separated into different motivations.

After this classification was explained, it is time to show the causal distribution of fires in Galicia. It is easy to see that the vast majority of the fires taking place in Galicia are arsons (Figure 40). It is also remarkable that the second group with more fires is the unknown cause, but it is normal to think that most of the fires in this group would actually belong to the biggest group. Most of reproduced fires are also produced with an arson as its source, so it is really clear how important the problem is. More than 80% of fires produced in Galicia are arsons (without including any from causes 5 and 6). Negligent fires have a minor role in comparison to arsons and natural and accidental causes are almost insignificant.

The distribution for BH is quite similar, but in this occasion, reproduced fires overtakes negligence as 3rd most important cause. Attending to HFR, it is remarkable that every reproduced, natural or arson fire is, in mean, two times bigger than one caused by negligence or accident.

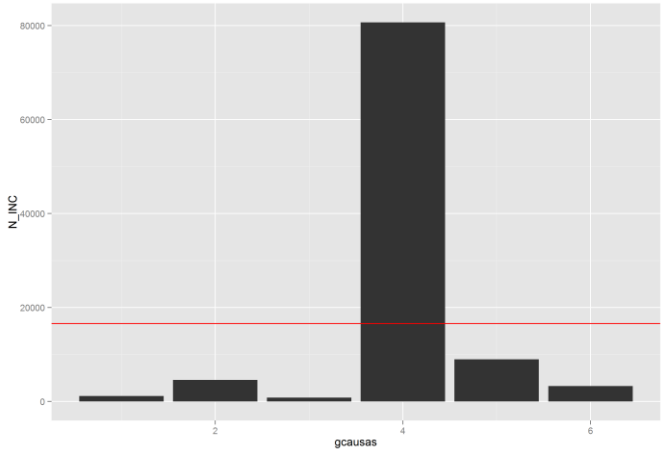


Figure 40: Distribution of fires divided by causes.

Although not very significant, negligence and accidents deserve an in-depth analysis. By studying the different sub-causes behind the negligence group, we can see that the major part of these fires are due to:

- Agricultural burns (1000 fires)
- Forestry work (560)
- Brush burnings (400)
- Garbage and landfills (400)
- Regenerating grasses (300)
- Stakes (250)
- Smokers (100).

Around 1600 fires from all the 4500 are due to other or no specified negligence. Most of fires behind the 850 accidental fires are caused due to different type of accidents involving vehicles.

Even if the last sub-classification was interesting to mention, it is much more important to study the motivations behind arsons. In the plot below (Figure 41), a comparison between all the NF distinguishing by motivation can be observed. Without considering the last two (no data and other motivations), we can find that the most important motivations are:

- farmers eliminating brush and residues (26800 fires in last 15 years)
- shepherds and breeders to regenerate pastures (7700 fires)
- actual arsonists (6850 fires)
- vandalism (1840 fires)
- hunters due to facilitate hunting (1300)
- fires produced for driving away animals (970)
- revenges (500)
- people for changing land uses (300)
- dissensions (100)

It is remarkable that many more fires are being produced due to the two firsts motivations (34500 fires) than by actual arsonists (6850), the third most important motivation after arsons in Galicia. Contrary to most people's opinion, it is also important to mention that just 300 fires were started due to people who wanted to get changes of land uses.

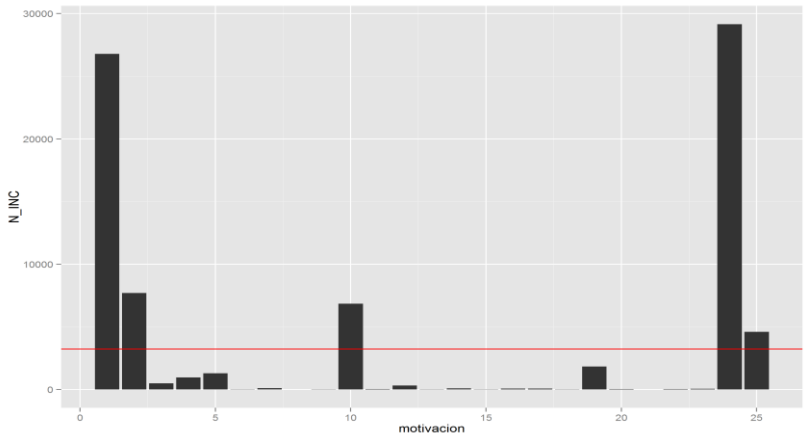


Figure 41: Distribution of number of arsons divided by motivations.

So, the most problematic points concerning arsons (they were previously shown as the biggest problem in Galicia) are produced by **rural agents** (farmers, shepherds and breeders) representing **more than the 75% of arsons** from those ones classified.

This is **clearly the biggest problem** at a global level in Galicia and the reasons for these as the main causes of fires are the following: Galicia has an enormous forest area, the dispersion of the population is brutal (more than half of the urban centers in Spain are there), private property of this huge territory is divided into a lot of tiny smallholdings (parents dividing their lands to their children, and those to their children and so on from a long time ago), rural exodus by new generations leads to the abandonment of the countryside, and the fire culture extended in the rural Galicia since time immemorial. Farmers do not have money, time or willingness to clean up their land of bushes and they use fire as a cheaper and faster technique to eliminate bushes or regenerate pasture.

Some measures have been taken for the Galician government in order to reduce this kind of fires, but they seem to be inefficient for the moment. Awareness programs and harder punishments have been performed, but just a small part of perpetrators are judged, and just a small part of these last ones are imprisoned. In most cases, even if it is known who the perpetrator was, it is difficult to prove it, and most people in small rural areas do not want to accuse his neighbor.

By only significantly reducing these kind of fires, Galicia would reach almost a normal level in comparison to some other regions of Spain. The only solution here would be to create a new territorial planning, but this involves long term measures, something that clashes against

current politics, only worried about short term results. The importance of this phenomenon is clearly shown in some other parts of Galicia, like the Mariña Lucense, at the north of the province of Lugo, where the farmers are well organized, and hardly any fire is taking place.

6.1.1.3. Causal vs Temporal Analysis

Concerning the causal distribution of fires along the time, the following results have been obtained:

In last 8 years, natural and accidental fires were below 100 fires per year each and negligent and arsons were, in mean, reduced by a factor 2 from the previous 8 years' period. Undetermined and reproduced follow a very similar pattern to arsons, showing that most of the fires into these classes are derived by arsons.

Monthly analyzing, most of fires due to natural causes focus mainly on the summer months (where more storms are occurring in this region) as well as the accidental ones, negligent fires are most uniformly distributed covering Spring and Summer and the arsons exhibit the two peaks showed in the general analysis, since they are the dominant cause. Undetermined and reproduced follows a similar pattern as well. These distributions can be seen in the figure below in the following order for every cause group: 1-2-3, 4-5-6.

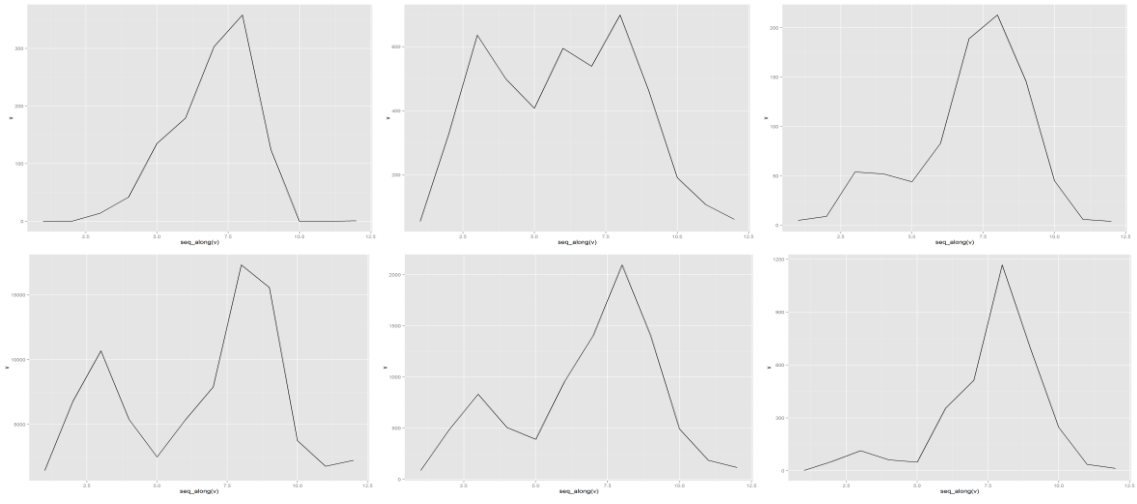


Figure 42: Distribution of fires divided by months for all the 6 big causes.

Attending to significant sub-causes, some of them exhibit an unusual distribution like forestry work (most of these fires are produced from February to June) or authorized bushes burnings with two peaks, one in March greater than the one in August. Looking now to motivations, motivation 1 and 2, as main contributors to arsons, show the famous peaks mentioned in the temporal general analysis, in motivation 2 it is greater in March while in motivation 1 is greater in summer, as well as for motivation 10, 12 and 19 (these last corresponding to arsonists, land use modifications and vandalism respectively).

Differing by weekdays is really remarkable the increase of arsons during the weekends (around 11.000, in mean from Monday to Friday and 13.500 in mean in Saturday and Sundays). This curious trend is mainly produced by the same trend in motivations 1 2, 10 and 19.

In terms of hours, the fire distribution also differs by cause. For example, the natural ones are very concentrated between 16 and 21h while negligent are more disperse (maximum from 12 to 20h) or arsons (from 14 to 24h). In this case reproduced fires differ from arsons and are more prone to be produced from 12 to 20h because the conditions are more favorable at these hours. Same motivations than before are behind this hourly distribution for arsons.

By doing this analysis, **some patterns along time can be distinguish from cause to cause**. It may not seem really helpful in a practical way to do this analysis for Galicia, but it may be really useful to known what is happening in a general manner in order to manage personal and material resources in small places where some causes or motivations are behind the majority of fires.

6.1.1.4. Causal vs Weather Analysis

A weather analysis was also made for every particular cause, but results were not very satisfactory. By using just 10 meteorological stations it is very difficult to specify the exact meteorological conditions in the point where the fire was started. Almost every important cause or motivation was following a similar pattern to the one shown below for arsons (Figure 43). On it, it can be seen that most of fires are between a temperature of 10 to 20ºC for a 100% relative humidity and goes up till 20ºC to 30ºC for a humidity of 75%. In this part is where most of the fires seem to occur. This happens because relative humidity and

temperature are inversely proportional relational parameters. For lower relative humidity fires seem to vary more randomly with temperature.

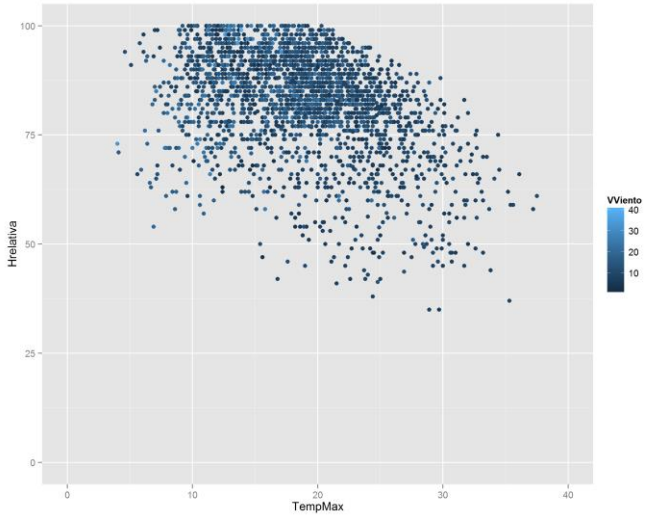


Figure 43: Distribution of fires attending to maximum temperature, relative humidity and wind speed.

At least, from this chart, it can be deduced that fires taking place with temperatures lower than 10ºC are really infrequent and that there are some cases varying with weather variations. For example, in Figure 44, arsons (in black) and natural fires (white) are superposed in order to see the weather conditions generally ruling ones and others. As it could be thought, storms producing natural fires are happening for more extreme conditions with higher temperatures than usual.

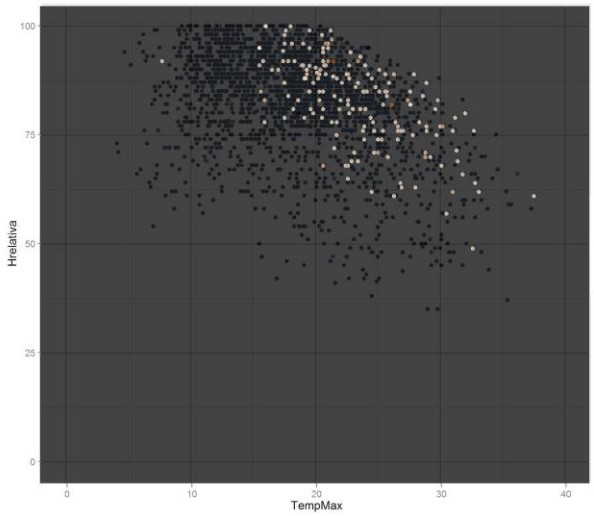


Figure 44: Natural fires (white) and arsons (black) distribution attending to maximum temperature and relative humidity.

6.1.1.5. Geographical Analysis

In addition, a geographical general analysis was made in order to study how the fires are distributed along the territory and to see if fires are very unevenly distributed across the area or not. After, a more in-depth analysis is made in order to locate these more problematic places, and to distinguish geographically between time and causes.

To start, a comparison between the four Galician provinces is made in terms of HSR (burned hectares for surface ratio). The plot below shows the percentage for every province of the HSR. As long as the x-axis is growing, just the biggest fires are taken into account. So, for x equals 0, all fires are taken into account and for x equals 10, just fires burning more than 2000 ha are being considered. From the figure, enormous differences can be observed from province to province. Considering all the fires, we can see a percentage of fires higher than 0.4 for Ourense, 0.3 for Pontevedra, almost 0.2 for A Coruña and lower than 0.1 for Lugo. This is, for same sizes, Ourense would be 5 times more burned than Lugo and 2.5 more than A Coruña. These are really big differences for big territories as the provinces we are talking about.

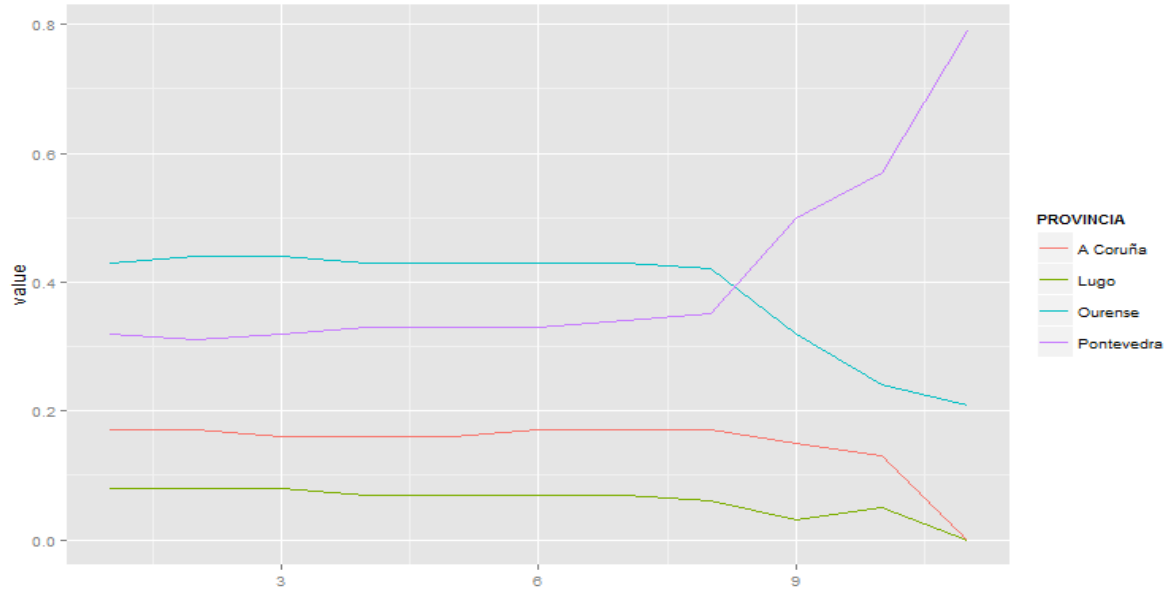


Figure 45: Hectares-surface ratio for the 4 provinces.

It can also be elucidated from the plot that the majority of big fires are concentrated in Pontevedra and almost none of them are occurring in A Coruña or Lugo. The FSR (fires-surface-ratio) is also similar. The HFR (hectares burned by fire ratio) is more similar between

the provinces but particularly higher for Pontevedra, possibly due to those big fires mentioned before.

Going in a lower level analysis, more specifically at a municipal level, this fire dispersion also can be observed. The mean for the HSR is 14.5 ha/km2 burned since 2000 but, by observing the plot below, it is easily noticeable that a lot of disparities are being produced from municipality to municipality, with some municipalities showing values around 100 ha/km2 (this is the equivalent to its whole territory burnt in last 15 years) but, on the other hand, half of the municipalities with this ratio lower than 8 ha/km2. So the disparity of the geographical fires distribution is enormous, varying a lot from term to term. Along the x-axis a color appears representing the province to which the municipal term belongs.

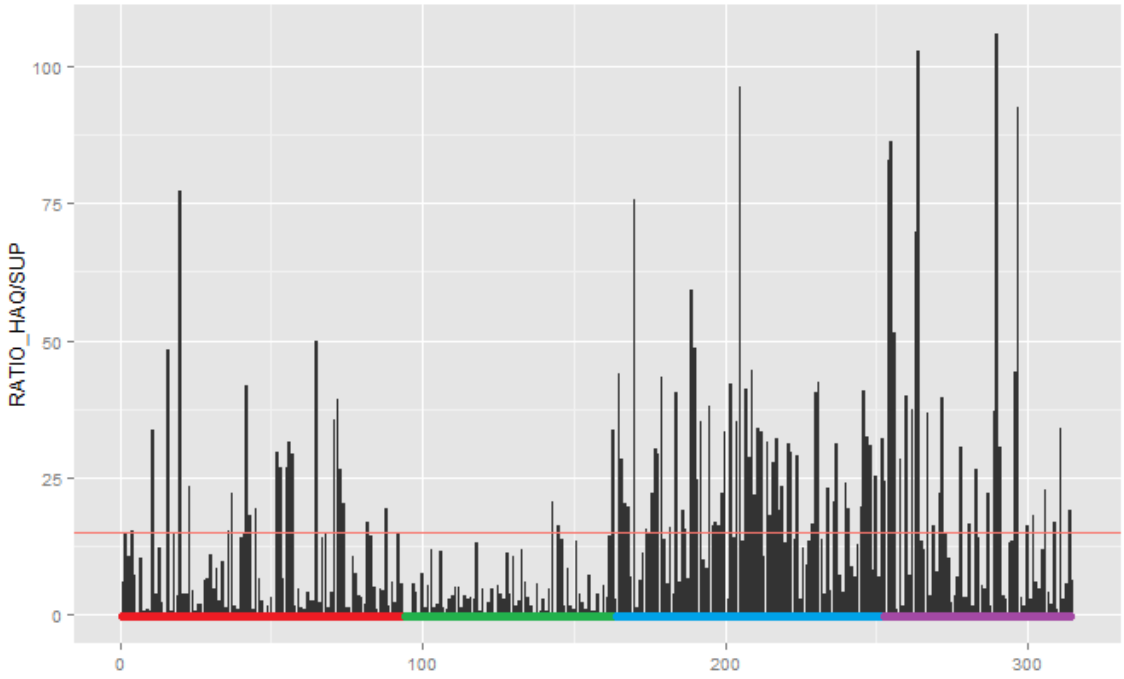


Figure 46: Hectares-surface ratio for every municipality.

In the FSR case, the differences are not so big but also remarkable. The mean is around 1 fire/km2 but there are also some municipalities showing a ratio bigger than 4 fires/km2 while half of them show less than 0.75 fires/km2. It is also important to say that not necessary municipalities with more NF are the more with more BH. In order to illustrate this, the HFR is

studied also for every municipality, showing some with more than 100 ha/fire while the mean is in 15.8 ha/fire[1].

From this analysis, the main obtained result is the **enormous geographical dispersion, in both at a municipal level and even at a provincial level.** In order to see where the most problematic municipalities are, some plots were created. On them, a point is placed in the Galician map for every municipality, with a redder or greener color depending on the higher or lower that the ratio measured is for that municipal term (Figure 47).

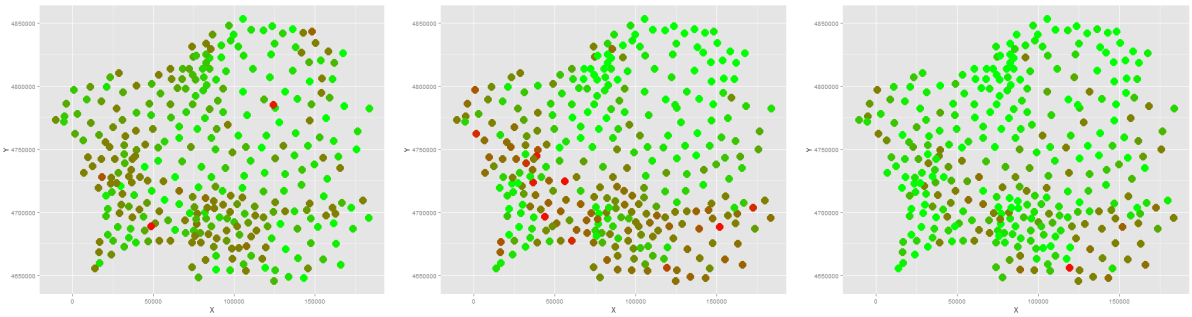


Figure 47: FSR, HSR and HFR for every municipality.

The HSR shows that the biggest problems occur in Ourense, the interior part of Pontevedra and the southwestern part of A Coruña. For the FSR the picture is similar to the first one while the HFR seems to be more regularly distributed across the four provinces. By doing this analysis and detecting where the most problematic municipalities are, a better distribution of resources, high level talking, can be made.

6.1.1.6. Geographical vs Causal and Temporal Analysis

Once this general analysis is made for the Galician geographic distribution, it is time to check if there are some relations between causes and the geographic distribution of fires. As it was

[1] For this general geographical study, just fires of more than 1ha were taking into account. It is a must to store data from all the fires bigger than 1ha, but it is also recommended to store data from smaller fires. Since these small fires may be stored or not depending on the place, it was preferred to work just with the big fires at this section.

seen in the *Causal vs Temporal Analysis* section, some causes follow different patterns along months or even hours. If we could find also a relationship between causes and geography, the resources could be distributed in a more efficient way depending on time. And we can:

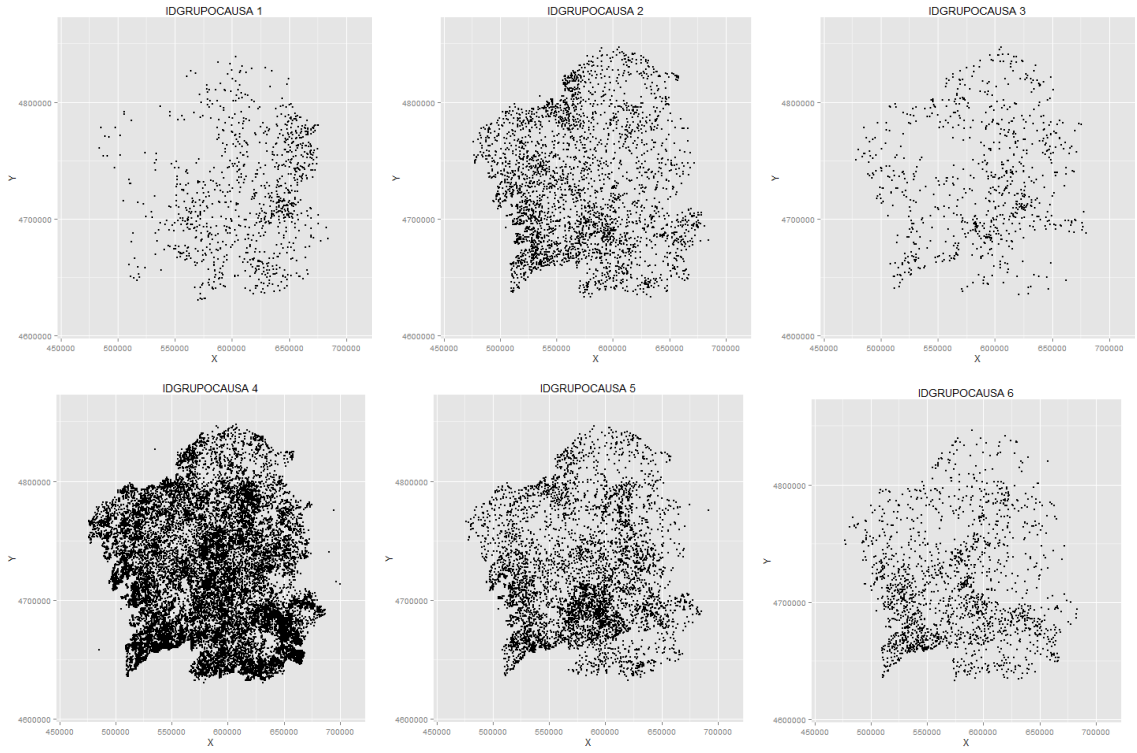


Figure 48: Geographical distribution of fires differing by cause (1-2-3,4-5,6).

Galician map differs a lot if separating causes (Figure 48). This new analysis was made without splitting the fires into municipalities but painting directly the coordinates where every fire started. It can be seen that natural produced fires are happening more commonly in Lugo, in Ourense and Pontevedra border, and in Lugo and A Coruña border, mainly corresponding to mountain areas. On the contrary, most of negligent fires, are being produced on the most populated areas of the region, more specifically in Pontevedra shore and in the metropolitan area of the cities of A Coruña and Ferrol. Accidental fires seem to be more randomly distributed and arsons, since they represent the big majority of all the fires, follow a similar pattern than the one obtained without cause distinction. The same occurs with the reproduced fires. The map representing the fires with unknown cause is also relevant because it shows in which part of Galicia the investigation efforts must be improved.

The problem of the previous analysis, as it was seen at the beginning of the document, is that the vast majority of fires are arsons, so distinguishing by this causes is not really decisive. The good point is that, in a similar way, another analysis can be made but distinguishing between the different motivations behind arsons.

This new analysis also shows an important dispersion attending to this criterion. For example, if observing Figure 49, differences can be observed by comparing motivation 1 (a) and motivation 2 (b) maps, but what is really remarkable is the geographical distribution of fires started by arsonists (c) that are focus mainly in the province of Pontevedra and southwest of A Coruña. After observing this phenomenon, a deeper analysis was made, and it was concluded that every single year the number of fires of this type is much bigger in Pontevedra than in any other province in Galicia. Another map from motivation 19, vandalism (d), shows that most of the activity due to this motivation is produced near the biggest population centers.

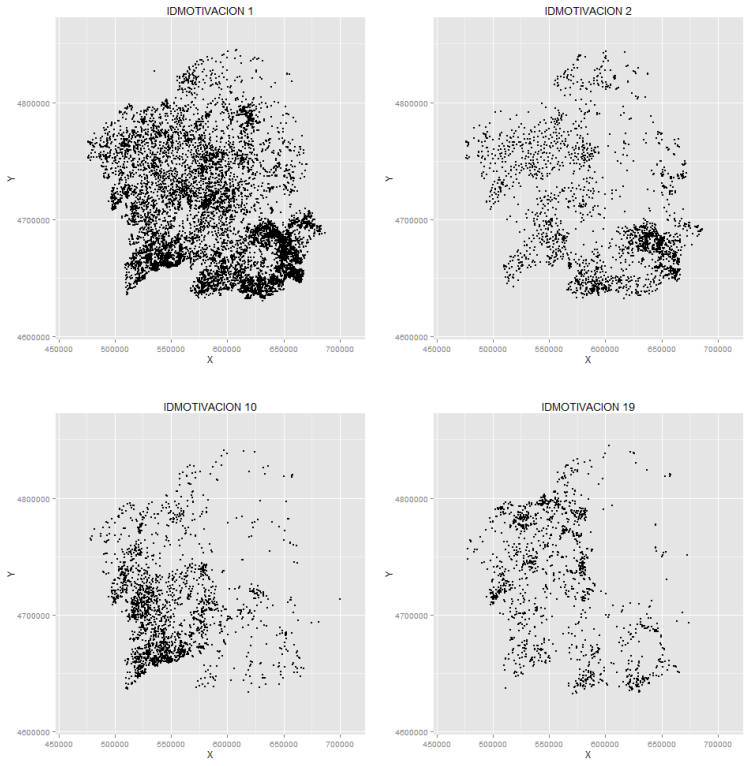


Figure 49: Geographical distribution of fires differing by motivation (a),(b)-(c),(d).

If analyzing geographically by months some differences can also be observed (Figure 50), but not so clearly in this case. For example, if we focus in the two main annual peaks produced in

Galicia (March and August) some distinctions can be made, so resources should not be equally distributed in different periods of the year. In both cases, some differences are also produced with an intermediate month as June. In the hourly study, a strange pattern was find in the south of the province of Pontevedra, where some significant activity was registered between 2:00 and 7:00 AM, probably due to arsonist identified before.

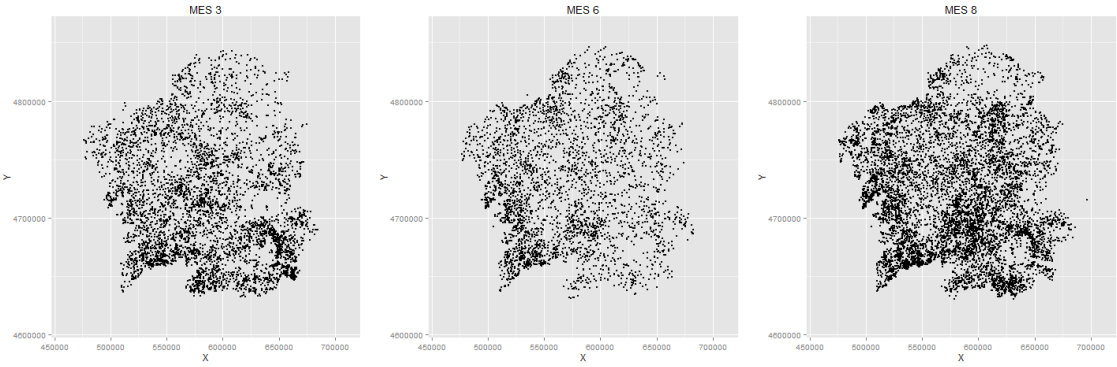


Figure 50: Geographical distribution of fires differing by month (March, June, August).

So, the main conclusions coming out from the different subsections composing this first results part are the following:

- 6.1.1.1. Temporal General Analysis
    - 2 main annual activity peaks were detected: March and August – September.
    - Some differences were also found if looking for different weekdays or hours.

- 6.1.1.2. Causal General Analysis
    - Arsons identified as the greatest cause of fires in Galicia by an overwhelming majority.
    - Main motivations behind arsons were also identified.

- 6.1.1.3. Causal vs Temporal Analysis
    - Important Differences at the monthly distribution of fires depending on causes or motivations.
    - Differences also between causes for different hours.

- 6.1.1.4. Causal vs Weather Analysis
    - Differences found just for some specific causes when relating to weather conditions.

- 6.1.1.5. Geographical Analysis
    - Enormous differences from province to province and from municipality to municipality were detected

- 6.1.1.6. Geographical vs Causal and Temporal Analysis
    - Big differences appear also when looking to the distribution of fires from different causes over the territory.
    - Not so important, when comparing the temporal distribution of fires across the geography.

Identifying this dispersion in the distribution of fires in Galicia according to different causes, times or places, is one of the key concepts of this project. By doing this, some of this initial apparent randomness can be model, and help to reach a better distribution of the firefighting resources from a high level point of view.

6.1.2. Municipality Report

Once it was showed that some temporal, causal and geographical patterns were found from a high level analysis for Galicia, it is the moment to explain the main output of the descriptive study: the municipality fires distribution report. The idea behind it, is to **show to every municipal term where their main problems are, what are their weaknesses, how the fires are distributed along their geography and how their situation is in comparison with the rest of Galicia**. Some captures of an example report for one municipal term are shown in this section and a whole report example is shown in the Annex I.

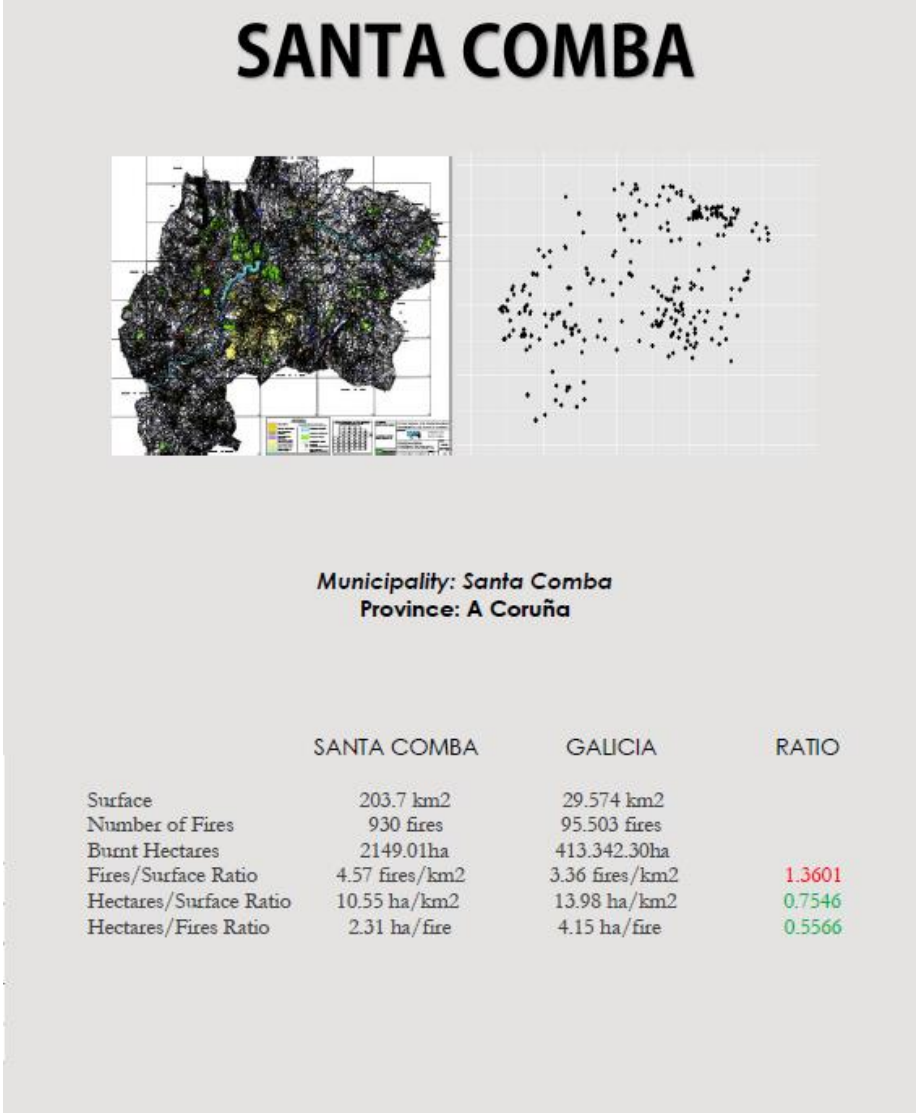


Figure 51: Example of a municipal report: general information.

First, in this report, some general information about surface, NF, BH, HSR, FSR and HFR on the municipal term at issue is given, as well as the same data related to Galicia, in such a way that a first comparison can be made (Figure 51). This information would show, in general terms, if the municipality is in a good level of preventing and extinguishing fires in comparison with the mean of the whole territory.

After this general analysis is made, an analysis of the temporal distribution of fires in the municipality along months (Figure 52), weeks, years, weekdays and hours is provided. With this analysis, local authorities could see if the volume of fires is going down or up through last years, which months, weeks or hours are usually the most dangerous in their region and

find if whether some weekdays are more dangerous than others or not. By acquiring this knowledge, resources could be more intelligently distributed.

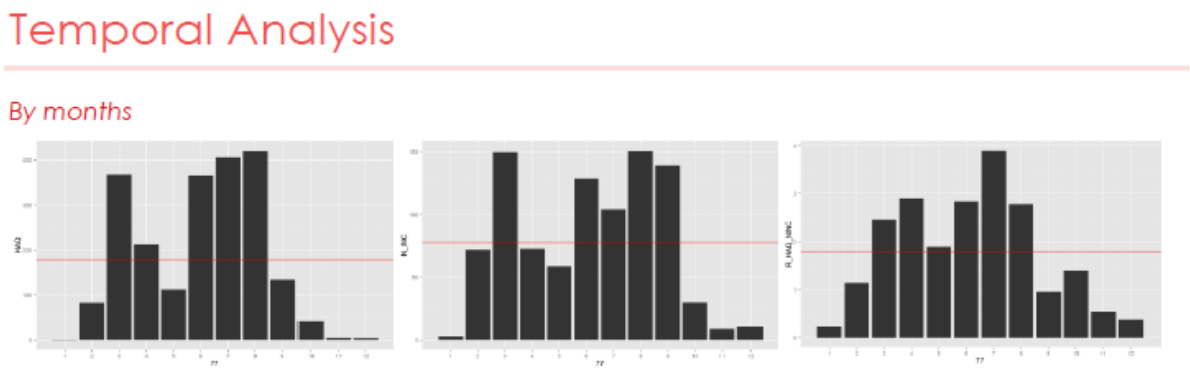


Figure 52: Example of a municipal report: temporal analysis.

The causal analysis will show, at the beginning, a decomposition from all the fires produced in the municipality between the different causes, and those causes between different sub-causes and motivations (Figure 53). After this, a table showing a deep comparison of every single cause, sub-cause and motivation in the municipal term with Galicia is showed (Figure 54).

In this table, information about HSR, FSR and HFR of every of these different categories are compared to the general terms of the entire autonomous community. By doing this comparison, a ratio is given for every cause, sub-cause and motivation and HSR, FSR and HFR. If this ratio is bigger than 1, the situation is worse in the municipal term than in Galicia for the measured parameter, if around 1 it is normal, and if lower, the situation is good. With this table in local authority's hands, main causes of fires can be known at a local level, so more aggressive and direct policies can be conducted.

Causal Analysis

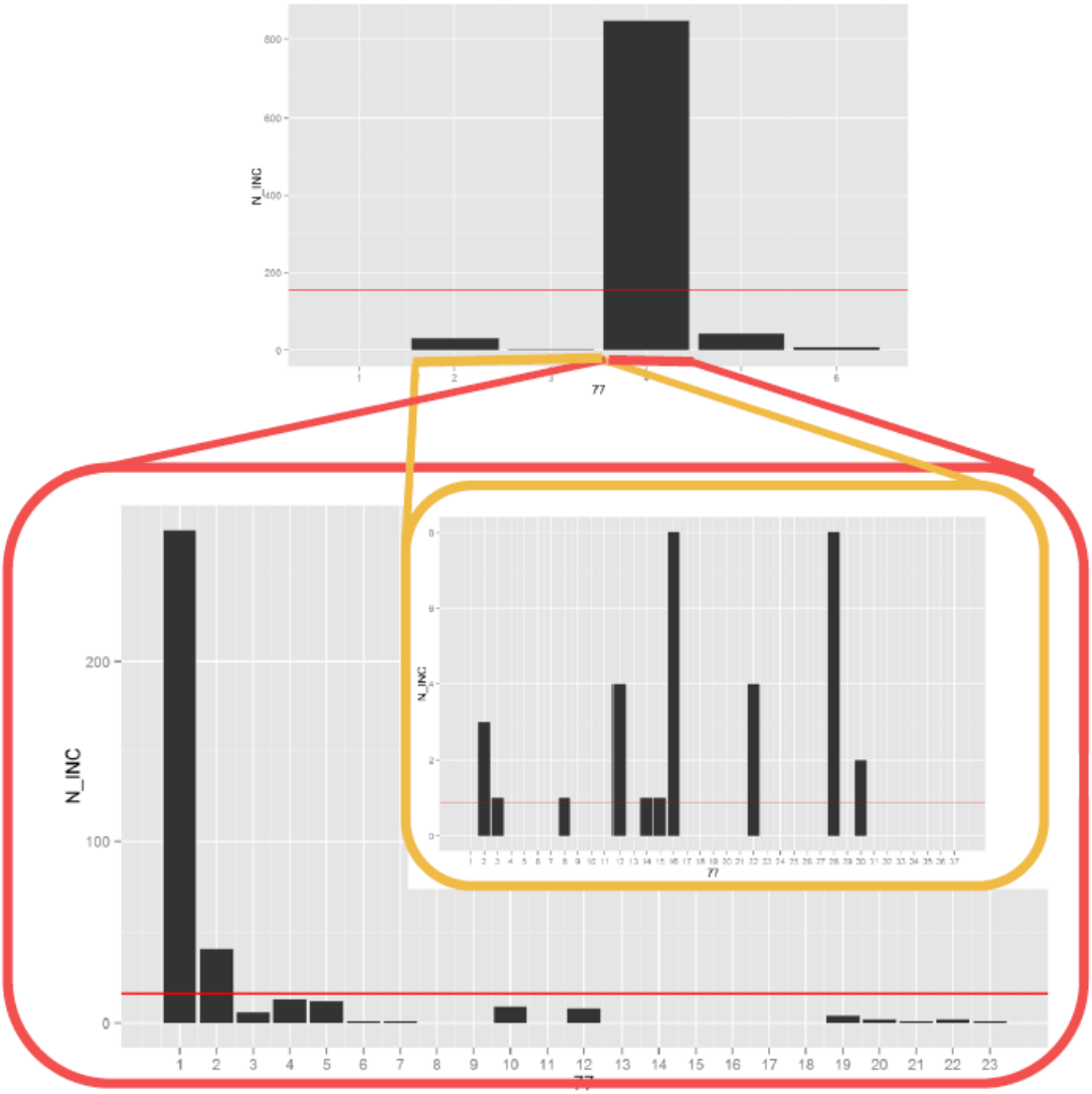


Figure 53: Example of a municipal report: causal decomposition.

CAUSE	SANTA COMBA					GALICIA			COMPARISON		
	BH	NF	HSR	FSR	HFR	HSR	FSR	HFR	HSR	FSR	HFR
1	0	0	0	0	0	0.18	0.04	4.65	0.00	0.00	0.00
2	40.61	31	0.2	0.15	1.31	0.36	0.15	2.32	0.56	1.00	0.56
3	1	2	0	0.01	0.5	0.07	0.03	2.38	0.00	0.33	0.21
4	2014.79	847	9.89	4.16	2.38	11.77	2.73	4.32	0.84	1.52	0.55
5	80.54	42	0.4	0.21	1.92	1.00	0.30	3.30	0.40	0.70	0.58
6	12.07	8	0.06	0.04	1.51	0.60	0.11	5.34	0.10	0.36	0.28

Figure 54: Example of a municipal report: causal analysis.

Last, a geographical distribution of fires divided also by causes (Figure 55), sub-causes and motivations is shown in form of map, so both authorities and firefighters can identify the hottest areas in terms of forest fires. Last but not least, a temporal analysis is made for the most problematic causes, sub-causes or motivations of the municipality in comparison to Galicia (Figure 56). This will provide more relevant information in order to attack any of these cases individually.

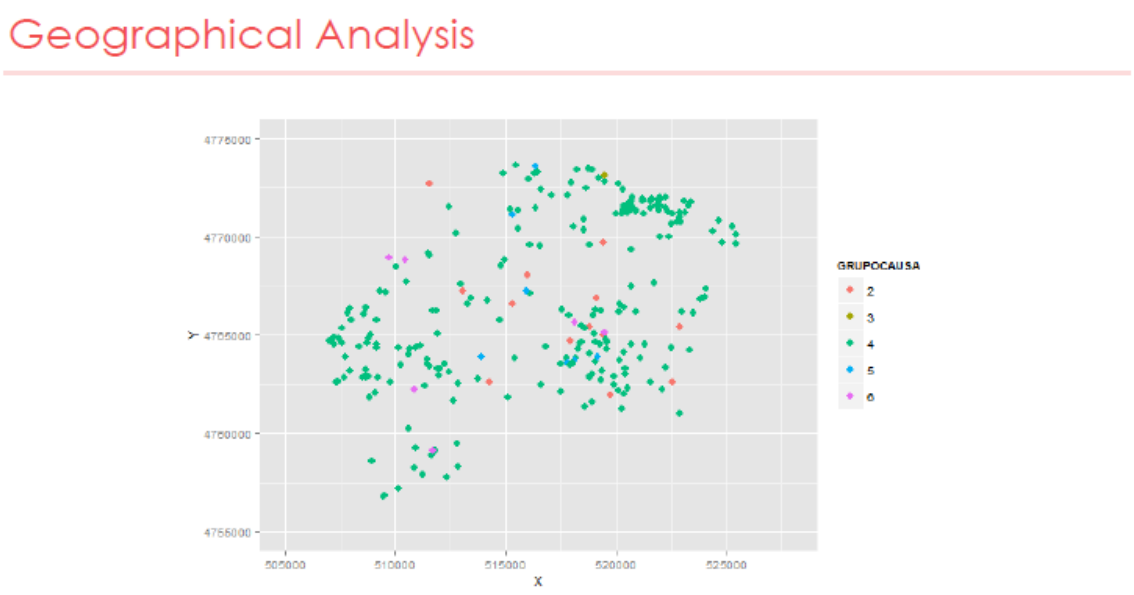


Figure 55: Example of a municipal report: geographical analysis.

To summarize, this report would show to every municipality how its situation is if compared with the rest of Galicia, and indicate where their weak points are at a temporal, causal and geographic level. Plus, it offers a more detailed perspective on the most important causes behind fires in that municipal term.

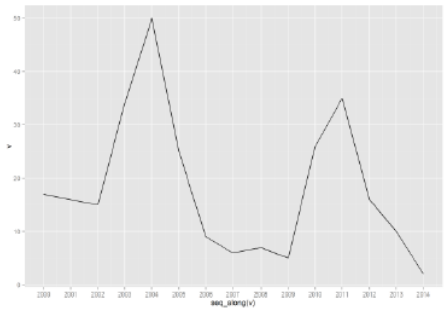
The creation of this report could be easily automatized with RStudio software since it allows documents creation from the HTML plugin it incorporates. So, this report could be delivered to every single municipality and also be published on the Internet.

Causal vs Temporal Analysis

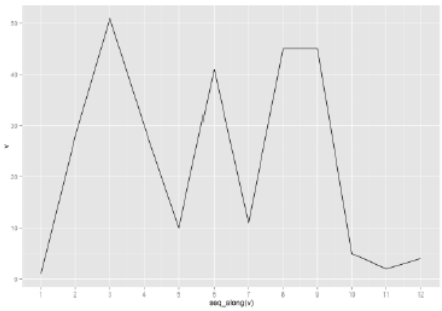
Main problem identified in Santa Comba is motivation #1: Farmers eliminating brush and residues, since it is the only point overcoming the Galician mean with a significant FSR (>0.1 fires/km2).

For this reason, some more information on **Motivation #1** is given:

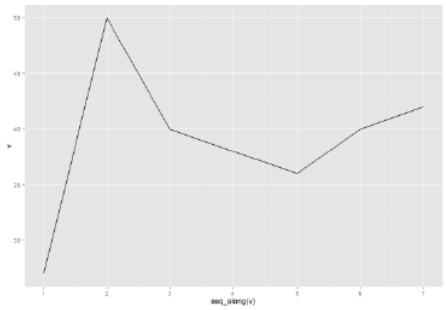
By year:



By month:



By weekday:



By hour:

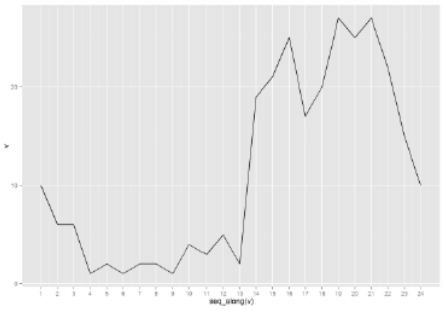


Figure 56: Example of a municipal report: causal vs temporal analysis.

6.2. Predictive System

In order to understand the results of the prescriptive system, it is interesting to explain some important metrics used with this kind of supervised learning systems. The metrics used to evaluate this project are the Confusion Matrix and the area under the ROC curve (AUC), but before getting to this, some other intermediate metrics should be explained.

For every binary classification predictive algorithm (also called detection problem), like this one, there are some common measures of the algorithm performance that must be mentioned. In the testing phase, the algorithm predicts an output for every set of inputs and compare its result with the real result (the output was previously known). By doing this with every set of

inputs in the testing set, the algorithm can catalog the results into some different groups (Figure 57):

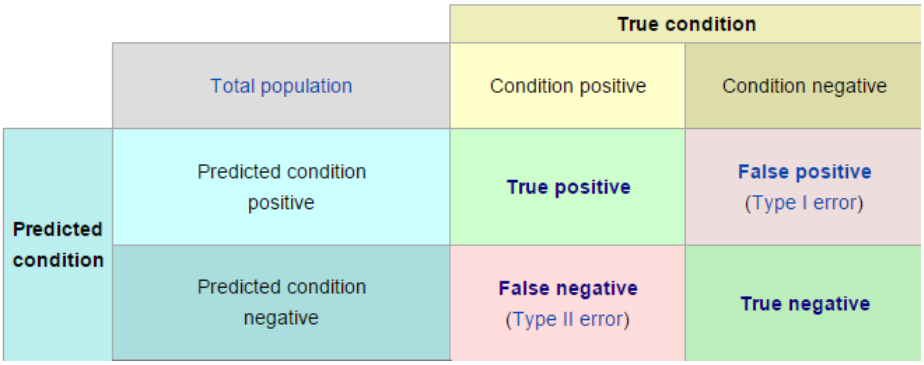


Figure 57: Confusion Matrix.
(Source: https://en.wikipedia.org/wiki/Confusion_matrix)

- True Negatives: algorithm says no, and it was, in fact, a no.
- True Positives: algorithm says yes, and it was a yes.
- False Negatives: algorithm saying no when, actually, it was a yes.
- False Positives: algorithm saying yes when it was a no.

From those parameters, some other important values can be taken: sensitivity and specificity are the most important ones. **Sensitivity** (or true positive rate) is the parameter responsible for measuring the proportion of positives that are correctly identified as such. In this case, it would be the percentage of days with fires correctly identified by the algorithm over the total number of days with fires. **Specificity** (also known as true negative rate) measures the proportion of negatives that are correctly identified as such. In our example, the percentage of the number of days without fires correctly identified as not dangerous over the total number of days without fires.

In this kind of problems there is always a trade-off between these two parameters. If a high sensitivity is required, for example in critical situations where is crucial the detection of a given object, the specificity will be decreased most of the times, because some false positive will also occur. This is, for detecting every time the present object, you are paying some false alarms occurrences that may not create a big problem.

In order to try to maximize the relation between these two parameters, the concept of Receiver Operating Characteristic (ROC) appears. The **ROC**, or ROC curve is a graphical representation of the sensibility or TPR (True Positive Rate, described as the proportion of positive examples correctly classified as belonging to the positive class) against 1- specificity or FPR (False Positive Rate, as the proportion of negative examples misclassified as belonging to the positive class) of a classifier. This metric illustrates the performance of a binary classifier as its discrimination threshold is varied. ROC analysis is used to see how well a classifier can separate positive and negative examples and to identify the best threshold for separating them.

This concept is much more useful than a simple accuracy analysis ([True Positives + True Negatives] / Total Number of Samples). To give an example of this, in a municipal term with 5% days with fires (at least 275 fires in the last 15 years is not a negligible number for a small municipality), the algorithm responsible for getting a good success rate can achieve a 95% accuracy just by saying that there will not be a fire any day. So, this classifier does not seem as an optimal solution for this case.

For every set of inputs, the Logistic Regression algorithm is not directly producing a binary output (0 or 1), actually, it produces a probability (higher or lower depending on the inputs introduced). At first thought, setting the threshold at 0.5 may seem the best idea for a binary classifier, but ROC analysis shows that sometimes, the model can be optimized by varying this threshold. By doing ROC analysis the optimal threshold can be obtained. In this way, ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones.

This analysis calculates the sensitivity and specificity using some different cutoffs values. This means that it calculates many pairs of sensitivity and specificity. If a high threshold is selected, the specificity of the test will increase, but by losing sensitivity. On the other hand, if the chosen threshold is low, the sensitivity will increase but by losing specificity.

The optimal operating point concerning ROC analysis (if assuming equal costs of positive and negative misclassification) is the one in the ROC curve closest to the upper-left corner (point of 1 sensitivity and specificity). This optimal threshold is automatically picked by ROC criterion making the algorithm better.

Another concept underlying this ROC analysis is the Area Under the ROC Curve (**AUC**). The area under the obtained curve is also an indicator that quantifies the overall ability of the test to discriminate between positive and negative input cases. A random test, just like flipping a coin, has an area of 0.5 and a perfect test, the one separating perfectly positive from negative cases, has an area of 1.

To give an example, if considered the situation where the cases have already been classified, if a fire case and a non-fire case are randomly picked, the case shedding a higher output probability should be the fire case. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is, the test correctly classifies the two cases in the random pair). This measure is a good indicator, since it is unifying the most important indicators (specificity and sensitivity) in a single one.

Some others interesting parameters to take into account are the false negative rate, false positive rate, positive predictive value, negative predictive value, false omission rate, false discovery rate, positive and negative likelihoods, and diagnostic odds ratio. The formulas and meaning for all of them are explained in the next figure:

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	

Figure 58: Extended Confusion Matrix.
(Source: https://en.wikipedia.org/wiki/Confusion_matrix)

Results are not the same for every municipality. They depend on the number of total fires in the municipality and also on the amount of randomness that the algorithm is not able to model concerning fires in the past. Plus, some variation is also produced from test to test as long as samples taken for training or testing are randomly chosen. As an example of results obtained, two confusion matrix for Viana do Bolo (a big municipality in Ourense around 270km2 with

1555 fires in the last 15 years) and Beade (a small municipal term also in Ourense of just 6.4 km2 and with 41 fires in the same period of time) are shown. The results for Beade are shown in Figure 59.

In this case, the training set was formed for 6016 samples, around half of them (3014 to be exact) corresponding to days without fires and the other half (3002 samples) corresponding to days where a place had been started. Even if just there were 19 real samples of days with fires, these 19 sample were replicated (Random Oversampling technique) in order to balance the number of samples, as explained in the Methodology section for getting acceptable results. It is also important to remark that only samples from 2004 were taken in this case because there was no meteorological information before that year for that area.

```
> trainingSobremuestreo(Ninc1_32_10,c(1:9))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 619   0
         1 134   5

               Accuracy : 0.8232
                 95% CI : (0.7942, 0.8497)
    No Information Rate : 0.9934
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0574
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.000000
            Specificity : 0.822045
         Pos Pred Value : 0.035971
         Neg Pred Value : 1.000000
             Prevalence : 0.006596
         Detection Rate : 0.006596
   Detection Prevalence : 0.183377
      Balanced Accuracy : 0.911023

       'Positive' Class : 1

> municipios[172,]
    Municipio Superficie CP ID N_INCENDIOS HA_QUEMADAS RATIO_NINC/SUP RATIO_HAQ/SUP
172     Beade        6.4 32 10          41        8.43        6.40625          1.32
```

Figure 59: Confusion Matrix for Beade.

The testing set was formed by 753 samples where no fires had been produced and by 5 days with fires. The results obtained in this particular case were the following:

- 619 true negatives: this is, 619 cases where the algorithm is predicting that no fires will take place on the municipality and no fires are taking place.

- 134 false positives: this value tells us the number of days that the algorithm would say a fire is taking place but no fires are produced.

- 0 false negatives: This zero means that, for this example, there was no occasion where the algorithm is predicting that a fire would not take place and a fire starts.

- 5 true positive: in 5 occasions, the algorithm said it would be a fire and, actually, it was a fire.

Under this circumstances, an accuracy of 82.3% is acquired, **a sensitivity of 100% and a specificity of 82.2%**. So, every time a fire is taking place the algorithm can detect it. Or, to put it another way, it is sure that when the algorithm is saying that there will not be any fire, any fire would start. The only bad new (everything cannot be perfect) is that 18% of times where the predictor says a fire will take place, it will not. A perfect prediction in both ways is very difficult to reach, and in this case, it is much more important obtaining a low rate of false negatives (firefighters are told that no fire will take place and a fire start) than of false positives (firefighters may be alert and no fires are taking place). Obviously, a **100% sensitivity for every case cannot be guaranteed.** This was just a random trial, and some false negatives may occur for other trials in this municipality or some other municipalities with a higher fire rate.

In the other extreme, the algorithm shows a worse performance when working in Viana do Bolo (Figure 60). The confusion matrix shows that 83 days from the 106 days with fires are detected but the bad side is that in 23 days, the predictor is saying that a fire will not take place but it actually takes. Plus, the percentage of false positives has raised a little. In summary the sensitivity and specificity got down to a 78% and 66% respectively. As it was said, this is just a random sample, now, the general case will be studied by showing the AUC for those two cases.

```
> trainingSobremuestreo(Ninc1_32_86,c(1:8))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 433  23
         1 219  83

               Accuracy : 0.6807
                 95% CI : (0.6462, 0.7138)
    No Information Rate : 0.8602
    P-Value [Acc > NIR] : 1

                  Kappa : 0.252
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7830
            Specificity : 0.6641
         Pos Pred Value : 0.2748
         Neg Pred Value : 0.9496
             Prevalence : 0.1398
         Detection Rate : 0.1095
   Detection Prevalence : 0.3984
      Balanced Accuracy : 0.7236

       'Positive' Class : 1

> municipios[248,]
       Municipio Superficie CP ID N_INCENDIOS HA_QUEMADAS RATIO_NINC/SUP RATIO_HAQ/SUP
248 Viana do Bolo     270.41 32 86        1555     8769.14       5.750527         32.43
```

Figure 60: Confusion Matrix for Viana do Bolo.

Using the AUC as a metric, is more efficient in this case since it combines the two main parameters given by the Confusion Matrix (specificity and sensitivity) in one. Moreover, different to the Confusion Matrix metric, the result given by the AUC is not dependent on the individual realization of a simulation, offering a more solid value.

When checking the AUC results, at Figure 61, we can also see that the algorithm works much better for the case of Beade (black line) than for Viana do Bolo (the red one), shedding an AUC of 88.6 and 76.7% respectively. Below, a plot with the ROC curve for every single municipality in Galicia is showed in Figure 62.

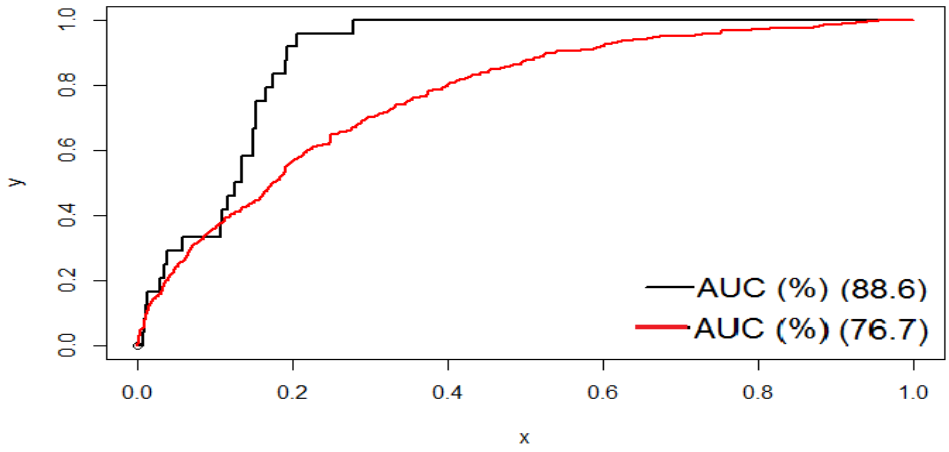


Figure 61: ROC and AUC for Beade and Viana do Bolo.

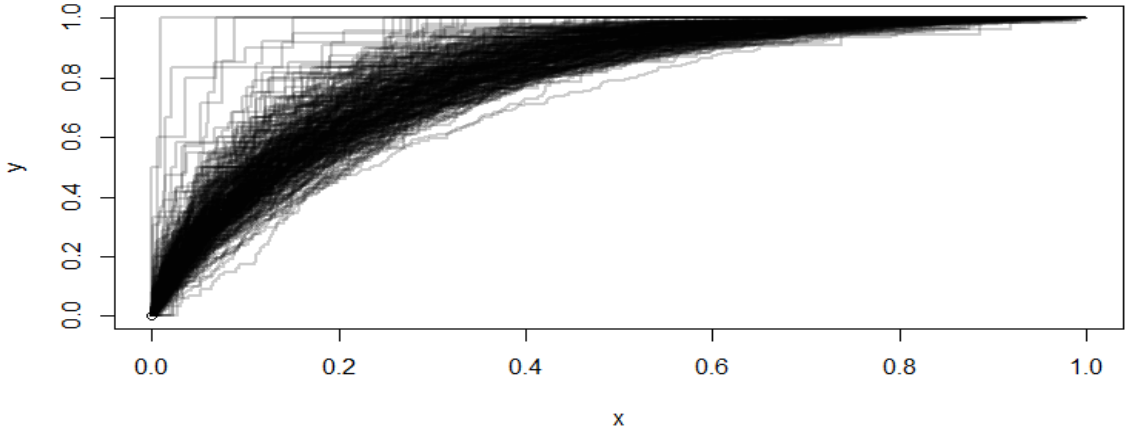


Figure 62: ROC for every municipality in Galicia.

All these ROC curves one in top of another cannot give us a lot of information. Since the Confusion Matrix varies from simulation to simulation, the AUC offers the same value every time. So, in addition, a histogram with the fixed AUC value for every municipality is given (Figure 63). It can be seen from it that **most of the AUC values are around 0.85**, with a minimum in 0.7 and a maximum close to 1.

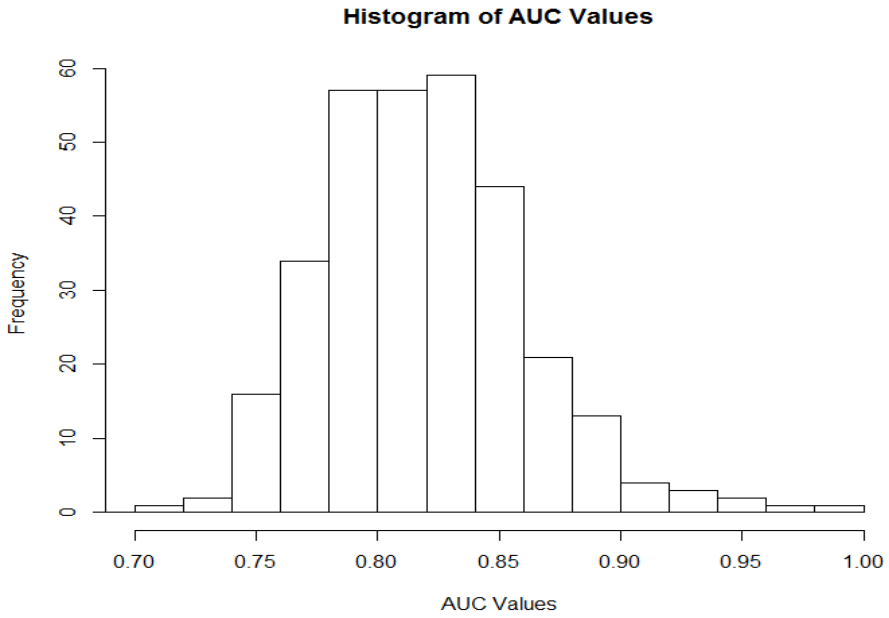


Figure 63: Histogram of the AUC values for every municipality.

To have an idea of how good or bad the algorithm is in terms of AUC, these values can be compared to a commonly used table for accuracy classification in diagnostic tests (Table 2). However, this table is just a rough indicator, since an AUC of 0.9, for example, can be a really bad value in one field of study but being the state of the art in another. Put it in another words, this value will be good or bad depending on the application it is being used for.

AUC RANGE	**Classification**
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.7 < AUC < 0.8	Fair
0.6 < AUC < 0.7	Not good
0.5 < AUC < 0.6	Fail

Table 2: Commonly used AUC classification.

In conclusion, the logistic regression model was chosen because it seemed the most appropriate for solving a binary classification problem like this. The Random Over-Sampling technique was key to get an acceptable success rate and the ROC analysis tends to increase this rate. If using the AUC metric, much more reasonable than just measuring the accuracy in this case, the results show around an 85% of success at predicting whether a fire will take place on a municipality and given day. This rate varies quite a bit depending on the municipal term where it is used, but even for the worst cases, it seems a good predictor.

6.3. Prescriptive System

The prescriptive system will add up the information given by the descriptive and the predictive system by offering conclusions so posterior decisions can be taken. Since the predictive algorithm is not 100% reliable, it is very important taking into account the information given by the descriptive block and the local knowledge about the area when making decisions.

As a first example, some final measures will be tried to extract from the report given for a municipality in Ourense called A Mezquita. The report shows in this case an extremely worrying amount of arsons in this term, representing the 95% of the total number of fires occurred. The problem is remaining throw years with more than 40 arsons every year from 2000 to 2013, having maximums in 2005 and 2011 with values around 125 fires. Studying the motivations behind these arsons, we can find even more worrying values, since motivation #1 (farmers eliminating brush and residues) and motivation #2 (shepherds and breeders regenerating pastures), the most important in all the territory, have in this municipal term values of 5 and 8 times higher than for the whole community respectively, in terms of HSR.

The problem is clearly focus in those two motivations and some policies should be made by experts for curbing them. Some solutions to these problems could be to create some composting areas in the municipal term, for creating compost from pruned branches, dried grass or leaves, that it is what farmers usually burn when cleaning their lands, also an awareness campaign should be realized and a harder control at the dangerous periods. Besides, a more in-depth analysis is made.

As it can be seen on the following plots (that would be shown in the municipality report), these fires are very concentrated between January and March and August and October. Over these months the surveillance service should be significantly elevated. Another relevant data is that both motivations fires are highly concentrated from 13h to 22h, so surveillance and firefighter's resources should be mainly active on the afternoon and evenings.

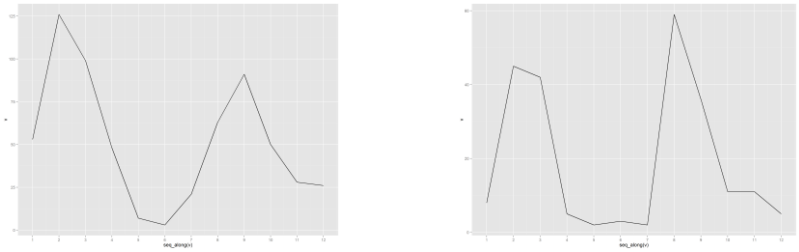


Figure 64: Distribution of motivations 1 (a) and 2 (b) fires in A Mezquita over months.

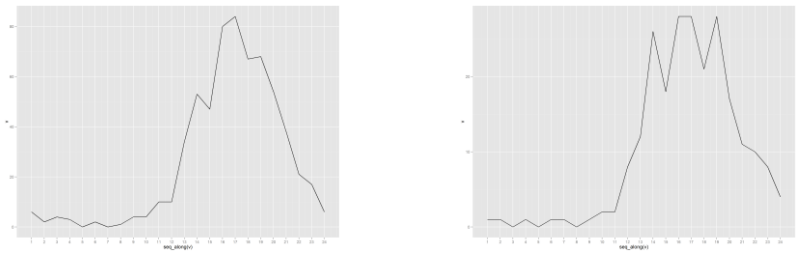


Figure 65: Distribution of motivations 1 (a) and 2 (b) fires in A Mezquita over hours.

Also, since the geographical distribution of these fires is known (Figure 66), these resources should be distributed in a more efficient way. The picture below shows the fire distribution of motivation #1 fires (in orange) and #2 (blue) over the map of A Mezquita where some clear hot areas can be distinguished.

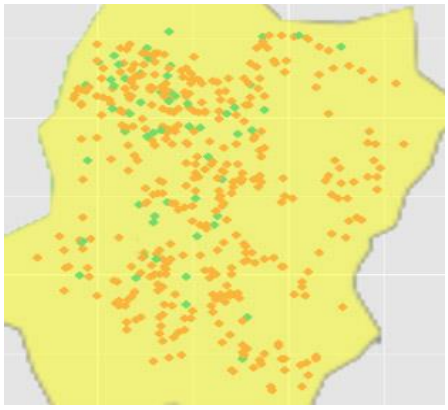


Figure 66: Geographic distribution of motivations 1 and 2 fires in A Mezquita.

Another study case as a prescriptive analysis can be made for the municipality of As Neves in Pontevedra. In this case, arsons are also behind most of the total number of fires declared on the area (just 19 from the 1214 fires produced were from causes 1,2 or 3). But the distribution of arsons between different motivations is different in this case. Motivation #1 appears as one of the main reasons behind fires in this municipal term, with a factor 6 over the ratio in Galicia, but the first reason appears to be motivation #10: arsonists, representing more than half of the burned hectares in the municipality. This amount of fires are almost 30 times the regular quantity of fires of this type per surface in Galicia.

When analyzing data, it can be seen that this arsonist activity was abruptly reduce at two important points: 2005 and 2011. At 2005, with the arrest of an arsonist, fires were reduced significantly and in 2011 a mentally disabled man was also captured for being the responsible of most of the fires in the previous years. This is a clear example of how the descriptive report and the local knowledge of the area must be mixed up to reach to good conclusions. So in this terms, only measures against farmers should be taken.

This prescriptive system also seeks to fix a more approximated local daily risk level, by combining the analysis given by the descriptive system, the output of the algorithm, the local knowledge of the area, and having IRDI into account. For example, for the case of A Mezquita studied before, a positive result given by the algorithm on May might be less significant than a positive in February, once the descriptive system's results are known for this village.

To sum up, in this prescriptive section, it was shown that a deep study of the municipality report, combined with the local knowledge about the area can result in getting to good conclusions and taking good preventive measures. Moreover, if combined with the algorithm from the predictive system and IRDI, a more precise risk indicator could be set in every municipality.

7. FUTURE LINES

This chapter has the purpose of explaining which next steps should be taken in order to get this project improved. Due to the time limitation imposed by the nature of the project, not all the contemplated concepts could be implemented into the system and some of them remained finally out of the scope of the project. **Even if the thesis is complete and presents a full meaning by its own as it actually is** (the goals initially set were accomplished), **some future lines are set in this section.**

Some of these concepts would lead to improvements at the descriptive system, by doing a more in-depth analysis, and some others would be applied to the prescriptive system, by introducing some modifications in the Machine Learning algorithms, metrics or model used.

First of all, the descriptive system could be notably improved. A more in-depth general analysis concerning the relation between the number of fires and the meteorological conditions could be made. As it was told in the introduction chapter, in some occasions, for the same month in different years, the absolute number of fires differs a lot from one year to another. It would be interesting to analyze if there are some meteorological changes from year to year behind these changes in the number of incidents over the same month.

Also, a deeper analysis should be made in order to **study the anomalous months**, this is, months with an extreme and unusual amount of fires. Studying individually every of these months in a meteorological, geographic, causal and temporal manner could lead to get a lot of important new information. Some of these months are, for example, March 2000, August 2006, February 2008 or October 2011.

At a lower level, it would be interesting to use the forest mass distribution for every province and municipality, so more indicative statistics could be extracted. Currently, the comparison between municipalities is being made with total area of the municipality, but for some of them, this area may be covered or completely empty of forests. Plus, by knowing the distribution of the different forest species in every municipality, it could be known what are the most dangerous ones, and doing another causal and temporal analysis according to the specie.

If the current database could be combined with a **GIS** (Geographic information system) map containing a lot of different layers with information about the terrain (altitude, slope), vegetation (trees species), land divisions according to owners and even some meteorological information for different areas and periods, many more new patterns could appear.

On the other hand, the algorithm could also be improved in order to predict with a higher success rate. The basic unit does not have to be the municipality, and it could be set at a lower level helping even more to concretize the dangerous areas. This could be done because at the fire records it is also detailed a smaller division inside a municipality in which the fire took place. Some more accuracy could be possibly gotten if adding a variable to the input set, by indicating the number of previous consecutive days without rains.

Also, the Random Over-Sampling technique used for overcoming the initial problem caused by the unbalanced nature of the data set could be changed for other more sophisticated techniques. A similar method used with the same goal is the **Synthetic Minority Over-Sampling** technique, that generates new artificial minority examples by interpolating samples between the existing minority examples rather than simply duplicating the original examples. [18]. Likewise, Tomek's Link method (TLINK) could be used in order to remove noise or borderline examples.

At evaluation level, the use of AUC (Area Under the ROC Curve) offers us a great metric to know how good our predictor is. However, it was recently shown that the AUC has a deficiency since it implicitly uses different misclassification cost distributions for different classifiers. Specifically, using the AUC is equivalent to averaging the misclassification loss over a cost ratio distribution which depends on the score distributions. Since the score distributions depend on the classifier itself, employing the AUC as an evaluation measure

actually means measuring different classifiers using different metrics. To overcome this incoherence, the "H measure" could be used instead, which uses a symmetric Beta distribution to replace the implicit cost weight distribution in the AUC. [18].

Some other changes could also be applied to the algorithm model. As it is known, some changes appear along different causes or motivations depending on the meteorological conditions or the period of the year. Based on this idea, the algorithm could have as output, a **binary classifier** indicating whether a fire will take place or not **for every cause** (Figure 67) instead of a general one. By doing this, and using the municipality report, firefighters could concretize even more the area where the fire should take place, or the hour of the day when it is more probable to have it.

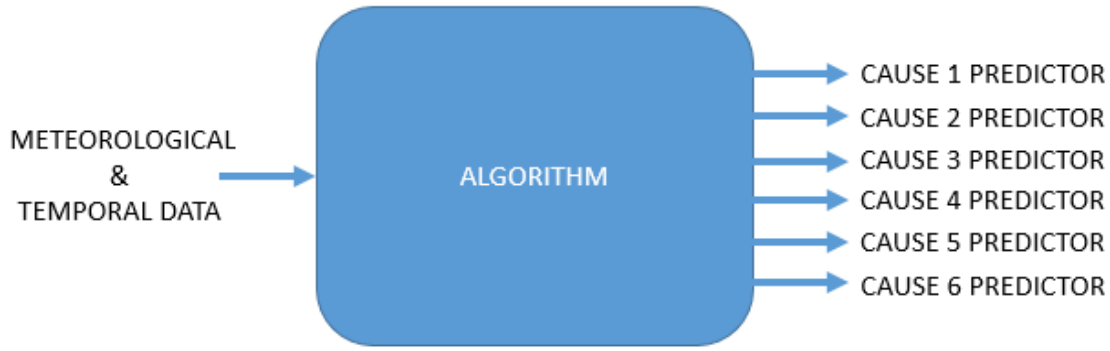


Figure 67: Predictor distinguishing by causes scheme.

Another improvement concerning the algorithm model would be to **merge the current system with a Forecasting** algorithm (Figure 68). This kind of algorithm works pretty good when working with data dependent over time, like for example in stock prediction problems. Somehow, temporal data is already introduced to the algorithm by setting as inputs the month and week for the given day, but the mentioned algorithm would work much better since it is able to find some trends (consistently patterns over time), seasonal (patterns that repeat periodically) and cycles patterns (when data rises and falls over non fixed periods). This can be done by studying the dependency between nearby observations, so it would be important in this case, to divide carefully the training and testing set in bunches of samples and not randomly as it was done with the current algorithm. It is though that combining the existing algorithm with this one would improve significantly the results.

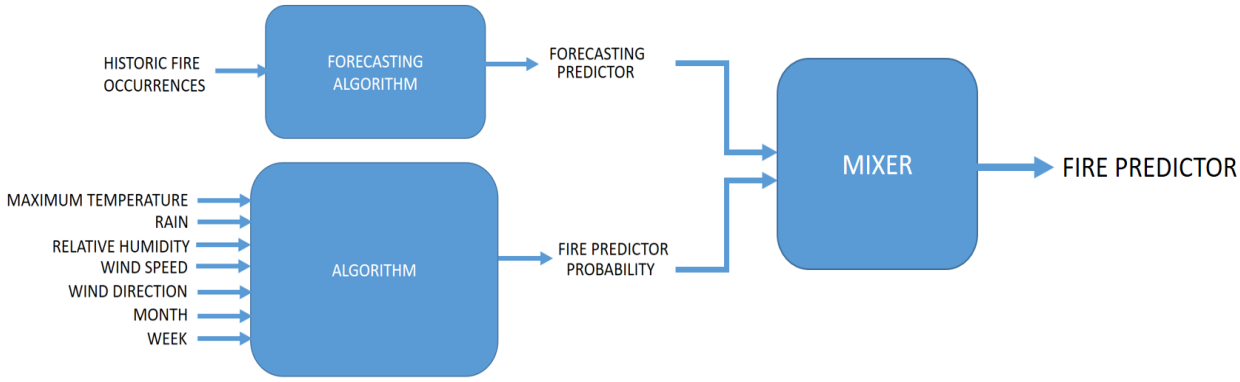


Figure 68: Predictor combined with Forecasting.

8. CONCLUSIONS

The aim of this chapter is to enumerate the main ideas and results produced by the fulfilment of this project, as well as to give some more conclusions concerning forest fires in general.

At the beginning of the document, **the magnitude of the forest fires problem in Galicia was showed**, making obvious that important measures are needed, above all, at prediction. **Traditional politics are not helping** to solve the problem and the existing fire risk indicator in which authorities are based on for distributing the available resources (IRDI) is not optimal, not by a long chalk.

In order to enhance this resources distribution, a system based on Machine Learning techniques was proposed and it was divided into three different sections: a descriptive, a predictive and a prescriptive part.

The descriptive analysis showed, at a high level, that **important variations appeared when studying the fires according to different causes, times or locations.** By knowing more in-depth the fire distribution differing by these parameters, a more efficient distribution of resources could be performed. At a low level, a report with very valuable information was also created for every municipality indicating their weaknesses and how the situation is for them in comparison to the rest of Galicia. This information was not available, till now, to firefighters and local authorities.

Besides, a predictor with an acceptable success rate (around **85% effectivity**) was created in order to predict whether a fire will start in a given municipality and day or not. The algorithm was based on meteorological and temporal inputs and produced using a Logistic Regression

Model and applying Random Over-Sampling. The metrics used for proving the effectiveness of it were the Confusion Matrix and the Area under the ROC Curve also known as AUC.

It is also thought that by putting together the **information from the descriptive and predictive systems in local hands**, and combining it with their knowledge about the area, a more accurate work could be achieved, manifested into final specific measures and a more accurate daily level alert.

From a more general perspective, and as a final comment, it is remarkable to mention that deforestation is one of the main causes of climatic change at a global level, and forest fires are among the most important causes when talking about deforestation. New technologies such as Machine Learning, satellite technology or low power networks could help a lot in this sense, among others.

Nowadays we are living a very dangerous period concerning pollution levels and climatic change. Due to the astonishing population growth and the technological revolution occurred in the last century, we are taking advantage of technology to over-exploit our planet resources. This evolution is providing us a better standard of living, but it should not be at the expense of destroying the planet. **Technology and nature should live together in a harmonious way and this project wants to be an example of it**. Applying technology with good purposes will be critical for getting a sustained growth for new generations.

REFERENCES & BIBLIOGRAFHY

[1] Javier Sevillano, "*Precipitaciones anuales y mensuales en ciudades españolas y europeas y otros datos climáticos",* Dec 2015. <http://javiersevillano.es/PrecipitacionMediaAnual.htm#provincia>.

[2] Ministerio de Medio Ambiente y Medio Rural y Marino, *Evita el fuego… la diversidad es vida*.

[3] <https://faltadeingenieria.wordpress.com/2012/04/03/los-incendios-forestales-en-galicia/>.

[4] <http://www.stopexpolio.com/superficie-forestal-en-cifras/>.

[5] Xunta de Galicia. Consellería do Medio Rural e do Mar, *Pladiga 2015,* Jun 2015.

[6] <http://mediorural.xunta.es/areas/forestal/incendios_forestais/irdi/>.

[7] Marey-Pérez, M.F., Rios-Pena, L., Franco-Vázquez, L. *Metodología para la validación de los diferentes índices meteorológicos de riesgo de incendio para Galicia.* <http://www.congresoforestal.es/actas/doc/6CFE/6CFE01-339.pdf>.

[8] <http://ingenieroforestalalvarez.blogspot.com.es/2010/05/un-sistema-predice-que-dia-comienzan.html>.

[9] Government of Alberta. *Modeling the spread of wildfire*, Aug 2013, <http://aep.alberta.ca/lands-forests/forest-management/documents/BraggCreek-ModelingSpreadofFire-Aug03-2012.pdf>.

[10] W.W. Hargrove, R.H. Gardner, M.G. Turner, et al. *Simulating fire patterns in heterogeneous landscapes,* Jul 2010.

[11] Paul Cortez, Anibal Morais, *A Data Mining Approach to Predict Forest Fires using Meteorological Data.*

[12] R.A. Bradstock, J.S. Cohn, A.M. Gill, et al. *Prediction of the probability of large fires in the Sydney region of south-eastern Australia using fire weather,* Jun 2009.

[13] <http://integraciones.com/index.aspx?IdCon=237&TC=2&IdIdioma=1&IdOrigen=76&AP=False&IdIO=251&IdS=76&IdItemOrigen=243&IdSubItemOrigen=243>.

[14] C. Pennypacker, M. Jakubowski, M. Kelly, M. Lampton, C. Schmidt, S. Stephens, R. Tripp, FUEGO – Fire Urgency Estimator in Geosynchronous Orbit – A Proposed Early-Warning Fire Detection System <https://fuego.ssl.berkeley.edu/>.

[15] Katie Lobosco, *Drones can change the fight against wildfires*, CNN Tech. Aug 2013. <http://money.cnn.com/2013/08/19/technology/innovation/fire-fighting-drones/>.

[16] Bill Gabbert, *Experimental device blasts a fire with compressed air and water,* Nov 2015. <http://wildfiretoday.com/2015/11/06/experimental-device-blasts-a-fire-with-compressed-air-and-water/>.

[17] <https://en.wikipedia.org/wiki/Logistic_regression>.

[18] Nguyen Thai-Nghe, *Predicting Student Performance in an Intelligent Tutoring System*, Dec 2011.

Jeff Leek, Roger Peng, Brian Caffo, *Data Science Specialization*, Johns Hopkins University. [Online course]. Coursera. http://www.coursera.org.

Hastie T, Tibshirani R, Friedman J, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*. Aug 2008.

Yáñez Armesto A, Castro López F, Lombardía Fdez C, et al. *Manual de Prevención e Defensa contra os Incendios Forestais de Galiza*. 2007.

3er Inventario Forestal Nacional. Descripción de los códigos de la base de datos de campo.

Asociación Forestal de Galicia. *O problema actual dos incendios forestais en Galicia.* Apr 2014.

Rosa Almudena Seco Granja. *Aplicación de un Sistema de Información Geográfica al análisis de los datos de incendios forestales en España.* Jun 2009.

Hernando C, Guijarro M, Díez C, et al. *Laboratorio de Incendios Forestales CIFOR-INIA.*

Consellería del Medio Rural y del Mar, *Ley 3/2007, de 9 de abril, de prevención y defensa contra los incendios forestales de Galicia.* Apr 2013.

ANNEX I

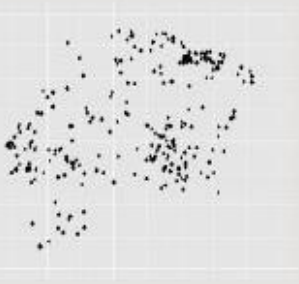
Municipality Report Example

Municipality Fires Distribution Report

Santa Comba (A Coruña)

General Information

SANTA COMBA

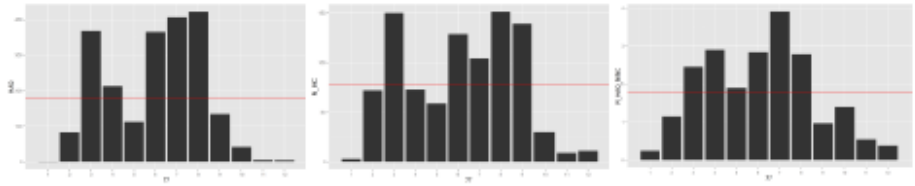


Municipality: Santa Comba
Province: A Coruña

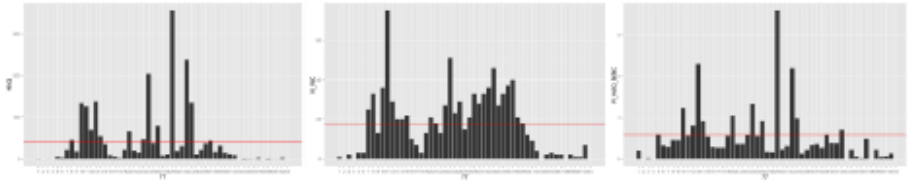
	SANTA COMBA	GALICIA	RATIO
Surface	203.7 km2	29.574 km2	
Number of Fires	930 fires	95.503 fires	
Burnt Hectares	2149.01ha	413.342.30ha	
Fires/Surface Ratio	4.57 fires/km2	3.36 fires/km2	1.3601
Hectares/Surface Ratio	10.55 ha/km2	13.98 ha/km2	0.7546
Hectares/Fires Ratio	2.31 ha/fire	4.15 ha/fire	0.5566

Temporal Analysis

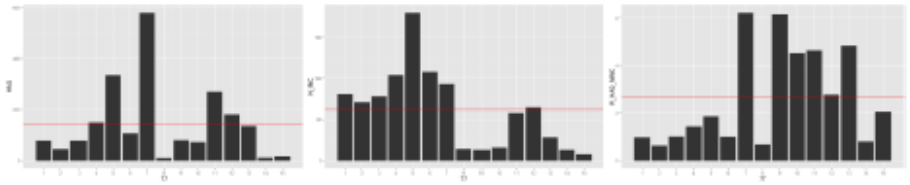
By months



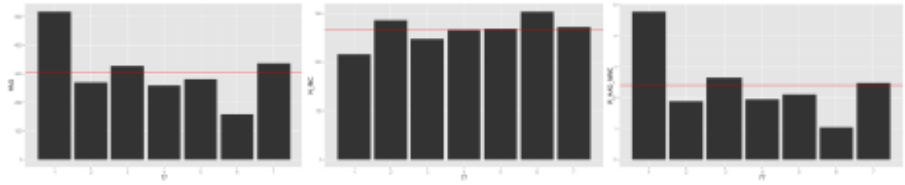
By weeks



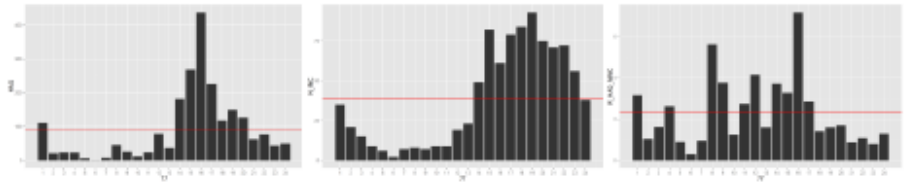
By years



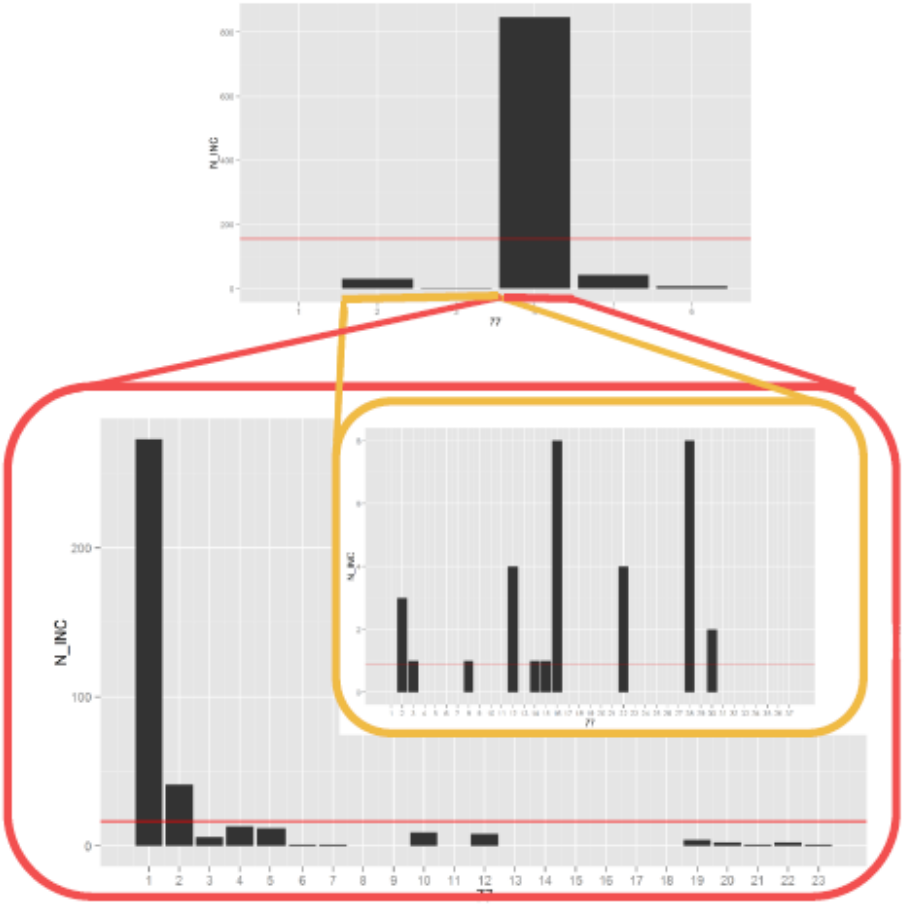
By weekdays



By hours



Causal Analysis



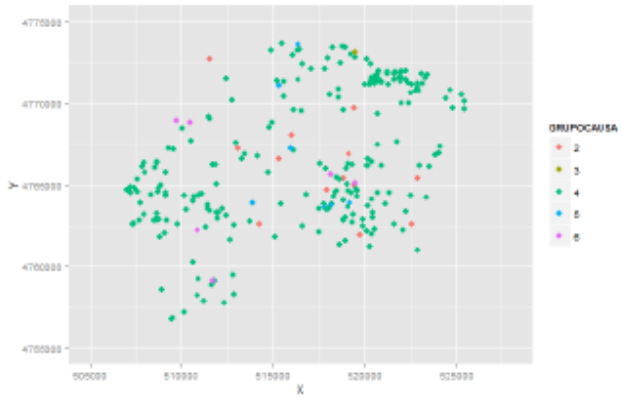
Comparison Against Galicia

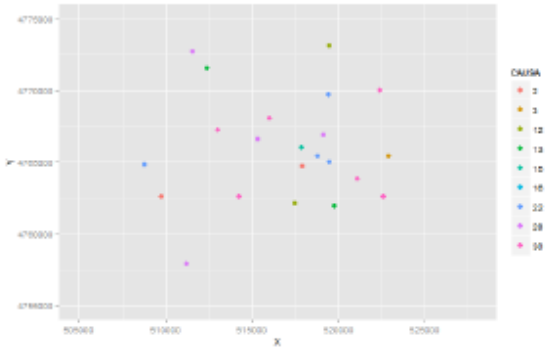
CAUSE			SANTA COMBA			GALICIA			COMPARISON		
	BH	INF	HSR	FSR	HFR	HSR	FSR	HFR	HSR	FSR	HFR
1	0	0	0	0	0	0.18	0.04	4.65	0.00	0.00	0.00
2	40.61	31	0.2	0.15	1.31	0.36	0.15	2.32	0.56	1.00	0.56
3	1	2	0	0.01	0.5	0.07	0.03	2.38	0.00	0.33	0.21
4	2014.79	847	9.89	4.16	2.38	11.77	2.73	4.32	0.84	1.52	0.55
5	80.54	42	0.4	0.21	1.92	1.00	0.30	3.30	0.40	0.70	0.58
6	12.07	8	0.06	0.04	1.51	0.60	0.11	5.34	0.10	0.36	0.28

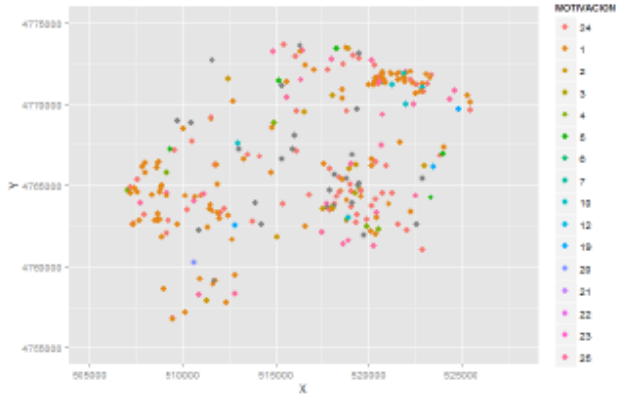
SPECIFIC CAUSE			SANTA COMBA			GALICIA			COMPARISON		
	BH	INF	HSR	FSR	HFR	HSR	FSR	HFR	HSR	FSR	HFR
1	0	0	0	0	0	0.18	0.18	4.65	0.00	0.00	0.00
2	2.44	3	0.01	0.01	0.81	0.04	0.04	1.26	0.25	0.25	0.64
3	1.5	1	0.01	0	1.5	0	0	3.04	0.00	0.00	0.49
4	0	0	0	0	0	0	0	0.25	0.00	0.00	0.00
5	0	0	0	0	0	0	0	0.21	0.00	0.00	0.00
6	0	0	0	0	0	0	0	0	0.00	0.00	0.00
7	0	0	0	0	0	0	0	0.31	0.00	0.00	0.00
8	0.01	1	0	0	0.01	0.02	0.01	2.59	0.00	0.00	0.00
9	0	0	0	0	0	0	0	0.78	0.00	0.00	0.00
10	0	0	0	0	0	0	0	0.25	0.00	0.00	0.00
11	0	0	0	0	0	0	0	0.08	0.00	0.00	0.00
12	9.51	4	0.05	0.02	2.38	0.04	0.02	2.27	1.25	1.00	1.05
13	0	0	0	0	0	0.03	0.01	3.16	0.00	0.00	0.00
14	0.5	1	0	0	0.5	0.03	0	8.37	0.00	0.00	0.06
15	0.01	1	0	0	0.01	0	0	0.4	0.00	0.00	0.03
16	5.54	8	0.03	0.04	0.69	0.03	0.01	2.66	1.00	4.00	0.25
17	0	0	0	0	0	0.03	0.01	2.4	0.00	0.00	0.00
18	0	0	0	0	0	0	0	0	0.00	0.00	0.00
19	0	0	0	0	0	0	0	0.36	0.00	0.00	0.00
20	0	0	0	0	0	0	0	0	0.00	0.00	0.00
21	0	0	0	0	0	0	0	1.39	0.00	0.00	0.00
22	16.52	4	0.08	0.02	4.13	0.07	0.04	1.71	1.14	0.50	2.42
23	0	0	0	0	0	0.01	0	18.02	0.00	0.00	0.00
24	0	0	0	0	0	0.01	0	1.74	0.00	0.00	0.00
25	0	0	0	0	0	0	0	0.01	0.00	0.00	0.00
26	0	0	0	0	0	0	0	0.09	0.00	0.00	0.00
27	0	0	0	0	0	0	0	0.03	0.00	0.00	0.00
28	4.58	8	0.02	0.04	0.57	0.06	0.01	4.22	0.33	4.00	0.14
29	0	0	0	0	0	0.01	0	4.1	0.00	0.00	0.00
30	1	2	0	0.01	0.5	0.03	0.01	1.9	0.00	1.00	0.26
31	0	0	0	0	0	0.02	0.01	3.09	0.00	0.00	0.00
32	0	0	0	0	0	0	0	2.26	0.00	0.00	0.00
33	0	0	0	0	0	0	0	0.76	0.00	0.00	0.00
34	0	0	0	0	0	0	0	1.11	0.00	0.00	0.00
35	0	0	0	0	0	0	0	0.11	0.00	0.00	0.00
36	0	0	0	0	0	0.01	0	3.15	0.00	0.00	0.00
37	0	0	0	0	0	0	0	18.49	0.00	0.00	0.00

MOTIVATION			SANTA COMBA			GALICIA			COMPARISON		
	BH	NF	HSR	FSR	HFR	HSR	FSR	HFR	HSR	FSR	HFR
1	658.76	273	3.23	1.34	2.41	2.43	0.91	2.68	1.33	1.47	0.90
2	31.39	41	0.15	0.2	0.77	1.58	0.26	6.06	0.09	0.77	0.13
3	3.06	6	0.02	0.03	0.51	0.05	0.02	2.9	0.00	0.00	0.18
4	54.35	13	0.27	0.06	4.18	0.23	0.03	7.08	0.00	0.00	0.59
5	45.63	12	0.22	0.06	3.8	0.38	0.04	8.65	0.00	0.00	0.44
6	2.5	1	0.01	0	2.5	0.02	0	15.67	0.00	0.00	0.00
7	4	1	0.02	0	4	0.02	0	4.82	0.00	0.00	0.83
8	0	0	0	0	0	0	0	0.96	0.00	0.00	0.00
9	0	0	0	0	0	0	0	0.73	0.00	0.00	0.00
10	13.75	9	0.07	0.04	1.53	1.38	0.23	5.95	0.00	0.00	0.26
11	0	0	0	0	0	0.01	0	7.13	0.00	0.00	0.00
12	1.99	8	0.01	0.04	0.25	0.02	0.01	1.91	0.50	4.00	0.13
13	0	0	0	0	0	0	0	2.67	0.00	0.00	0.00
14	0	0	0	0	0	0.01	0	3.01	0.00	0.00	0.00
15	0	0	0	0	0	0	0	4.03	0.00	0.00	0.00
16	0	0	0	0	0	0.02	0	8.68	0.00	0.00	0.00
17	0	0	0	0	0	0	0	0.54	0.00	0.00	0.00
18	0	0	0	0	0	0	0	0.34	0.00	0.00	0.00
19	5.62	4	0.03	0.02	1.4	0.35	0.06	5.62	0.00	0.00	0.25
20	0.7	2	0	0.01	0.35	0	0	2.06	0.00	0.00	0.00
21	0.04	1	0	0	0.04	0	0	0.84	0.00	0.00	0.05
22	0.55	2	0	0.01	0.28	0	0	1.54	0.00	0.00	0.18
23	1.4	1	0.01	0	1.4	0	0	0.6	0.00	0.00	2.33

Geographical Analysis





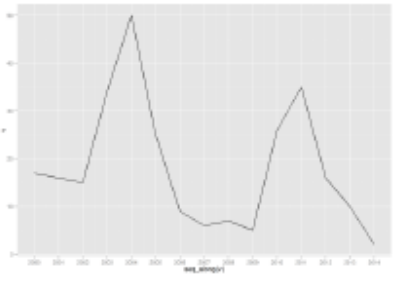


Causal vs Temporal Analysis

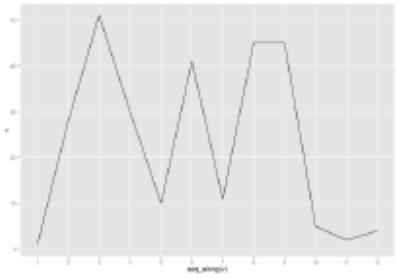
Main problem identified in Santa Comba is motivation #1: Farmers eliminating brush and residues, since it is the only point overcoming the Galician mean with a significant FSR (>0.1 fires/km2).

For this reason, some more information on **Motivation #1** is given:

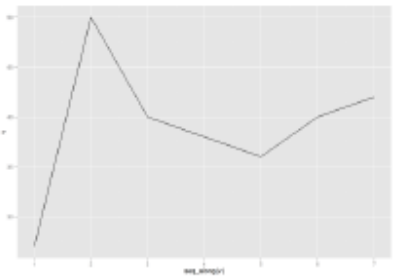
By year:



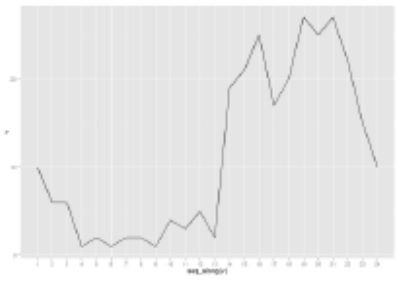
By month:



By weekday:



By hour:



Leyenda

Causas

1- Natural
2- Negligencia
3- Accidente
4- Intencionado
5- Desconocido
6- Reproducido

Causas específicas

1- Quema agrícola (sin especificar)
2- Quema agrícola (quema de rastrojos)
3- Quema agrícola (quema de restos de poda)
4- Quema agrícola (quema de lindes y bordes de fincas)
5- Quema agrícola (quema de bordes de acequias)
6- Quema agrícola (otras quemas agrícolas)
7- Quema para reg. Pastos (sin especificar)
8- Quema para reg. Pastos (quemas de matorral)
9- Quema para reg. Pastos (quemas de herbáceas)
10- Quema para reg. Pastos (otras quemas para pastos)
11- Trabajos forestales
12- Hogueras
13- Fumadores
14- Quema de basura
15- Escape de vertedero
16- Quema de matorral (sin especificar)
17- Quema de matorral (matorral próximo a edificaciones)
18- Quema de matorral (para limpieza de caminos o sendas)
19- Quema de matorral (focos de animales nocivos)
20- Quema de matorral (otras)
21- Otras negligencias (sin especificar)
22- Otras negligencias (actividades apícolas)
23- Otras negligencias (fuegos artificiales)
24- Otras negligencias (globos)
25- Otras negligencias (juegos de niños)
26- Otras negligencias (restos de poda de urbanización)

27- Otras negligencias (otras)
28- Ferrocarril
29- Líneas eléctricas
30- Motores y máquinas (sin especificar)
31- Motores y máquinas (cosechadoras)
32- Motores y máquinas (vehículos ligeros y pesados)
33- Motores y máquinas (accidentes de vehículos)
34- Motores y máquinas (maquinaria fija)
35- Motores y máquinas (otros)
36- Maniobras militares

Motivaciones

1- Provocados por campesinos para eliminar matorral y residuos agrícolas
2 - Provocados por pastores y ganaderos para regenerar el pasto
3 - Provocados por venganzas
4 - Provocados para ahuyentar animales (lobos, jabalíes)
5 - Provocados por cazadores para facilitar la caza
6 - Provocados contra el acotamiento de la caza
7 - Disensiones en cuanto a la titularidad de los montes públicos o privados
8 - Represalia al reducirse las inversiones públicas en los montes
9 - Obtener salarios en la extinción de los mismos o en la restauración
10 - Provocados por pirómanos
11 - Para hacer bajar el precio de la madera
12 - Para obtener modificación en el uso del suelo
13 - Provocados por grupos políticos para crear malestar social
14 - Animadversión contra repoblaciones forestales
15 - Provocados por delincuentes, etc. para distraer a la G. Civil o Policía
16 - Rechazo a la creación o existencia de espacios naturales protegidos
17 - Ritos pseudoreligiosos y satanismo
18 - Para contemplar las labores de extinción
19 - Vandalismo
20 - Para favorecer la producción de productos del monte
21 - Forzar resoluciones de consorcios o convenios
22 - Resentimiento por expropiaciones
23 - Venganzas por multas impuestas

ANNEX II

EXAMPLE CODE

Twenty R files were used in total for accomplishing this Project. Just Which.r, Plot.r and Meteogalicia.r are shown as examples.

Which.r

```
####### WHICH POR HECTAREAS QUEMADAS #######

#incs en los que ardió más o menos de x ha
# y <- "mas" o "menos"
#mode <- "total", "arb", "noarb"
which_masmenos_xha <- function(x1 ,x2 = x1, y="mas", mode="total"){
    if(x1 == x2){
    if(mode == "total"){
    if(y == "mas"){
        which_masmenos_xha <- incendios$TOTAL >= x1
    }
    if(y == "menos"){
        which_masmenos_xha <- incendios$TOTAL <= x1
    }

    }
    if(mode == "arb"){
        if(y == "mas"){
            which_masmenos_xha <- incendios$TOTALAR >= x1
        }
        if(y == "menos"){
            which_masmenos_xha <- incendios$TOTALAR <= x1
        }

    }
    if(mode == "noarb"){
        if(y == "mas"){
            which_masmenos_xha <- incendios$TOTALNAR >= x1
        }
        if(y == "menos"){
            which_masmenos_xha <- incendios$TOTALNAR <= x1
        }

    }
    } else {
```

```
            if(mode == "total"){
                which_masmenos_xha <- incendios$TOTAL >= x1 & incendios$TOTAL <= x2
            }
            if(mode == "arb"){
                which_masmenos_xha <- incendios$TOTALAR >= x1 & incendios$TOTALAR <= x2
            }
            if(mode == "noarb"){
                which_masmenos_xha <- incendios$TOTALNAR >= x1 & incendios$TOTALNAR <= x2
            }
        }
        which_masmenos_xha
}
                                        ####### WHICH POR LOCALIZACION #######

#incs ocurridos en la provincia de A Coruña
which_provincia <- function(prov) {

    which_provincia <- incendios$IDPROVINCIA == prov
    which_provincia
}

#incs ocurridos en Santa Comba dados el código de provincia y municipio
which_municipio <- function(prov,muni) {

    which_municipio <- incendios$IDPROVINCIA == prov & incendios$IDMUNICIPIO == muni
    which_municipio
}

#incs ocurridos en Castriz dados código de provincia, municipio y parroquia
which_parroquia <- function(prov, muni, parro) {

    which_parroquia <- incendios$IDPROVINCIA == prov & incendios$IDMUNICIPIO == muni &
incendios$IDENTIDADMENOR == parro
    which_parroquia
}

#incs ocurridos entre coordenadas x1,x2,y1,y2
which_coordenadas <- function(x1,x2,y1,y2){

    which_coordenadas <- incendios$X >= x1 & incendios$X <= x2 & incendios$Y >= y1 & incendios$Y <= y2
    which_coordenadas
    }


                                        ####### WHICH POR FECHA #######
#incs ocurridos en el año 2005
# o entre el 2005 y 2008
# o 2005 y anteriores o 2005 y posteriores
which_anho <- function(anho1, anho2 = anho1, y="exact") {
    if(anho1 == anho2){
    if(y=="exact"){
```

```
        which_anho <- incendios$ANHO == anho1
        }
        if(y=="ypost"){
            which_anho <- incendios$ANHO >= anho1
        }
        if(y=="yant"){
            which_anho <- incendios$ANHO <= anho1
        }

        }
        else{
            which_anho <- incendios$ANHO >= anho1 & incendios$ANHO <= anho2
        }
        which_anho
}

#incs ocurridos en el mes de enero de cualquier año
# o entre los meses de enero y mayo de cualquier año
# o en mayo y anteriores o mayo y posteriores
which_mes <- function(mes1, mes2 =mes1, y="exact") {



        if(mes1 == mes2){
            if(y=="exact"){
                which_mes <- incendios$MES == mes1
            }
            if(y=="ypost"){
                which_mes <- incendios$MES >= mes1
            }
            if(y=="yant"){
                which_mes <- incendios$MES <= mes1
            }

        }
        else{
            which_mes <- incendios$MES >= mes1 & incendios$MES <= mes2
        }
        which_mes
}

# incs ocurridos en la semana 1 de cualquier año
which_semana <- function(sem1, sem2 = sem1, y="exact") {

        if(sem1 == sem2){
            if(y=="exact"){
                which_semana <- incendios$SEMANA == sem1
            }
            if(y=="ypost"){
                which_semana <- incendios$SEMANA >= sem1
            }
```

```
            if(y=="yant"){
                which_semana <- incendios$SEMANA <= sem1
            }

        }
        else{
            which_semana <- incendios$SEMANA >= sem1 & incendios$SEMANA <= sem2
        }
        which_semana
    }



    #incs ocurridos en 5 de julio de cualquier año
    # o del 5 de julio al 7 de agosto
    # o 5 julio y posteriores o 5 julio y anteriores
    which_diaMes <- function(dia1, mes1, dia2 = dia1, mes2 = mes1, y) {

        if(dia1 == dia2 & mes1 == mes2){
            if(y=="exact"){
                which_dia <- incendios$MES == mes1 & incendios$DIA == dia1

            }
            if(y=="ypost"){
                which_dia <- incendios$MES > mes1 | (incendios$DIA >= dia1 & incendios$MES == mes1)

            }
            if(y=="yant"){
                which_dia <- incendios$MES < mes1 | (incendios$DIA <= dia1 & incendios$MES == mes1)

            }

        }
        else{
            if(mes1 == mes2){
                which_dia <- (incendios$DIA >= dia1 & incendios$DIA <= dia2 & incendios$MES == mes2)
            }
             else{
            which_dia <- (incendios$MES > mes1 & incendios$MES < mes2) | (incendios$DIA >= dia1 &
incendios$MES == mes1) | (incendios$DIA <= dia2 & incendios$MES == mes2)
            }

        }
        which_dia
    }



    #incs ocurridos en enero de 2000
    which_mesAnho <- function(mes1, anho1, mes2 = mes1, anho2 = anho1, y){
        if(mes1 == mes2 & anho1 == anho2){
        if(y=="exact"){
```

```
            which_mesAnho <- incendios$MES == mes1 & incendios$ANHO == anho1

        }
        if(y=="ypost"){
            which_mesAnho <- incendios$MES >= mes1 & incendios$ANHO == anho1

        }
        if(y=="yant"){
            which_mesAnho <- incendios$MES <= mes1 & incendios$ANHO == anho1
        }
        }
        else{
            which_mesAnho <- (incendios$ANHO == anho1 & incendios$MES >= mes1) | (incendios$ANHO > anho1 & incendios$ANHO < anho2) | (incendios$ANHO == anho2 & incendios$MES <= mes2)
        }

        which_mesAnho
    }

    #incs ocurridos en 5 de julio de 2001
    #o posteriores a 5 de julio de 2001 (hasta el 31 de diciembre de 2001)
    #o anteriores (hasta 1 enero de 2001)
    #o entre 5 de julio de 2001 y 2 de agosto de 2002
    which_diaMesAnho <- function(dia1, mes1, anho1, dia2 = dia1, mes2 = mes1, anho2 = anho1, y) {

        if(dia1 == dia2 & mes1 == mes2 & anho1 == anho2) {

            if(y=="exact"){

                which_diaMesAnho <- incendios$DIA == dia1 & incendios$MES == mes1 & incendios$ANHO == anho1
            }
            if(y=="ypost"){

                which_diaMesAnho <- incendios$ANHO == anho1 & ((incendios$MES > mes1) | (incendios$MES == mes1 & incendios$DIA >= dia1))

            }
            if(y=="yant"){

                which_diaMesAnho <- incendios$ANHO == anho1 & ((incendios$MES < mes1) | (incendios$MES == mes1 & incendios$DIA <= dia1))

            }

        }
        else {
            if(anho1 == anho2){
            if(mes1 == mes2){
```

```
                    which_diaMesAnho <- ((incendios$DIA >= dia1 & incendios$DIA <= dia2 & incendios$MES ==
mes2)) & (incendios$ANHO >= anho1 & incendios$ANHO <= anho2)
                }
                else{

                    which_diaMesAnho <- ((incendios$MES > mes1 & incendios$MES < mes2) | (incendios$DIA >=
dia1 & incendios$MES == mes1) | (incendios$DIA <= dia2 & incendios$MES == mes2)) & (incendios$ANHO >= anho1 &
incendios$ANHO <= anho2)
                }
                }
                else{
                    which_diaMesAnho <- (incendios$ANHO == anho1 & ((incendios$MES > mes1) | (incendios$MES
== mes1 & incendios$DIA >= dia1))) | (incendios$ANHO > anho1 & (incendios$ANHO < anho2)) | (incendios$ANHO ==
anho2 & ((incendios$MES < mes1) | (incendios$MES == mes1 & incendios$DIA <= dia1)))
                }
            }
            which_diaMesAnho
        }



        #incs detectados a las 13h
        #o entre las 13h y las 16h
        #o 13 y anteriores o 13 y posteriores
        which_hora <- function(hora1, hora2 = hora1, y="exact") {

            if(hora1 == hora2){
                if(y=="exact"){
                    which_hora <- incendios$HORA == hora1
                }
                if(y=="ypost"){
                    which_hora <- incendios$HORA >= hora1
                }
                if(y=="yant"){
                    which_hora <- incendios$HORA <= hora1
                }

            }
            else{
                which_hora <- incendios$HORA >= hora1 & incendios$HORA <= hora2
            }
            which_hora
        }

        #incs por dia de la semana
        which_diaSemana <- function(day){
            if(day == 7){
                day <- 0
            }
            which_diaSemana <- wday(incendios$DHDETECCION) -1 == day
            which_diaSemana
```

```
}



#incs dependiendo de la clase de dia
#1 festivo
#2 sabado
#3 laborable vispera de festivo
#4 laborable
which_clasedia <- function(clase) {

      which_clase <- incendios$IDCLASEDIA == clase
      which_clase
}

#devuelve fechas validas de los registros
which_fechasvalidas <- function(x){
      index_v <- as.logical()
      for( i in 1:nrow(incendios)){
            index <- ifelse( str_length(x[i]) == 19, TRUE,FALSE)
            index_v <- append(index_v, index)
      }
      index_v
}

inicilizar <- function(){
fechasDetValidas <- which_fechasvalidas(incendios$DHDETECCION)
fechasLleValidas <- which_fechasvalidas(incendios$DHLLEGADA)
}
                                    ####### WHICH POR CAUSA #######
#incs con grupocausa = gcausa
      # rayo <- 1
      #negligencias <- 2
      #causas accidentales <-3
      #intencionado <- 4
      #causa desconocida <- 5
      #incendio reproducido <- 6
which_grupocausa <- function(gcausa) {

      which_gcausa <- incendios$IDGRUPOCAUSA == gcausa
      which_gcausa
}

#incs cuya causa es segura o supuesta
      #segura <- 1
      #supuesta <- 2
which_causasegura <- function(seguracausa=0) {

      which_seguracausa <- incendios$IDCAUSA == seguracausa
      which_seguracausa
```

```
}

#incs cuya causa concreta fue: 50 (fumadores)
which_causaconcreta <- function(gcausa, causa){

	which_causaconc <- incendios$IDCAUSAS == causa & incendios$IDGRUPOCAUSA == gcausa
	which_causaconc
}

#1 identificado
#2 no identificado
which_causante <- function(causante){

	which_causante <- incendios$IDCAUSANTE == causante
	which_causante
}

#0 <- sin datos, 1-23 motivaciones, 99 <- otras
#incs cuya motivación fue 3 (venganza)
which_motivacion <- function(motivacion){

	which_motivacion <- incendios$IDMOTIVACION == motivacion
	which_motivacion
}

                                        ####### WHICH POR METEO #######
#incs que empezaron con un nivel de peligro 1(mínimo) :4(máximo)
which_peligro <- function(pel) {

	which_pel <- meteoOrd$IDPELIGRO == pel
	which_pel
}


#x: valor
#var: "dsinlluvia", "tempmax", "hrelativa", "vviento", "pignicion"
#mode:menos, exact, mas
which_temp <- function(x, var, mode="exact"){
	if(var == "dsinlluvia"){
		if(mode == "exact"){
			which_temp <- meteoOrd$DULLUVIA == x
		}
		if(mode == "mas"){
			which_temp <- meteoOrd$DULLUVIA >= x
		}
		if(mode == "menos"){
			which_temp <- meteoOrd$DULLUVIA <= x
		}
	}
	if(var == "tempmax"){
		if(mode == "exact"){
```

```
            which_temp <- meteoOrd$TEMPMAX == x
        }
        if(mode == "mas"){
            which_temp <- meteoOrd$TEMPMAX > x
        }
        if(mode == "menos"){
            which_temp <- meteoOrd$TEMPMAX <= x
        }
    }
    if(var == "hrelativa"){
        if(mode == "exact"){
            which_temp <- meteoOrd$HRELATIVA == x
        }
        if(mode == "mas"){
            which_temp <- meteoOrd$HRELATIVA > x
        }
        if(mode == "menos"){
            which_temp <- meteoOrd$HRELATIVA <= x
        }
    }
    if(var == "vviento"){
        if(mode == "exact"){
            which_temp <- meteoOrd$VVIENTO == x
        }
        if(mode == "mas"){
            which_temp <- meteoOrd$VVIENTO >= x
        }
        if(mode == "menos"){
            which_temp <- meteoOrd$VVIENTO <= x
        }
    }

    if(var == "pignicion"){
        if(mode == "exact"){
            which_temp <- meteoOrd$PIGNICION == x
        }
        if(mode == "mas"){
            which_temp <- meteoOrd$PIGNICION >= x
        }
        if(mode == "menos"){
            which_temp <- meteoOrd$PIGNICION <= x
        }
    }

    which_temp
}

#incs con vientos predominantes entre dir1 y dir2
which_dirviento <- function(dir1, dir2){

    if(dir1 == 0 & dir2 == 360){
```

```
            which_dirv <- meteoOrd$DVIENTO >= 0 & meteoOrd$DVIENTO <= 360
        } else
        if(dir1 == 0 & dir2 == 0 | dir1 == 360 & dir2 == 360){

            which_dirv <- meteoOrd$DVIENTO == 0 | meteoOrd$DVIENTO == 360
        }

            if (dir1 > dir2) {
            which_dirv <- meteoOrd$DVIENTO >= dir1 | meteoOrd$DVIENTO <= dir2
        }
        else if(dir1 < dir2){
        which_dirv <- meteoOrd$DVIENTO >= dir1 & meteoOrd$DVIENTO <= dir2
        }
        which_dirv
        }

#incs con el combustible indicado (matorrales, bosque..)
which_combustible <- function(comb){

        which_comb <- meteoOrd$IDMODELOCOMBUSTIBLE == comb
        which_comb
}

                                    ####### WHICH POR DETECCION #######
#incs detectados por: 2 (vigilante fijo)
which_detectadopor <- function(detec){

        which_detect <- incendios$IDDETECTADOPOR == detec
        which_detect
}

#incs iniciados junto a: 1 (carretera)
which_iniciadojuntoa <- function(juntoa) {

        which_juntoa <- incendios$IDINICIADOJUNTOA == juntoa
        which_juntoa
}


                                    ####### WHICH POR ACTUACION #######

#incs que tardaron más o menos de time horas entre
#su deteccion, llegada, control o extincion
#time: tiempo en horas
#var1, var2: "det", "lle", "ctr", "ext"
#mode, "mas", "menos"
which_tiempoactuacion <- function(time, var1, var2, mode="mas"){

        if(var1 == "det" & var2 == "lle"){
            if(mode == "mas"){
```

```
                    which_time <- difftime(incendios$DHLLEGADA, incendios$DHDETECCION, units = "hours") >=
time
                }
                if(mode == "menos"){
                    which_time <- difftime(incendios$DHLLEGADA, incendios$DHDETECCION, units = "hours") <=
time
                }

            }
            if(var1 == "det" & var2 == "ctr"){
                if(mode == "mas"){
                    which_time <- difftime(incendios$DHCONTROLADO, incendios$DHDETECCION, units =
"hours") >= time
                }
                if(mode == "menos"){
                    which_time <- difftime(incendios$DHCONTROLADO, incendios$DHDETECCION, units =
"hours") <= time
                }

            }
            if(var1 == "det" & var2 == "ext"){
                if(mode == "mas"){
                    which_time <- difftime(incendios$DHEXTINCION, incendios$DHDETECCION, units = "hours")
>= time
                }
                if(mode == "menos"){
                    which_time <- difftime(incendios$DHEXTINCION, incendios$DHDETECCION, units = "hours")
<= time
                }

            }
            if(var1 == "lle" & var2 == "ctr"){
                if(mode == "mas"){
                    which_time <- difftime(incendios$DHCONTROLADO, incendios$DHLLEGADA, units = "hours")
>= time
                }
                if(mode == "menos"){
                    which_time <- difftime(incendios$DHCONTROLADO, incendios$DHLLEGADA, units = "hours")
<= time
                }

            }
            if(var1 == "lle" & var2 == "ext"){
                if(mode == "mas"){
                    which_time <- difftime(incendios$DHEXTINCION, incendios$DHLLEGADA, units = "hours") >=
time
                }
                if(mode == "menos"){
                    which_time <- difftime(incendios$DHEXTINCION, incendios$DHLLEGADA, units = "hours") <=
time
                }
```

```
        }
        if(var1 == "ctr" & var2 == "ext"){
            if(mode == "mas"){
                which_time <- difftime(incendios$DHEXTINCION, incendios$DHCONTROLADO, units = "hours") >= time
            }
            if(mode == "menos"){
                which_time <- difftime(incendios$DHEXTINCION, incendios$DHCONTROLADO, units = "hours") <= time
            }

        }
        which_time

    }


                                    ####### WHICH POR TIPO #######
    #1 de superficie, 2 de copas, 3 subsuelo, 4 sup y copas, 5 sup y sub, 6 cop y sub, 7 sup,cop y sub.
    which_tipofuego <- function(tipo){

        which_tipo <- incendios$TIPOFUEGO == tipo
        which_tipo
    }
```

Plot.r

```
    gen_plot_gen <- function(df){

        var <- (deparse(substitute(df)))
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/GENERAL/", var, sep = ""))
        xlabel <- strsplit(var,"_")[[1]][2]
        for(col in 1:ncol(df)) {
            media <- data.frame( x = c(-Inf, Inf), y = mean(df[,col], na.rm = T), media = factor(500) )

        g <- ggplot(df, aes_string(x = paste("seq_along(",names(df)[col],")"), y = as.name(names(df)[col]))) + geom_bar(stat="identity", width = 0.9) + xlab(xlabel) +geom_line(aes( x, y, linetype = media ), media, color="red")
          ggsave(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/GENERAL/", var,"/", names(df)[col],".png" ,sep="") ,g)
        }
    }

    gen_plot_nev <- function(prov,muni,param){

        var <- paste(prov, "_", muni , "_", param, sep = "")
        df <- (get(var))
        print(var)
```

```
        if(nrow(df) == 25){
            df <- df[-c(24,25),]
        }
        split <- strsplit(var, "_")
        provmuni <- paste(split[[1]][1],"_", split[[1]][2], sep = "")
        param <- paste(split[[1]][3],"_", split[[1]][4], sep = "")
        dir.create("C:/Pablo/Proyecto_UAB_Incendios/images_R/CONCELLOS/")
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/CONCELLOS/", provmuni, sep = ""))
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/CONCELLOS/", provmuni,"/",param, sep = ""))
        xlabel <- strsplit(var,"_")[[1]][2]
        print(df)
        for(col in 1:ncol(df)) {
            media <- data.frame( x = c(-Inf, Inf), y = mean(df[,col], na.rm = T), media = factor(500) )

            g <-  ggplot(df, aes_string(x = paste("seq_along(",names(df)[col],")"), y = as.name(names(df)[col]))) + geom_bar(stat="identity", width = 0.9) + xlab(xlabel)  +geom_line(aes( x, y, linetype = media ), media, color="red")  + scale_x_continuous(breaks = c(1:nrow(df)), labels = c(1:nrow(df)))
            ggsave(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/CONCELLOS/", provmuni,"/", param ,"/",names(df)[col],".png" ,sep="") ,g)
        }
    }


    norm_plot_norm <- function(df){

        var <- (deparse(substitute(df)))
        print(var)
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/NORMALIZADO/", var, sep = ""))
        print(var)
        xlabel <- strsplit(var,"_")[[1]][2]
        for(col in 1:ncol(df)) {
            media <- data.frame( x = c(-Inf, Inf), y = mean(df[,col], na.rm = T), media = factor(500) )
            print("a")
            g <-  ggplot(df, aes_string(x = paste("seq_along(",names(df)[col],")"), y = as.name(names(df)[col]))) + geom_bar(stat="identity", width = 0.9) + xlab(xlabel)  +geom_line(aes( x, y, linetype = media ), media, color="red")
            ggsave(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/NORMALIZADO/", var,"/", names(df)[col],".png" ,sep="") ,g)
        }

    }




    ## OTROS ##

    plot_meses_anhos <- function(df){

        var <- (deparse(substitute(df)))
```

```
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/GENERAL/", var, sep = ""))
        xlabel <- strsplit(var,"_")[[1]][2]
        seq <- rep(1:12,nrow(general_mesAnho)%%12, len=nrow(general_mesAnho))
        for(col in 1:ncol(df)) {
            media <- data.frame( x = c(-Inf, Inf), y = mean(df[,col], na.rm = T), media = factor(500) )

            g <-  ggplot(df, aes_string(x = paste("seq_along(",names(df)[col],")"), y = as.name(names(df)[col]))) + geom_bar(stat="identity", width = 0.9, colour= seq) + xlab(xlabel)  +geom_line(aes( x, y, linetype = media ), media, color="red")
            ggsave(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/GENERAL/", var,"/", names(df)[col],".png" ,sep="") ,g)
        }
    }


    plot_prov_ninc_ratio <- function(){
        dfm <- melt(tb.provincias_ninc_ratio, id.vars=c("PROVINCIA"))
        g <-ggplot(dfm, aes(x=as.numeric(variable), y=value, colour = PROVINCIA)) + geom_line()
        g <- g + ggtitle("RATIO Nº DE INCENDIOS POR PROVINCIA / SUPERFICIE")
        g

    }

    plot_prov_haq_ratio <- function(){
        dfm <- melt(tb.provincias_haq_ratio, id.vars=c("PROVINCIA"))
        g <-ggplot(dfm, aes(x=as.numeric(variable), y=value, colour = PROVINCIA)) + geom_line()
        g <- g + ggtitle("RATIO HA QUEMADAS POR PROVINCIA / SUPERFICIE")
        g

    }

    plot_muni_ninc_ratio <- function(){
        dfm <- melt(tb.municipios_ninc_ratio, id.vars=c("MUNICIPIO"))
        g <-ggplot(dfm, aes(x=as.numeric(variable), y=value, colour = MUNICIPIO)) + geom_line()
        g <- g + ggtitle("RATIO Nº DE INCENDIOS POR MUNICIPIO / SUPERFICIE") + ylim(0,0.025)
        g

    }


    plot_prov_ninc_mas_xha_ratio <- function(x){

        cadena <- paste("RA_SUP_N_", x, "ha", sep = "")
      g <-  ggplot(tb.provincias_ninc_ratio, aes_string(x="PROVINCIA", y = cadena)) + geom_bar(stat="identity", width = 1) + xlab("MES_ANHO")
            g

    }


    plot_muni_ninc_mas_xha_ratio <- function(x){
```

```
        cadena <- paste("RA_SUP_N_", x, "ha", sep = "")
        cutoff <- data.frame( x = c(-Inf, Inf), y = mean(tb.municipios_ninc_ratio[,cadena], na.rm = T), cutoff = factor(50) )
        g <- ggplot(tb.municipios_ninc_ratio, aes_string(x="MUNICIPIO", y = cadena)) + geom_bar(stat="identity", width = 1) +geom_line(aes( x, y, linetype = cutoff ), cutoff, color="red")

            g

    }


    plot_muni_haq_mas_xha_ratio <- function(x){
        cadena <- paste("RA_SUP_HA_", x, "ha", sep = "")
        cutoff <- data.frame( x = c(-Inf, Inf), y = mean(tb.municipios_haq[,cadena], na.rm = T), cutoff = factor(50) )
        g <- ggplot(tb.municipios_haq, aes_string(x="MUNICIPIO", y = cadena)) + geom_bar(stat="identity", width = 1) +geom_line(aes( x, y, linetype = cutoff ), cutoff, color="red")

        g

    }

    plot_muni_1ha_haq_ninc_ratio <- function(){

        cutoff <- data.frame( x = c(-Inf, Inf), y = mean(tb.muni_1ha$RA_HAQ_NINC, na.rm = T), cutoff = factor(50) )
        g <- ggplot(tb.muni_1ha, aes_string(x="MUNICIPIO", y = "RA_HAQ_NINC")) + geom_bar(stat="identity", width = 1) +geom_line(aes( x, y, linetype = cutoff ), cutoff, color="red")

        g

    }


    plot_muni_gcausa_porc <- function(num){
        numS <- paste("C", num, sep = "")
        cutoff <- data.frame( x = c(-Inf, Inf), y = mean(tb.municipios_gcausa_porc[,numS], na.rm = T), cutoff = factor(50) )
        g <- ggplot(tb.municipios_gcausa_porc, aes_string(x="MUNICIPIO", y = numS)) + geom_bar(stat="identity", width = 1) +geom_line(aes( x, y, linetype = cutoff ), cutoff, color="red")

        g

    }

    #
    plot_muni_gcausa_superficie <- function(num){
        numS <- paste("C", num, sep = "")
        cutoff <- data.frame( x = c(-Inf, Inf), y = mean(tb.municipios_gcausa_superf[,numS], na.rm = T), cutoff = factor(50) )
        g <- ggplot(tb.municipios_gcausa_superf, aes_string(x="MUNICIPIO", y = numS)) + geom_bar(stat="identity", width = 1) +geom_line(aes( x, y, linetype = cutoff ), cutoff, color="red")
```

```
    g

}

plot_gcausas_meses_porc <- function(num){

    str <- paste("C", num, sep = "")
    str1 <- paste("seq_along(",str,")",sep = "")
    g <- ggplot(gcausas_meses_ninc_porc, aes_string(x=str1 , y=str)) + geom_bar(stat="identity") + xlab('id')
    g
}

plot_causas_meses_porc <- function(num){
    x <- c(1:12)
    xrange <- range(x)
    y <- c(0:100)
    yrange <- range(y)
    plot(xrange, yrange)
    colors <- rainbow(6)

    for (i in 1:6) {
        str <- paste("C", i, sep = "")
        y <- gcausas_meses_ninc_porc[[str]]
        lines(x, y, col=colors[i])
    }
}



plot_mapa_municipios_HAQ_SUP <- function(){
    #rojo significa muchas haq/sup y azul significa pocas.
    rbPal <- colorRampPalette(c('green', 'red'))
   MMOcol <- rbPal(10)[as.numeric(cut(as.numeric(MuniMergedOrdered$`RATIO_HAQ/SUP`),breaks = 20))]
    g <- ggplot(MuniMergedOrdered, aes(x=X, y=Y)) + geom_point(colour = MMOcol, size=3)

        g
}


plot_mapa_municipios_HAQ_SUP <- function(table, var){
    #rojo significa muchas haq/sup y azul significa pocas.

    n <- 10
    rbPal <- colorRampPalette(c('green', 'red'))
    media <- mean(table[[var]], na.rm = T)
    print(media)
    max <- max(table[[var]], na.rm = T)
    breaks1 <- seq(from = -0.1, to = media, length.out = n)
    breaks2 <- seq(from = media , to = max,length.out = n)
    breaks <- c(breaks1[-n], breaks2)
```

```
        MMOcol <- rbPal(n*2-1)[as.numeric(cut(as.numeric(table[[var]]),breaks ))]
        g <- ggplot(table, aes(x=X, y=Y)) + geom_point(colour = MMOcol, size=6) + ggtitle(var)

        g
    }

    #guarda varios plots con cada municipio mas rojo o verde segun años, causas...
    #wh es para tener en cuenta solo incendios de más de x ha.
    save_plot_municipios_whichs <- function(table,Nvar,wh = 0){

        vars <- c("RATIO_HAQ/NINC", "RATIO_NINC/SUP", "RATIO_HAQ/SUP")

        for(i in 1:length(vars)){
            plot_mapa_municipios_NINC_SUP(table, Nvar, vars[i])
        }

    }


    plot_mapa_municipios_NINC_SUP <- function(table, Nvar, var){
        library(lubridate)
        #rojo significa muchas haq/sup y azul significa pocas.
        rbPal <- colorRampPalette(c('green', 'red'))
        n <- 10

        media <- mean(table[[var]])
        print(media)
        max <- max(table[[var]])
        breaks1 <- seq(from = -0.1, to = media, length.out = n)
        breaks2 <- seq(from = media , to = max,length.out = n)
        breaks <- c(breaks1[-n], breaks2)

        MMOcol <- rbPal(n*2-1)[as.numeric(cut(as.numeric(table[[var]]),breaks ))]
        g <- ggplot(table, aes(x=X, y=Y)) + geom_point(colour = MMOcol, size=3)
        tablename <- deparse(substitute(table))
        varname <- gsub("/", "_" ,var)
        dir.create("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/municipios/")
        anho <- strsplit(Nvar, "_")[[1]][2]
        num <- strsplit(Nvar, "_")[[1]][3]
        dir.create(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/municipios/", anho, sep = ""))
        ggsave(paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/municipios/",anho,"/",  varname,  "_",
anho,"_", num,".png" ,sep="") ,g)

    }

    plot_mapa_incendios_which_causa <- function(wh = T){

        tabla <- incendios[wh,]
```

```
        p <- ggplot(tabla, aes(x=X, y=Y, col = factor(IDCAUSAS, labels= c(2,3,12:13,15:16,22,28,30)))) + geom_point(size=3)+  xlim(505000,528000) + ylim(4755000,4775000) + labs(color = "CAUSA", c(1:9))
        p
    }

    plot_mapa_incendios_which_motivacion <- function(wh = T){

        tabla <- incendios[wh,]

        p <- ggplot(tabla, aes(x=X, y=Y, col = factor(IDMOTIVACION, labels= c(1,2))))  + labs(color = "CAUSA", c(1:25)) + geom_point(size=3) + xlim(650000,675000) + ylim(4640000,4660000)
        p
    }
    # te pinta el nincs por mes y año coloreandote los meses si le pasas 15_11_general_mesAnho
    plot_anhosmeses_colores <- function(param){

        tb <- cbind(param, "MES" =  c(1:12))
    tb
        g <- ggplot(tb, aes(seq_along(N_INC), N_INC, col = factor(MES, labels=meses_vector))) + geom_bar(stat = "identity")+ labs(color = "MES") + scale_x_discrete(name="Time", breaks=seq(1,180,12), labels=c(2000:2014))
        g
    }

    #guarda varios plots con la localizacion de incendios segun años, causas...
    #wh es para tener en cuenta solo incendios de más de x ha.
    plot_mapa_incendios <- function(wh = 0){

        if(!(wh == "")){
            which <- which_masmenos_xha(wh)
            incendiosmeteo <- incendios[which,]
            }


        #  params <- names(incendiosmeteo)[c(12)]
            params <- names(incendiosmeteo)[c(13:21,25,26,27)]

            for(i in 1:length(params)){

                param <- params[i]

            Sfactor <- paste("factor", param, sep = "")
            factor <- get(Sfactor)
            plot_list = list()

            for(i in 1:length(factor)){

                p <- ggplot(incendiosmeteo[incendiosmeteo[[param]] == as.integer(factor[i]),], aes(x=X, y=Y)) + geom_point(size=1)+ xlim(460000,710000) + ylim(4610000,4860000) + ggtitle(paste(param,factor[i]))
                plot_list[[i]] = p

            }
```

```
                # Save plots to tiff. Makes a separate file for each plot.
                folder <- paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/" , sep = "")
                subfolder <- paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/+", wh,"ha/" , sep = "")
                subfolder2 <- paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/+", wh,"ha/", param, "/",
sep = "")
                dir.create(folder)
                dir.create(subfolder)
                dir.create(subfolder2)

                for (i in 1:length(factor)) {
                    file_name = paste("C:/Pablo/Proyecto_UAB_Incendios/images_R/mapas/+", wh,"ha/", param, "/",
param, factor[i], ".png", sep="")
                    png(file_name)
                    print(plot_list[[i]])
                    dev.off()
                }
                }
```

Meteogalicia.r

```
getAllfromEst <- function(est=10087){
    for(j in 5:5){
        if (j == 1){
            param <- "tempmax"
        }
        if (j == 2){
            param <- "hrelativa"
        }
        if (j == 3){
            param <- "vviento"
        }
        if (j == 4){
            param <- "dirviento"
        }
        if (j == 5){
            param <- "chuvia"
        }
    for(anho in 2004:2014){
            var <- paste(param,anho, sep= "")
            assign(var,getFinalTable(anho,param,est=10087),.GlobalEnv)
        }
    }
    hrelativa2013
}
```

```
#CONSIGUE LOS DATOS SOBRE UN PARÁMETRO DE UNA ESTACION METEOROLÓGICA DE CADA DIA DE UN AÑO DADO
getFinalTable <- function(anho, param, est = 10087){
    URL <- crearUrl(est,anho, param)
    file <- downloadFile(URL, est, anho, param)
    raw_table <- read.csv(file, header = F, blank.lines.skip = T, sep= "", stringsAsFactors = F)
    final_table <- arreglarMeteoCSV(raw_table, param)
    final_table
}

#consigue la url de lo que queremos obtenter de la web. crearUrl(0,2000,"vviento")
crearUrl <- function(est, anho, param){
    codParam <- know_codParam(param)
    fileUrl <- paste("http://www2.meteogalicia.es/galego/observacion/estacions/historicosAtxt/DatosHistoricosTaboas_diarioAFicheiro.asp?est=",est,"&param=",codParam , "&data1=1/1/",anho,"&data2=1/1/",anho+1,"&tiporede=automaticas", sep = "")
    fileUrl
}

downloadFile <- function(URL, est, anho, param){

    folder <- paste("meteogalicia/", est,"/", sep = "")
    dir.create(folder, showWarnings = F)
    file <- paste(folder,param,anho,".csv" , sep = "")
    download.file(URL, destfile = file)
    file

}


#devuelve codigo del parámetro a partir de un string. x ej: "tempmax" devuelve 84
know_codParam <- function(param){

    codParam = 0

    if(param == "tempmax"){
        codParam <- 84
    }
    if(param == "hrelativa"){
        codParam <- 86
    }
    if(param == "vviento"){
        codParam <- 81
    }
    if(param == "dirviento"){
        codParam <- 10124
    }
    if(param == "chuvia"){
        codParam <- 10001
    }
```

```
    codParam

}



arreglarMeteoCSV <- function(raw_table, param){
    library(stringr)
    fixed_table <- raw_table

    if(param == "tempmax"){
    fixed_table <- fixed_table[-c(1:16),]
    row.names(fixed_table) <- c(1:nrow(fixed_table))
    for(i in seq(2, nrow(fixed_table), by = 2)){
        fixed_table[i-1,6] <- fixed_table[i,1]

    }
    }
    if(param == "vviento" || param == "hrelativa"){
        fixed_table <- fixed_table[-c(1:16),]
        row.names(fixed_table) <- c(1:nrow(fixed_table))
        for(i in seq(2, nrow(fixed_table), by = 2)){
            fixed_table[i-1,6] <- fixed_table[i,2]

        }

    }
    if(param == "dirviento"){
        fixed_table <- fixed_table[-c(1:16),]
        row.names(fixed_table) <- c(1:nrow(fixed_table))
        for(i in seq(2, nrow(fixed_table), by = 2)){
            fixed_table[i-1,6] <- fixed_table[i,3]

        }

    }
    if(param == "chuvia"){

        fixed_table <- fixed_table[-c(1:16),]
        print(raw_table)
        row.names(fixed_table) <- c(1:nrow(fixed_table))
        for(i in 1:nrow(fixed_table)){
            fixed_table[i,6] <- fixed_table[i,5]

        }
    }
    fixed_table <- fixed_table[-seq(2, nrow(fixed_table), by = 2),]
    row.names(fixed_table) <- c(1:nrow(fixed_table))
    if(nrow(fixed_table)==365){
```

```
            row2902 <- fixed_table[1,]
            row2902[1] <- 9
            anho <- str_sub(ene1, start = -4)
            row2902[2] <- paste("29/02/", anho, sep = "")
            row2902[6] <- -9999
            fixed_table <- rbind(fixed_table[1:59,], row2902, fixed_table[60:365,])

        }



        names(fixed_table)[6] <- paste(fixed_table[1,3])

        row.names(fixed_table) <- c(1:nrow(fixed_table))

        fixed_table <- fixed_table[,c(1,2,6)]

        final_table <- NaNValues(fixed_table)

        final_table[,3] <- as.numeric(gsub(",",".",final_table[,3]))

        final_table

    }

    NaNValues <- function(fixed_table){

        for(i in 1:nrow(fixed_table)){
           if(fixed_table[i,3] == -9999){
              fixed_table[i,3] = NaN
        }
        }
        fixed_table
    }


    meanValues <- function(final_table){

        round(mean(final_table[,3], na.rm = T), digits = 2)

    }



    #esta en Ninc1Meteopordia
    AllMeteoVarsPerDay <- function(prov, muni, ninc1 = T){
        library(lubridate)
          all <- data.frame(a = numeric(0), b = numeric(0), c = numeric(0), d=numeric(0), e = numeric(0), f=
numeric(0), g= numeric(0), h=numeric(0))
          NAs <- data.frame(matrix(NA, nrow = 366, ncol = 6))
```

```
    vars <- data.frame(matrix(NA, nrow = 366, ncol = 6))
    vectorMeses <- crearSeqMeses()


    cont <- 0
    for(k in 1:15){
        tempanho <- get(paste("tempmax", k+1999, sep = ""))
        hrelanho <- get(paste("hrelativa", k+1999, sep = ""))
        vvientoanho <- get(paste("vviento", k+1999, sep = ""))
        dirvientoanho <- get(paste("dirviento", k+1999, sep = ""))
        chuviaanho <- get(paste("chuvia", k+1999, sep=""))
        if(nrow(tempanho) == 366 & nrow(hrelanho) == 366){
    vars[,1] <- as.numeric(tempanho[,3])
    vars[,2] <- as.numeric(hrelanho[,3])
    vars[,3] <- as.numeric(vvientoanho[,3])
    vars[,4] <- as.numeric(dirvientoanho[,3])
    vars[,5] <- as.numeric(chuviaanho[,3])
    vars[,6] <- vectorMeses

    all <- rbind(all, vars)}
        else{
            cont <- cont + 1
            all <- rbind(all, NAs)

        }


    }

    assign("cont",cont,.GlobalEnv)

###aqui estaa
    all <- IncTempHrel_stc(all, prov, muni, ninc1)

    names(all)[7] <- "Mes"
   vectorMesesFactor <- MapMeses(all)

      vectorMesesOrdered <- sort(vectorMesesFactor)
    all[,7] <- (vectorMesesOrdered)
    all[,7] <- factor(all[,7], c(1:12))
      all[,7] <- as.numeric(all[,7])



    all[,8] <- crearSeqSemanas()
    names(all)[8] <- "Semana"
  vectorSemanasFactor <- MapSemanas(all)
   vectorSemanasOrdered <- sort(vectorSemanasFactor)
  all[,8] <- (vectorSemanasFactor)
  all[,8] <- factor(all[,8], c(1:53))
   all[,8] <- as.numeric(all[,8])
```

```
        vectorAnhoFactor <- MapAnhos()
        vectorAnhoOrdered <- sort(vectorAnhoFactor)
        all[,9] <- (vectorAnhoFactor)
        all[,9] <- factor(all[,9], c(2000:2014))
        all[,9] <- as.numeric(all[,9])
        all[,10] <- crearSeqDates()
        vectorCausas <- buscaCausas(all, prov, muni)
        all[,11] <- factor(vectorCausas, c(0:6))
        if(ninc1 == T){
        #all[,1] <- factor(all[,1])
        }
        names(all[11]) <- "Causa"
        all <- all[complete.cases(all),]

         all

}

crearSeqMeses <- function(){
      vectorMeses <- numeric()
      for(i in 1:12){
            diasenmes <- as.numeric(dias_en_mes(i,2000))
            vectorMeses <- c(vectorMeses,  rep(i,diasenmes))
      }
      vectorMeses
}

crearSeqSemanas <- function(){

      vectorSemanas <- numeric()
      vectorTotal <- numeric()
      for(i in 1:53){

            vectorSemanas <- c(vectorSemanas,  rep(i,7))
      }
      vectorSemanas <- vectorSemanas[1:366]

      vectorSemanas
}

crearSeqDates <- function(){
      seqDatesTotal <- numeric(0)
      class(seqDatesTotal) <- "Date"
      for(i in 2000:2014){

            seqDates <- seq.Date(from = as.Date(paste(i ,"-1-1", sep = "")), to   = as.Date(paste(i ,"-12-31", sep =
"")), by = 1)
            if((i %% 4)){
                  feb <- as.Date(paste(i ,"-2-28", sep = ""))
                  seqDates <- c(seqDates[c(1:59)], feb, seqDates[c(60:365)])
```

```
                # print(i)
            }
            # print(seqDates)
            seqDatesTotal <- c(seqDatesTotal, seqDates)
        }
        as.Date(seqDatesTotal)

}

buscaCausas <- function(table, prov = 0, muni = 0){
        vectorCausas <- numeric(0)
        for ( i in 1:nrow(table)){
            causa <- 0
        if(table$N_INC[i] == 1){
            fecha <- table$V10[i]
            dia <- day(fecha); mes <- month(fecha); anho <- year(fecha)
            if(prov != 0 && muni != 0){
            which <- which_municipio(prov,muni) & which_diaMesAnho(dia,mes,anho, y= "exact")
            }
            else{
                which <-  which_diaMesAnho(dia,mes,anho, y= "exact")

            }
            index <- which(which)[1]
            causa <- incendios$IDGRUPOCAUSA[index]
        }
            vectorCausas <- c(vectorCausas, causa)
        }
        vectorCausas
}

#meteo por dias
IncTempHrel_stc <- function(vars, prov, muni, ninc1){

        incPD <- data.frame(a = numeric(0), b = numeric(0), c = numeric(0))

        simpgeneral <- paste("simp_general_diaMesAnho", "_", prov, "_" ,muni, sep = "")
        general_diaMesAnho <- get(simpgeneral)
        incPD <- cbind(as.integer(general_diaMesAnho$N_INC), vars)

        if(ninc1 == T){
            incPD <- NincTo1(incPD)
        }

        names(incPD) <- c("N_INC", "TempMax", "Hrelativa", "VViento", "DirViento","Chuvia")
        incPD
}



#pasa nº d incendios a 1
```

```
NincTo1 <- function(incPD){

    incPD[,1] <- sign(incPD[,1])
    incPD
}


tablaPuente <- function() {
    library(lubridate)
    incPuente <- data.frame(matrix(data = 0, nrow = nrow(incendios),ncol = 2))
    vectorDiasTotales <- as.character(Ninc1MeteoPorDia$V10)
    incPuente[,1] <- c(1:nrow(incendios))

    for(i in 1:nrow(incendios)){
    fecha <- strsplit(as.character(incendios$DHDETECCION[i]), split = " ")
    SfechaInc <- fecha[[1]][1]

    incPuente[i,2] <- which(SfechaInc == vectorDiasTotales)[1]

    }
    incPuente

}


Meteo2Incendios <- function(){

    TempMax <- numeric(0)
    Hrelativa <- numeric(0)
    VViento <- numeric(0)
    DirViento <- numeric(0)
    Chuvia <- numeric(0)

    factorIDPROVINCIA <- names(table(incendiosmeteo$IDPROVINCIA))
    factorMES<- names(table(incendiosmeteo$MES))
    factorANHO<- names(table(incendiosmeteo$ANHO))
    factorSEMANA<- names(table(incendiosmeteo$SEMANA))
    factorHORA<- names(table(incendiosmeteo$HORA))
    factorIDCLASEDIA<- names(table(incendiosmeteo$IDCLASEDIA))

    factorIDCAUSA<- names(table(incendiosmeteo$IDCAUSA))
    factorIDGRUPOCAUSA<- names(table(incendiosmeteo$IDGRUPOCAUSA))


    factorIDCAUSAS <- names(table(incendiosmeteo$IDCAUSAS))
    factorIDMOTIVACION<- names(table(incendiosmeteo$IDMOTIVACION))
    factorIDCAUSANTE<- names(table(incendiosmeteo$IDCAUSANTE))
    factorIDDETECTADOPOR<- names(table(incendiosmeteo$IDDETECTADOPOR))
    factorIDINICIADOJUNTOA<- names(table(incendiosmeteo$IDINICIADOJUNTOA))
    factorTIPOFUEGO<- names(table(incendiosmeteo$TIPOFUEGO))
```

```
        for (i in 1:nrow(tablePuente)){
            index <- tablePuente[i,2]
        TempMax[i] <- Ninc1MeteoPorDia$TempMax[index]
        Hrelativa[i] <- Ninc1MeteoPorDia$Hrelativa[index]
        VViento[i] <- Ninc1MeteoPorDia$VViento[index]
        DirViento[i] <- Ninc1MeteoPorDia$DirViento[index]
        Chuvia[i] <- Ninc1MeteoPorDia$Chuvia[index]

        }
        incendiosmeteo <- cbind(incendios, TempMax, Hrelativa, VViento, DirViento, Chuvia)
        incendiosmeteo$IDPROVINCIA <- factor(incendiosmeteo$IDPROVINCIA, c(15,27,32,36))
        incendiosmeteo$MES <- factor(incendiosmeteo$MES, c(1:12))
        incendiosmeteo$ANHO <- factor(incendiosmeteo$ANHO, c(2000:2014))
        incendiosmeteo$SEMANA <- factor(incendiosmeteo$SEMANA, c(1:53))
        incendiosmeteo$HORA <- factor(incendiosmeteo$HORA, c(0:23))
        incendiosmeteo$IDCLASEDIA <- factor(incendiosmeteo$IDCLASEDIA, c(1:4))
        incendiosmeteo$IDCAUSA <- factor(incendiosmeteo$IDCAUSA, c(1:12))
        incendiosmeteo$IDCAUSAS <- factor(incendiosmeteo$IDCAUSAS, factorIDCAUSAS)
        incendiosmeteo$IDGRUPOCAUSA <- factor(incendiosmeteo$IDGRUPOCAUSA, c(1:6))
        incendiosmeteo$IDMOTIVACION <- factor(incendiosmeteo$IDMOTIVACION, factorIDMOTIVACION)
        incendiosmeteo$IDCAUSANTE <- factor(incendiosmeteo$IDCAUSANTE, factorIDCAUSANTE)
        incendiosmeteo$IDDETECTADOPOR <- factor(incendiosmeteo$IDDETECTADOPOR,
factorIDDETECTADOPOR)
        incendiosmeteo$IDINICIADOJUNTOA <- factor(incendiosmeteo$IDINICIADOJUNTOA,
factorIDINICIADOJUNTOA)
        incendiosmeteo$TIPOFUEGO <- factor(incendiosmeteo$TIPOFUEGO, factorTIPOFUEGO)


        incendiosmeteo
    }
```