MASTER

ACTUARIAL SCIENCE

MASTER'S FINAL WORK

INTERNSHIP REPORT

MODELLING WILDFIRES IN MAINLAND PORTUGAL
AN APPROACH WITH MACHINE LEARNING

PEDRO RAIMUNDO DOS SANTOS LOPES

SUPERVISION:

TIAGO MARQUES FARDILHA
ONOFRE ALVES SIMÕES

NOVEMBER - 2024

Acknowledgments

Undertaking this thesis was truly a journey, throughout which the support given was unbelievable. I would like to express my special thanks to my faculty advisor, Professor Onofre Simões, for his incomparable guidance and motivation throughout the course of my Master degree, with special attention to the conclusion of this work. I benefited enormously from his patience and advices, allowing me to continuously improve my thesis.

I am also very grateful to my supervisor at Fidelidade, Tiago Fardilha, for giving me the orientation and tools necessary to complete this project. Your support and feedback is greatly appreciated, since it allowed me to develop clearer ideas for this work.

My sincere thanks to Rui Esteves and Fébio Levezinho for giving me the opportunity of doing this internship and for the precious hints and ideas they gave me.

To my family who supported me unequivocally, throughout my academic journey. A special thanks goes to them.

Last but not least, I have to acknowledge ICNF (*Instituto de Conservação da Natureza e das Florestas*) for providing me the dataset used in this work. They played a crucial role in my thesis and I am more than grateful to them.

"Data really powers everything that we do."
— Jeff Weiner

Abstract

Each year, wildfires cause billions of dollars of claims in the global insurance sector. Climate change scenarios suggest a potential increase in these losses, as rising temperatures and more frequent droughts intensify the underlying risk. Portugal, for its part, has one of the highest densities of wildfire ignitions among southern European countries, with this phenomenon posing the greatest threat to the sustainability of our forests.

Given the spatial and temporal uncertainty associated with wildfire occurrences, quantifying the underlying risk can be a challenging task that requires the use of advanced analytical methods. In this study, we analyzed 89 839 ignitions that occurred in Mainland Portugal over a 22-year period. Generalized Linear Models and Random Forests (RF) were employed to estimate the expected burned area of a wildfire and the likelihood of an ignition developing into a severe event, using a set of potentially explanatory variables. The obtained results provided insights into the key determinants within each modelling strand.

The estimated RF models were also used to predict the spatial patterns of ignitions at the national level under a climate scenario. Risk maps for the municipalities of Mainland Portugal were produced based on the resulting geographical predictions, indicating that the highest risk is predominantly concentrated in the inner central region of the country.

The estimated impact of the climate scenario considered in each modelling strand was further assessed. Risk maps reflecting the predicted variations due to the scenario were created, with the Alentejo region expected to be the most affected one.

Keywords: Wildfire; Burned Area; Severe Ignition; Random Forest; Climate Scenario; Risk Map.

Resumo

Todos os anos, os incêndios florestais causam a nível global perdas seguradas na ordem dos biliões de dólares. Cenários de alterações climáticas sugerem um possível aumento nessas perdas, à medida que a subida das temperaturas e a maior frequência de secas intensificam o risco subjacente. Portugal, por sua vez, possui uma das maiores densidades de ignições de incêndio entre os países do sul da Europa, sendo este fenómeno a maior ameaça à sustentabilidade das nossas florestas.

Dada a incerteza espacial e temporal associada aos incêndios, quantificar o risco subjacente constitui um processo complexo, podendo requerer o uso de métodos de analítica avançada. Neste estudo, analisámos 89 839 ignições ocorridas em Portugal Continental ao longo de 22 anos. Foram utilizados Modelos Lineares Generalizados e Florestas Aleatórias (RF) para estimar a área ardida esperada de um incêndio e a probabilidade de uma ignição evoluir para um incêndio grave, usando um conjunto de variáveis potencialmente explicativas. Os resultados obtidos forneceram informações sobre os principais determinantes em cada vertente de modelação.

Os modelos RF estimados foram também usados para prever os padrões espaciais das ignições a nível nacional, sob um cenário climático. Mapas de risco para os concelhos de Portugal Continental foram construídos com base nas previsões geográficas resultantes, indicando que o risco maior está predominantemente concentrado na região interior centro do país.

O impacto estimado do cenário climático considerado em cada vertente de modelação foi também avaliado. Mapas de risco refletindo as variações previstas devido à imposição do cenário foram implementados, com a região do Alentejo a ser estimada como a mais afetada.

Palavras-chave: Incêndio; Área Ardida; Ignição Grave; Floresta Aleatória; Cenário Climático; Mapa de Risco.

List of Figures

List of Tables

Contents

1 Introduction

1.1 Motivation

The present report follows on from the studies carried out during the Actuarial Science Master's curricular program. It summarizes the analysis developed from February to August in an internship at *Fidelidade*, where I joined the *Direção de Estatística e Estudos Técnicos de Não Vida* (DET). The project was organized in a progressive manner where two main topics were defined, namely the severity of a wildfire, i.e, the expected burned area caused by it, and the probability of an ignition developing into a severe one. The expected dimension of severe wildfires, although appealing, could not be fully embraced under the existing constraints on time and work dimension.

1.2 Context

1.2.1 Literature review

The topic of wildfires is not a new one, as it represents an extremely sensitive matter that has been affecting our society and environment. According to Copernicus Atmospheric Monitoring Service (CAMS), implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF) on behalf of the European Union, wildfire events have far-reaching consequences for both air quality and greenhouse gas emissions ([21]). Insurance companies are largely exposed to the underlying risk of this phenomena. [18] reports that the global insured losses related to wildfire claims were registered at almost 5 billions of dollars in 2023. It is, in fact, a subject that has been capturing media attention, due to some unprecedented wildfire seasons, namely the ones of Canada (2023) and Australia (2019/20), that resulted in widespread devastation of local ecosystems and communities.

Given their chaotic nature, alongside the uncertainty regarding when and where will they occur, the wildfire forecasters have been developing methods to quantify the underlying risk with the aim of improving the ability of local communities and agencies to manage and respond to wildfires effectively. As [21] reports, the traditional fire forecasting has been relying on a method that links weather conditions with fire activity to create an index of fire risk, known as Canadian Fire Weather Index (FWI). However, this approach has its limitations, namely the tendency of overestimation in areas with limited fuel. Moreover, since the FWI was originally developed for Canadian forests, its extrapolation to different ecosystems becomes complex.

In addition, we must mention that the dataset that ICNF (*Instituto da Conservação da Natureza e das Florestas*) provided, and will be presented later, highlighted a serious concern in relation to the possible causes of an ignition. The man-made wildfires represent almost 90% of the occurrences among the ones with known cause in the period 2001-2022. From these ones, 39% were intentionally caused, whereas the remaining ones were due to negligence. Therefore, the majority of the ignitions is due to the unpredictable behaviour of humans, which relates to the adversities on predicting wildfires. This fact raises the need of more sophisticated techniques to accurately assess wildfire risk across diverse landscapes. In fact, there has been a remarkable growth in the integration of machine learning into wildfire forecasting. ECMWF has devel-

oped a tool based on this techniques, known as Probability of Fire (PoF), to effectively predict ignition occurrence globally, up to ten days in advance. The key advantage of it is not only the robust and accurate predictions, but also the low computational burden of the underlying model ([21]).

Previous studies have also addressed the modelling of the ignition risk. It is meant by ignition risk the chance of a fire starting as determined by the presence and activity of any causative agent. Different approaches have been carried by several authors across different countries. Logistic regression has been one of the most used ([8], [9], [7], [27] and [20]). Artificial neural networks ([12]), classification and regression tree algorithms ([6] and [5]) and Hurdle models ([10]) have also been used to model ignition occurrence. In fact, neural networks are known to be more robust in modeling inconsistent or incomplete databases ([8]).

1.2.2 Main approaches and goals

In the current work, due to the nature of our dataset, it was not possible to model the probability of an ignition. It would be necessary to simulate geographical coordinates in Mainland Portugal (for instance, outside some buffer centered in the ignition points) for the points where ignitions did not occur. However, the extraction of real data relative to the explanatory variables for those points would not be feasible. For this reason, our work addresses two main topics that have not been approached by the majority of previous studies: i) the modelling of the burned area per ignition, and ii) the modelling of the probability of an ignition reaching 100 hectares (ha) of burned area. In fact, the wildfire severity is also a concerning topic, since there are many ignitions with small dimensions that do not generate such a loss as the generated by the large ones. To elucidate the reader, only 19% of the ignitions in our dataset reached 1 ha of burned area. According to [26], only these ones are considered wildfires. This differentiation is indeed important for statistical analyses, since the small ignitions can be quickly controlled and do not cause a significant impact in burned area.

Therefore, the present paper aims to reach two main goals. First, we intend to provide strong explanatory models that can elucidate about how each of the included predictors may influence not only our expectation relative to the severity of an ignition, but also the probability of an ignition being a severe one, i.e, of reaching 100 ha of burned area (threshold set by ICNF and reported in [16]). For this component, Generalized Linear Models (GLM) will be employed, as they provide great interpretability, through the coefficient analysis. Moreover, we are also interested in incorporating a predictive component in this study alongside climate scenarios. Hence, Random Forest (RF) models will be used to predict the spatial patterns of wildfires for Mainland Portugal, under those scenarios. Besides being less prone to overfitting than other machine learning techniques (particularly, boosting ones), a RF contains a more robust predictive ability than the GLM approach ([24]). Risk maps for Mainland Portugal will then be produced with the future predictions of the RF models, aggregated by municipality. Hence, we aim at contributing to the measurement of the risk of each municipality, in the context of our modelling strands.

The implementation of the models described was done through a dataset that ICNF provided. It contains 475 449 geographical coordinates in Mainland Portugal, corresponding to the location of each reported ignition in the period 2001-2022. Additional

information relative to these points was also available, such as the burned area (in hectares) of each occurrence. Since only the ignitions that we classify as wildfires were taken into account, 89 839 observations were considered for this study, as the ones that did not reach 1 ha of burned area were excluded. Other features relative to each observation at the time of the ignition were also provided. Hence, the models were built by testing the underlying variables. For the full description of the predictors incorporated, refer to Section 3.3. The analysis was conducted using R, a programming environment for statistical computing and plotting. As for the used packages, *tidyverse*, *ranger*, *pdp*, *leaflet* and *sf* were some of the most relevant ones.

It may be worth to emphasize that, in the earliest years of our time period, the geographical coordinates of the database may not always correspond to points where the ignition started. Our source from ICNF confirmed that they may represent the GPS (Global Positioning System) position of the first firefighters vehicle arriving to the wildfire, or simply the coordinates of the nearest toponymy (locality, geodesic vertex, etc.). Nevertheless, based on our prior knowledge on this subject and on the comparison of our conclusions with the ones from previous studies (e.g, [8], [7], [5]), we believe this situation did not significantly affect our results.

1.3 Document Structure

The outline of this thesis is as follows: in Chapter 2 we proceed to a descriptive analysis of the temporal evolution of statistical wildfire indicators, in the context of Mainland Portugal; in Chapter 3 a theoretical description of Generalized Linear Models and Random Forests is discussed, followed by the description of the variables; Chapter 4 presents the models employed for the burned area per ignition and the probability of an ignition reaching 100 ha; the models validation is approached in this stage; in Chapter 5 we perform a sensitivity analysis of the climate variables, followed by the methodologies employed in the climate scenarios and in the mapping of the resulting future predictions; conclusions and final thoughts are drawn in Chapter 6.

2 Wildfire Indicators in Mainland Portugal

2.1 Geographical Area covered by the study

The topic of wildfires represents a major concern at a national level and, among the natural disaster events, it is perhaps the one that most affects our country. The biodiversity of portuguese forests raises the need of efficient policies for forest management.

Portugal has demonstrated, over the years, some susceptibility regarding wildfires. In fact, it presents some natural features that increase the likelihood of this phenomena, such as the usual hot and dry summers. There are also many regions with an unfavourable topography (steep slope, for instance), especially the ones in the northeast of the country, as well as some vegetation characteristics that may evoke the risk of fire ignition, such as the evergreen vegetation with great resistance to dryness. These conditions also favour the fire propagation, representing then a critical contribution to burned surface indicators.

The study area of the current paper is Mainland Portugal. Regarding its main features ([20]), Continental Portugal has an area of approximately 89 000 km^2 and is located between $37°N$ and $42°N$ of latitude and between $6°W$ and $10°W$ of longitude. In terms of altitude, it ranges from the sea level to 2000 meters above it, whereas the higher elevations tend to locate in northeast regions. Mean annual temperatures tend to situate between $7°C$ and $18°C$. Mean annual precipitation ranges from 400 mm to 2800 mm. It is worth to emphasize that the south tends to have, on average, higher temperatures and lower levels of precipitation, when compared to the north.

Figure 1 (with data extracted from *Pordata*) illustrates a decrease tendency of the total forest area in Mainland Portugal from 1995 to 2010, which has been inverted with an increase from 2010 to 2015. Nevertheless, we observed a decrease of 2.46% of the total forest area from 1995 to 2015, with the wildfire phenomena having a preponderant contribution to it. In fact, Portugal has more critical indicators when compared with other European countries, such as Spain, France, Italy and Greece. According to the European Commission, since 1980 Portugal has been registering an increase tendency of the total burnt surface, contrarily to the decrease behaviour verified in the referred countries.

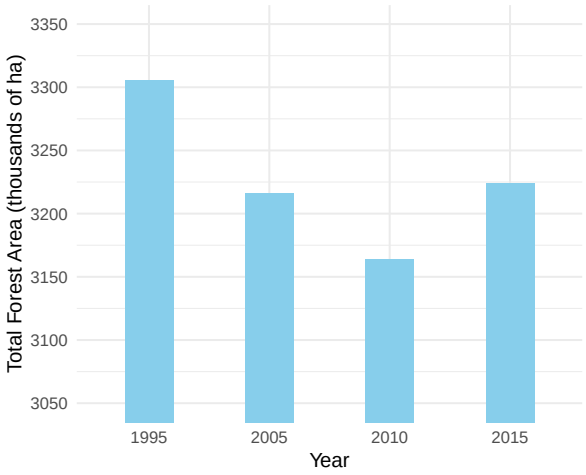


Figure 1: Evolution of the Total Forest Area in Mainland Portugal

2.2 Frequency and Severity Indicators

Portugal has been considerably affected in terms of total burnt surface due to wildfires. The annual average of this indicator from 2001 to 2022 is 138 033 ha, which represents 1.6% of the mainland territory. Figures 2 and 3 provide us some insights on this trend and enable us to conclude that 2017 was the most critical year in this respect. In fact, the most recent information we have on this concerns the year of 2022, when the burnt surface was only 1.2% of Mainland Portugal. This reflects a huge decrease relatively to 2017, when this indicator reached 6.1%. Several severe occurrences contributed to this result, namely the wildfires in the municipalities of Sertã, Pedrogão Grande and Lousã, that obviously had far-reaching consequences not only in terms of human damages, but also in terms of insured losses. In fact, the latest of the mentioned ones (Lousã, October of 2017) was the largest wildfire in Mainland Portugal from 2001 to 2022. The burned area caused by it reached 53 619 ha.

As it would be expected, the number of fire ignitions (displayed in Figure 3) has also been reaching huge values, resulting in an annual average of almost 21 028 ignitions from 2001 to 2022. Although this can be a relevant indicator, the main concern must be the burned surface or the number of ignitions with large severity, due to the fact that the fire frequency is highly affected by the ignitions with small dimensions. We can observe that 2017 was under the average, as it reached only 21 006 ignitions. However, it was one of the most tragic years in terms of not only total burnt surface, but also regarding the huge human consequences. On the other hand, the maximum value of 41 689 ignitions was observed in 2005. This was indeed a critical year, in which more than 346 000 ha were devastated, due to wildfires. Hence, further developments on this analysis regarding this year will follow.

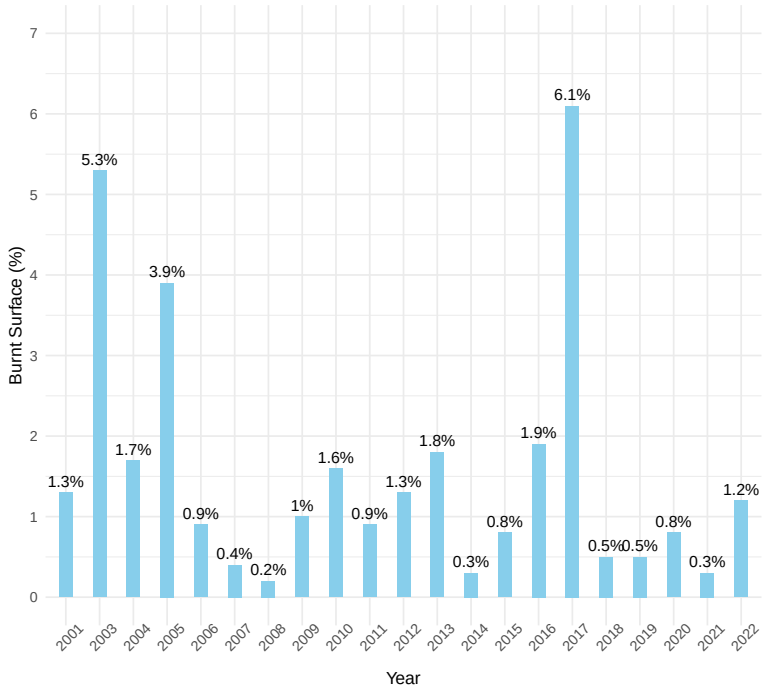


Figure 2: Evolution of wildfires in Mainland Portugal by proportion of burned area

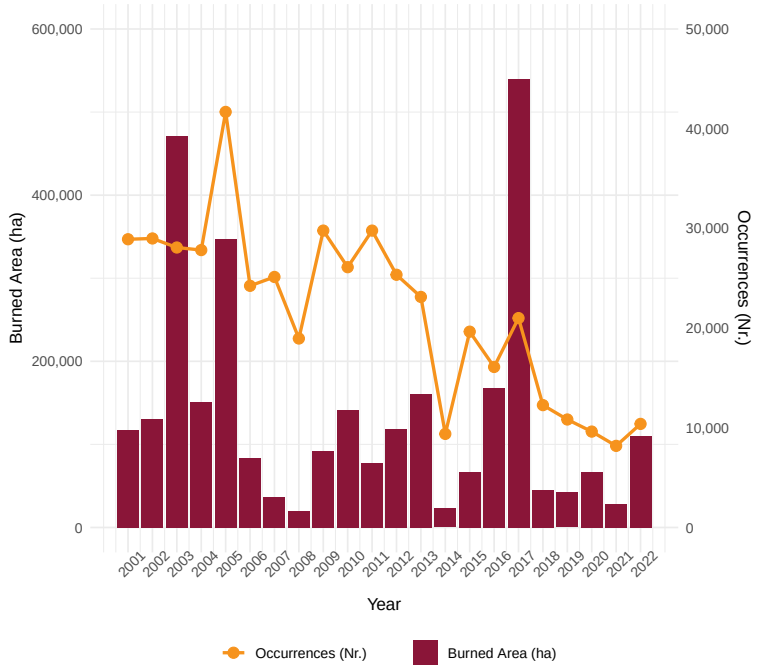


Figure 3: Evolution of wildfires in Mainland Portugal by burned area and number of occurrences

Moreover, Figure 4 provides an insight on the temporal evolution of an important indicator for our study, namely the annual number of wildfires exceeding 100 ha of burned area. The analysis of this indicator comes in line with what should be the main concern regarding the wildfire subject, namely the ignitions with large dimensions. In fact, [3] reports that the lower ones occur with greater likelihood, but end up to be controlled and extinct in the initial stages of the fire event, and the damages caused would be non significant.

As we observe, 2005 was one of the most critical years, both in terms of fire occurrences and severity per occurrence: 427 ignitions reaching a severity of 100 ha were incurred, which represents the maximum observed in the time period of our study. This year was, in fact, a major concern to all the entities involved in the wildfires subject. Thereby, we found convenient to detail our descriptive analysis in this specific year. According to ICNF, the ignitions in forest areas in 2005 were responsible for a burned area of 325 226 ha. In the context of Mainland Portugal, we must note that the man-made wildfires represent a proportion of almost 90% of the total wildfires. In fact, in 2005 we observed that more than 1/3 of the ignitions were intentionally caused.

According to [25], the wildfires with more than 100 ha were the most significant component of the total burned area, having represented 85.1% of the burned surface in 2005 and 93.1% in 2003. However, in both years, these large fires corresponded to a relatively reduced number of occurrences, since they did not exceed 1% of the total observed. In addition, a total of 3226 wildfires larger than 100 ha were observed from 2001 to 2022, corresponding to an annual average of 147. There were several years that revealed low values of this indicator such as 2008, 2014 and 2018, representing a downward contribution to the average. As previously said, Figure 4 proves that 2005 was the year that mainly pushed the mean of this measure in an upward direction.

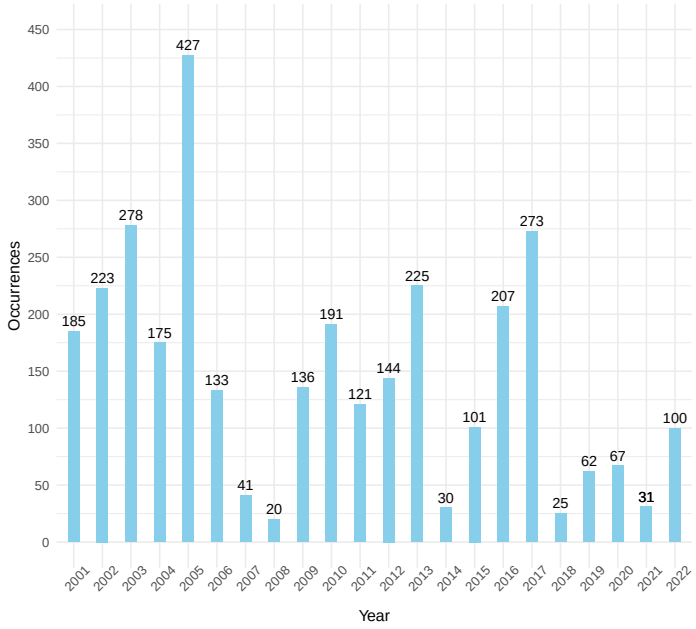


Figure 4: Evolution of wildfires in Mainland Portugal by number of ignitions reaching 100 ha

In fact, the climate conditions of 2005 were a huge enhancer of the wildfire statistics, namely the prolonged drought that affected the Portuguese mainland territory in that year. As it was stated by the former responsible of ICNF ([26]), in 2005 almost the whole all of the mainland territory was subject to high meteorological risk of wildfire, especially the district of Aveiro, which registered 22 012 ha of burned surface, the maximum of that year, among the 18 districts of the mainland territory. Moreover, one may say this year gave strength to the observed tendency on the usual cause of wildfires (natural or man-made) in Portugal, in the sense that only 2.2% of the total wildfires were due to natural causes.

In addition, 2017 was also one of the most noticeable years, with some critical consequences in human losses, and also for insurance companies. Figure 20 in Appendix A provides some insight on the burned area in 2017 at the district level, since it was a year when Portugal was highly affected by some severe incidences, with almost 49 000 ha burned in the district of Leiria. The ignition of Pedrógão Grande in June was the main contributor to this value, since the respective burned area (30 359 ha) represented 62% of the total for the district. The neighbor municipalities were also affected by this wildfire, some of them in other districts, such as Castelo Branco and Coimbra with a burned area in 2017 of 63 000 ha and 126 000 ha, respectively. In fact, Coimbra was the most concerning district regarding this indicator, also due to another fire incidence with big proportions that was triggered on the same day as the Pedrógão wildfire in the municipality of Góis. The competent authorities identified that the ignition source of Pedrógão incidence was dry thunderstorm, as it is mentioned in [29]. These two ignitions ended up to form a contiguous burnt area and had serious consequences in what regards financial losses, due to the assets that were affected. Also [29] reports that there were more than 500 destroyed houses, corresponding to an estimation of 500 million euros of total losses. Obviously, this situation had serious consequences to insurance companies, due to the several policies that incurred in a claim caused by these incidences.

3 Theoretical Foundation

3.1 Generalized Linear Models

In this section we follow the references [4] and [11], where further details can be found. Generalized Linear Models (GLM) are a well-known topic, regarding the implementation of statistical models, in which the aim is modelling the relationship between a dependent variable of interest and the relevant predictors. They represent a wide panoply of regression statistical methods with various applications, mainly regarding the data science area. We can consider the GLM an extension of the ordinary linear models ([11]), as they are particularly useful when the residuals (errors) appear to have a distribution different than the normal one. It is worth to emphasize that the distribution of the response variable is a good indicator of the residuals distribution.

Similarly to what is explained in [4] and [11], the main difference between a GLM and the classic linear models can be summed up by two essential aspects:

- The probability distribution of the response variable must be chosen from the exponential family (Gamma, Gaussian, Binomial, Poisson, Negative Binomial, Inverse Gaussian); Hence, the chosen one does not have necessarily to be the Gaussian distribution, as in an ordinary linear model.
- The relationship between the expected value of the response variable and the explanatory variables is specified via a link function.

Moreover, the probability distribution of each observation Y_i of the dependent variable must be specified according to the following expression:

$$f_{Y_i}(y_i; \theta_i; \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i; \phi; w_i)\right] \tag{1}$$

As for the parametrization, θ_i and ϕ represent, respectively, the location and scale parameters of the distribution of Y_i. The latter one, also called the dispersion parameter, is assumed to be constant for all the observations i. The parameter w_i corresponds to a prior weight specified in advance for each observation, and is denoted as the exposure parameter. The functions b(.) and c(.) are specified according to the chosen distribution for the response variable. In particular, b(.), denoted as the cumulative function, is twice differentiable with the second derivative a positive function, whereas c(.) is independent of the parameter θ_i. Thereby, we have that $E(Y_i) = \mu_i = b'(\theta_i)$ and $Var(Y_i) = \phi\frac{V(\mu_i)}{w_i}$, whereas $V_i = V(\mu_i) = b''(\theta_i)$ is called the variance function.

One of the key features of a GLM is its systematic component, denoted as the linear predictor, η. Basically, it consists in a linear combination of the p independent variables $(x_1, ..., x_p)$ included in the model, as defined in expression (2). Afterwards, the regression parameters (β_0, ..., β_p) must be estimated through maximum likelihood, using the available past data regarding the response and explanatory variables.

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2}$$

In an ordinary linear regression, the linear predictor would model directly the stochastic component, i.e., the response variable Y. However, with the GLM approach we

must incorporate a link function g(.), i.e., a differentiable and invertible function that maps the linear predictor with the expected value of the dependent variable. Thereby, the aim of the link function is to establish an association between the systematic and the stochastic components. The analytical reasoning is illustrated with expression (3), where u corresponds to the mean response variable:

$$\eta = g(u) \Leftrightarrow u = g^{-1}(\eta) \tag{3}$$

The link function must be chosen in accordance with the modelling strand that one aims to develop. As it was previously referred, one of the objectives of this paper consists in the implementation of a GLM approach in order to model the probability of an ignition reaching a severity of 100 ha. Thereby, a logistic regression will be implemented, since we are interested in a functional relation between the available predictors and the expected value of a binary variable (either the ignition reaches 100 ha or not). It is worth to emphasize the flexibility of this regression type, since it accepts a mixture of continuous and categorical variables, as well as non-normally distributed ones.

Therefore, a logistic function will be used to link the linear predictor with the expected value of the response variable, which in our case is the probability of an ignition being a severe one. Y represents the binary dependent variable, assuming the values of 1 (severe ignition) and 0 (non-severe ignition), and u symbolizes the respective expected value. The logistic link function is then defined as:

$$\eta = g(u) = log\left(\frac{u}{1-u}\right) \tag{4}$$

As for the probability of an ignition reaching a severity of 100 ha:

$$P(Y = 1) = u = g^{-1}(\eta) = \frac{e^{\eta}}{1 + e^{\eta}} \tag{5}$$

Since we are modelling the probability of a binary event, the model validation to be employed in Section 4.2 must be approached in a classification perspective, due to the fact that the past data regarding the response variable refers to binary outcomes. Hence, two different methods will be implemented with the aim of predicting a severe ignition through the predicted probabilities. We must emphasize that, with a logistic distribution, we are able to restrict the estimated values for the probability of success to the interval [0,1].

A similar approach can be developed if the purpose is modelling the burned area per wildfire. In this case, a Gamma distribution (particularly useful for modelling non-negative, continuous outcomes) will be used. This decision was mainly based on graphical inspection of the probability density function (p.d.f) of the Burned Area variable. In fact, the Gamma distribution is suitable for the right-skewed data (with a long tail towards larger areas) that we observe in Figure 5. This is consistent with the nature of fire data where small events occur frequently, but the large ones, though rarer, dominate the burned area totals. We have followed [28], that suggests the Gamma distributions are suitable for modelling proportions of burned area. We must highlight that the plotted variable in Figure 5 (the one we will use in the severity modelling) was subject to an outlier treatment. Although the basis of this decision

will be further explained in Section 4.1, we must mention that the ignitions exceeding a burned area of 100 ha were censored at 100, in order to avoid convergence issues in the computational process of the model fitting.

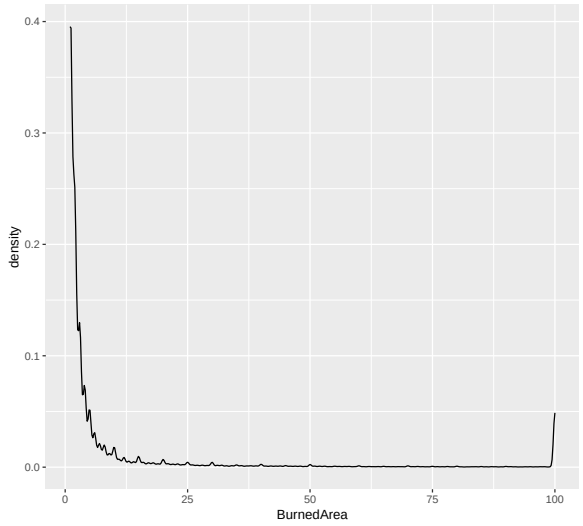


Figure 5: Burned Area density function

Therefore, the logarithmic link function will be employed, in order to provide us with the following coefficient interpretation: the parameter β_i allows us to calculate the percentage comparison of the category represented by x_i with the base level, *ceteris paribus*, regarding the expected burned area per ignition. We will further develop this in Section 4.1. Expression (6) provides the analytical reasoning:

$$\eta = g(u) = log(u) \Leftrightarrow u = e^{\eta} = e^{\beta_0+\beta_1 x_1+...+\beta_p x_p} \tag{6}$$

3.2 Random Forest Models

In this work, we find it convenient to employ, besides the GLM, a machine learning approach, since in the latest stage of the project the purpose will be to produce future predictions for the spatial patterns of wildfire ignitions in our study area. In fact, when the purpose of a model is to predict rather than explain, a machine learning approach is preferred over a regression based one ([24]). Therefore, a RF model will be employed.

This section follows the references [15] and [19], where further details can be found. Random Forest is a machine learning method of ensemble used in classification and regression tasks. An ensemble method is an approach that combines many simple "building block" models in order to obtain a single and potentially very powerful model. These building block models are sometimes known as *weak learners*, since they may lead to mediocre predictions on their own ([19]). The algorithm is mainly based in the construction of multiple independent decision trees, where each of them is grown in a randomly selected bootstrapped sample from the training data. The results of each tree are combined, in order to obtain more accurate predictions than the ones provided by GLM.

We start by introducing the concept of bagging as described in [19]. This procedure allows to surpass an adversity surrounding some tree-based methods, namely

the high variance of the results obtained when fitting decision trees to random subsets of the training data. Therefore, bootstrap aggregation, also called bagging, is a general-purpose procedure for reducing the variance of a statistical learning method. Recall that averaging a set of observations reduces the variance, i.e, given a set of n independent observations $Z_1, ..., Z_n$, each with variance σ^2, the variance of the mean $\overline{Z}$ of the observations is given by σ^2/n. One way to reduce the variance and increase the predictive accuracy of a statistical learning method would be to take several training sets from the population, build a prediction model using each of those and average the resulting predictions ([19]). Since it may not be feasible to have access to multiple training sets, with the bagging approach we must perform bootstrap selection. Basically, we sample with replacement from the training dataset, in order to generate B different bootstrapped training sets, although they may contain common points. The model is then trained on the bth bootstrapped training set, in order to get the respective prediction for that sample, denoted by $\hat{f}^{*b}(x)$. Finally, we average all the predictions to obtain the bagging estimate, defined by:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{7}$$

The bagging procedure can also be extended to a classification problem where the response variable Y is qualitative. In this case, the simplest approach would be as follows: for a given test observation, we record the class predicted by each of the B decision trees, and take a majority vote, i.e, the overall prediction corresponds to the most voted class among the B predictions.

A Random Forest is a substantial modification of bagging that builds a large collection of *de-correlated* trees, and then averages them ([15]). Similarly to bagging, a number of decision trees must be built in the bootstrapped training samples. However, in a RF procedure, on each split of a tree, a random sample of m predictors is chosen as split candidates among the full set of p explanatory variables ([19]). The split only uses one of those m predictors. While bagging considers $m = p$, in a RF model we typically choose $m \approx \sqrt{p}$. By not considering the majority of the predictors at each split in the tree, the RF procedure overcomes the adverse situation of having one very strong predictor in the data set, along with a number of other moderately strong predictors. In bagging, the predictions from the bagged trees would be highly correlated, and averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities ([19]). On the other hand, with RF, an average of $(p-m)/p$ of the splits will not even consider the strong predictor, and so other predictors have more significance on those splits. Hence, the main advantage of a RF over a bagging approach consists in what we may call "tree decorrelation", thereby reducing the variance of the average of the resulting trees and making them more reliable. Table 1 summarizes how a RF works.

In addition, a bagging or a RF model has the advantage of providing us with a variable importance measure through the computation of internal estimates. In the case of modelling a quantitative variable (RF for regression), the residual sum of squares (RSS) is computed, in order to provide an overall summary of the importance for each predictor. The RSS is calculated as the sum of the squared differences between

Random Forest for Regression or Classification

1. For $b = 1$ to B:
(a) Draw a bootstrap sample Z^* of size N from the training data.
(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
i. Select m variables at random from the p variables.
ii. Pick the best variable/split-point among the m.
iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\ \{T_b\}_1^B$

Source: [15]

Table 1: RF Algorithm

the observed values and the predicted values:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ , \tag{8}$$

where:

- y_i is the observed value of the target (response) variable for the i-th observation;
- $\hat{y}_i$ is the predicted value of the target variable for the i-th observation;
- n is the number of observations.

The importance of the RSS measure in RF for regression is that it quantifies the contribution of each feature (variable) to the model's predictive power. Thereby, we can record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all B trees ([19]). This measure helps to identify which variables are most influential in predicting the target variable, thus providing valuable insights into the model's behavior and the underlying data.

In a RF for classification, i.e, when modelling a qualitative variable, we can also record an importance measure for each predictor through the computation of the Gini index (also known as Gini impurity). Similarly to regression, we can add up the total amount that the Gini index is decreased by splits over a given predictor, averaged over all B trees. Expression (9) displays the formula of this measure:

$$G(t) = 1 - \sum_{i=1}^{m} p_i^2 \ , \tag{9}$$

where p_i is the proportion of instances of class i at node t, and m is the number of classes. The Gini index is a fundamental metric that quantifies the impurity of a node in a decision tree. By aggregating the reductions in Gini impurity across all trees, RF provides a robust measure of how important each feature is for making accurate classifications. This information can be crucial for understanding the model and for feature selection in machine learning methods.

3.3 Explanatory Variables

As referred in Chapter 1, a wide panoply of information was provided by ICNF. The dataset contains dynamic and structural features with respect to the place and time of each ignition. For the implemented models, both type of variables were selected and included. Although they had first to be tested as candidates, the prior selection process was based on previous knowledge (e.g, [8], [5], [9]). The dynamic predictors reflect temporary conditions subject to daily or even hourly change, like for instance the temperature or the relative humidity at the time and place of the ignition. The structural variables refer to intrinsic factors of the territory, which are relatively stable over time, such as the slope or the road density with respect to the point of interest. Hence, a total of 16 predictors were included in the models. The description of the variables is as follows:

- *Year* - Year of the ignition occurrence.
- *NSWD* - Number of simultaneous wildfires in the district of the ignition.
- *Density* - Population density of the municipality where the ignition occurred (persons per km^2).
- *District* - District where the ignition occurred.
- *Fire hazard* - Structural ICNF index of wildfire risk.
- *Month* - Month of the occurrence.
- *Altitude* - Mean altitude of the square 1×1 km of the ignition point (meters above sea level).
- *Land Cover* - Land use with respect to the ignition point.
- *Slope* - Mean slope of the square 1×1 km of the ignition point (%).
- *Road Density* - Road density of the square 1×1 km of the ignition point (meters per hectare).
- *Humidity* - Relative humidity of the air (%) at the time and local of the ignition.
- *Wind Direction* - Wind direction at the time and local of the ignition.
- *Precipitation* - Precipitation (millimeters) at the time and local of the ignition.
- *Temperature* - Temperature ($°C$) at the time and local of the ignition.
- *Wind* - Wind intensity (km/h) at the time and local of the ignition.

- *Dist. FD* - Straight line distance (meters) to the nearest fire department.

We must note that other variables were also tested, such as the meteorological indexes used in the calculation of the Canadian Fire Weather Index (FWI). However, due to their low explanatory power, they were not included in the models.

Before the GLM fitting, the variables were subject to a prior categorization, such that every level of a variable is compared to a base level. The quartiles of each variable were key factors for the creation of levels, with some posterior adjustments being made to achieve significance. Table 2 displays the treatment of the numerical predictors in the severity GLM, followed by the categorical ones. The base levels (highlighted in boldface) are the ones with the highest number of observations.

Variable \ Level	Level 1	Level 2	Level 3	Level 4
Year	$\leq$ **2009**	$]2009, 2016]$	> 2016	-
NSWD	$\leq$ **3**	$]3, 6]$	> 6	-
Density	< 40	$[40, 100[$	$\geq$ **100**	-
Fire hazard	< 2.5	$\geq$ **2.5**	-	-
Altitude	≤ 215	$]215, 509]$	$>$ **509**	-
Slope	< 7	**[7, 21[**	≥ 21	-
Road Density	$<$ **250**	≥ 250	-	-
Humidity	< 25	**[25, 50[**	≥ 50	-
Precipitation	$=$ **0**	> 0	-	-
Temperature	$<$ **25**	$[25, 32[$	$[32, 35[$	≥ 35
Wind	< 7	$[7, 10[$	**[10, 15[**	≥ 15
Dist. FD	< 1900	$[1900, 4700[$	$[4700, 6700[$	$\geq$ **6700**

Table 2: Treatment of the numerical variables - Severity GLM

As for the categorical predictors:

- *District*
  - **Porto and Braga**
  - Lisboa and Santarém
  - Évora and Setúbal
  - Leiria and Coimbra
  - Each of the 10 remaining districts corresponds to a level with no aggregations.
- *Month*
  - **July and August**
  - January and December
  - Each of the 8 remaining months corresponds to a level with no aggregations.
- *Land Cover*
  - **Shrubland**
  - Forest
  - Agriculture

- *Wind Direction*
  - **West**
  - East

Although the explanatory variables used for modelling the probability of a severe ignition coincide with the ones used in the severity, the categorization with respect to the logistic regression may differ for some of them. Table 3 refers to the variables that fit into that case. As for *Year*, *NSWD*, *District*, *Month*, *Altitude*, *Land Cover*, *Wind Direction*, *Precipitation* and *Dist. FD*, the categorization remains the same.

Variable \ Level	Level 1	Level 2	Level 3	Level 4
Density	< 75	$\geq$ **75**	-	-
Fire hazard	< 1.9	$[1.9, 3[$	$\geq$ **3**	-
Slope	$<$ **16**	≥ 16	-	-
Road Density	$<$ **175**	≥ 175	-	-
Humidity	< 40	$\geq$ **40**	-	-
Temperature	$<$ **30**	$[30, 35[$	≥ 35	-
Wind	< 7	**[7, 15[**	≥ 15	-

Table 3: Variable treatment - Logistic regression

Regarding the RF models, no prior categorization is considered. Thereby, the variables are tested in the exact way as they were provided (e.g, the 18 districts and the 12 months are all included separately). With this decision, we benefit from the high capacity of a RF in capturing interactions between variables.

The next section aims to describe the main details of the models employed for both the wildfire severity and the probability of an ignition reaching 100 ha. The results from the models validation approach are also displayed. We aim to provide a strong explanatory feature that highlights the influence of each variable regarding both modelling strands. The implementation of the GLMs were based on this objective. As for the RF models employed, since they have a predictive purpose, further developments will follow in Chapter 5.

4 Models and Results

4.1 Modelling the Burned Area per Ignition

First, we want to model the wildfire severity, i.e, the expected burned area per ignition. As already explained, two distinct methods were employed: a GLM and a Random Forest. Due to the fact that the majority of the ignitions have low dimensions, we started by facing convergence issues in the computational process. To deal with this problem, the technique that we employed will be presented later.

4.1.1 Preliminary analysis of the variables

Before we deepen the analysis into the technical insights of our models, we found it convenient to make a preliminary description of the patterns of the burned area per ignition, in relation to some of the selected variables of the models. Observed averages across the levels considered for the GLM were calculated.

- Firstly, we observed a decrease tendency of the average burned area per ignition with the increase of the population density in the municipality of the occurrence.
- By analysing the wind direction with respect to the ignition points, we observed that the ones in locations affected by easterly winds generated a burned surface, which was, on average, 16% greater than the ignitions related to westerly winds (36.2 *versus* 31.1 ha).
- As for the temperature at the time and local of the ignition, we display the graphical reasoning on Figure 21(a) in Appendix A: as expected, we observed an increasing tendency, whereas the ignitions belonging to the highest temperature level had the greatest average severity (331.8 ha).
- For the binary variable representing the occurrence of precipitation in the day of the ignition, the evidence is compatible with our prior knowledge, since the average severity of a wildfire in the precipitation scenario is 3.4% lower than in the no precipitation one.
- The distance from the ignition point to the nearest fire department also had the expected behaviour: the average burned area is strictly increasing with the referred variable, whereas the ignition points located further than 6700 meters from the nearest fire department had a mean severity of 49.1 ha. As for the ignitions related to the nearest level (less than 1900 meters), the mean burned surface was 15.2 ha. Figure 21(b) in Appendix A displays the graphical visualization regarding this predictor.

4.1.2 Results

We start with the implementation of a GLM that can provide us a strong explanatory ability. Although the main results of the RF models will be described in Chapter 5, some insights regarding variable importance will follow in this section. Thus, disparities between the GLM and the RF can be assessed. First, in order to avoid convergence issues in the computational process of the model fitting, a technique of outliers treatment was applied as we opted to censor the ignitions data. Hence, a burned area of

100 ha was set to all the observations with a severity larger than this threshold. This decision was due to the fact that our distribution would be extremely right skewed, since most of the ignitions have small dimensions. We remark that the chosen threshold is the one set by ICNF to designate a wildfire as a severe one, as mentioned in Chapter 1. Also, for the reasons given, we excluded the observations that did not reach 1 ha of burned surface. This decision resulted in a total of 89 839 wildfire ignitions. As for training and test techniques, we proceeded to a random splitting of the data, whereas the training set represents 80% and the remaining 20% is reserved for model validation.

Before the GLM fitting process, we performed several Chi-square tests of independence to our predictors, since we were interested in assessing possible associations between variables. This was done through the building of contingency tables between pairs of our categorical predictors. The Pearson's Chi-squared test is then applied to each of those tables. Hence, we verified that one of the most correlated pairs of explanatory variables is *Slope* and *Altitude* ($\chi^2 = 13795$, p-value $= 0$). This result strengthened our previous beliefs, motivating us to introduce an interaction component in the modelling framework, as we will see. Nevertheless, it is worth to mention that the results of the statistical tests performed should always be seen as an indication, as observations are bound to have some degree of temporal and spatial dependence ([8]). In fact, one of the advantages of a RF model resides on this topic, since it has an underlying optimization in the variables interaction to consider.

In line with this, two regressions, defined in Equations (10) and (11), have been considered as possible candidates. The main difference between them is that regression (10) is the simpler one with no interactions, whereas regression (11) employs the interaction term between the predictors *Slope* and *Altitude* (represented by *Slope* $\times$ *Altitude*). Thus, both regressions were calibrated alongside the RF model.

$$\begin{aligned}\log\left(Burned\ Area | Burned\ Area \geq 1\right) = \beta_0 &+ \beta_1 NSWD + \beta_2 Density \\ &+ \beta_3\, Year + \beta_4 District + \beta_5 Fire\ hazard \\ &+ \beta_6 Month + \beta_7 Altitude + \beta_8 Land\ Cover \\ &+ \beta_9 Slope + \beta_{10} Road\ Density \\ &+ \beta_{11} Humidity + \beta_{12}\, Wind\ Direction \\ &+ \beta_{13} Precipitation + \beta_{14}\, Temperature \\ &+ \beta_{15}\, Wind + \beta_{16} Dist.\ FD \end{aligned} \tag{10}$$

$$\begin{aligned}\log\left(Burned\ Area | Burned\ Area \geq 1\right) = \beta_0 &+ \beta_1 NSWD + \beta_2 Density \\ &+ \beta_3\, Year + \beta_4 District + \beta_5 Fire\ hazard \\ &+ \beta_6 Month + \beta_7 Land\ Cover \\ &+ \beta_8 Slope \times Altitude + \beta_9 Road\ Density \\ &+ \beta_{10} Humidity + \beta_{11}\, Wind\ Direction \\ &+ \beta_{12} Precipitation + \beta_{13}\, Temperature \\ &+ \beta_{14}\, Wind + \beta_{15} Dist.\ FD \end{aligned} \tag{11}$$

In order to provide a measure of model preference between regressions (10) and (11), the Akaike Information Criterion (AIC) of both models were calculated. The values of

417 443 and 417 379 were obtained in regressions (10) and (11), respectively. The lower value calculated for the latter means that regression (11) is the one which best balances goodness of fit and model complexity. Despite these results, regression (10) was the chosen one in the next developments. Although the AIC favored the model with the interaction term, the regression without interactions provides a greater simplicity and the ability to analyze each variable individually, making it more interpretable and practical for the study's objectives.

After the regression is calibrated, the focus goes to the interpretation of the estimated coefficients, as well as the assessment of the significance of each predictor. Results are displayed in Table 4.

First, the signs of estimated parameters were checked, to make sure they were compatible with some previous theoretical knowledge we had on the wildfires subject. The formulation of regression (10) allows us to interpret the coefficients as a percentage change in the burned area per ignition due to the marginal effect of any explanatory variable. Therefore, Equation (12) displays the formula that gives the percentage change on Y due to variable x_k:

$$100(e^{\beta_k} - 1) \ . \tag{12}$$

Table 4 displays the summary of some of the coefficients estimated for regression (10) (for all the coefficients, refer to Table 16 in Appendix A). These estimations result from a calibration to the whole dataset, as the more data we use, the more reliable is the explanatory feature of a regression model.

Coefficients	**Estimate**	**Std. Error**	**t value**	**p value**	**Sign. code**
(Intercept)	0.88944	0.03670	24.236	< 2e-16	***
Slope LEV1	-0.08228	0.02118	-3.884	0.000103	***
Slope LEV3	0.10273	0.01612	6.374	1.85e-10	***
Road Density LEV2	-0.08234	0.01819	-4.528	5.96e-06	***
Month Jan and Dec	-0.09946	0.04350	-2.286	0.022230	*
Month Feb	-0.32482	0.03328	-9.761	< 2e-16	***
Month Nov	-0.31893	0.04025	-7.924	2.33e-15	***
District Aveiro	0.37176	0.03898	9.537	< 2e-16	***
District Guarda	0.76434	0.03460	22.091	< 2e-16	***
Wind Direction East	0.12998	0.01426	9.113	< 2e-16	***
Altitude LEV 1	-0.22411	0.02511	-8.926	< 2e-16	***
Altitude LEV 2	-0.15275	0.01876	-8.143	3.90e-16	***
Dist. FD LEV1	-0.28712	0.02480	-11.579	< 2e-16	***
Dist. FD LEV2	-0.21245	0.01777	-11.956	< 2e-16	***
Dist. FD LEV3	-0.12084	0.01821	-6.636	3.24e-11	***

Table 4: Some of the estimated coefficients for regression (10)

By applying (12) to the results in Table 4, some conclusions follow:

- The burned area per ignition in the districts of Aveiro and Guarda is expected to be 45% and 115% greater than the verified one in the base level, Porto and Braga, respectively.

- Evidence of an upward effect of the variable *Slope* in fire propagation was also given by the model: the burned area is expected to increase by 8% between ignitions located in the area defined by the first level of *Slope* (under 7%) and the ones located in the area defined by the second level of the same variable (from 7% to 21%).

- It was also possible to quantify the effect of *Wind Direction*: the areas affected by easterly winds have an expected severity 14% greater than the ones affected by westerly winds. In fact, the former ones are normally associated to low levels of humidity, while the latter ones tend to be related to higher levels.

- As for the month of the ignition, we observe an unexpected behaviour for the January and December coefficient: the average burned area for this level is 9% lower than the one for the base level, July and August. We would expect a much stronger effect, in line with the coefficients verified for the levels November and February. The main contribution to this result was the anomalous January historic in 2021 and 2022: a total burned area of 1699 ha and 3312 ha were observed, corresponding to a mean per ignition of 21.5 ha and 25 ha, respectively.

- *Altitude* was found to have a positive influence on the wildfire severity: between the 1st level (under 215 meters) and the 3rd one (above 509 meters), we expect an increase of 20% in the severity of an ignition. [8] has also concluded a positive effect of this variable in the probability of ignition, namely due to some human activities at higher altitudes (renovation of pastures for livestock).

- As for *Road Density*, our results may reflect a better fire combat in areas with better road accessibility, as suggested by the negative effect of this variable. An interesting fact is the positive effect of this variable on previous studies for the probability of ignition. Nevertheless, this is expected since most of the wildfires are human caused and this variable is an indicator of human proximity and activity.

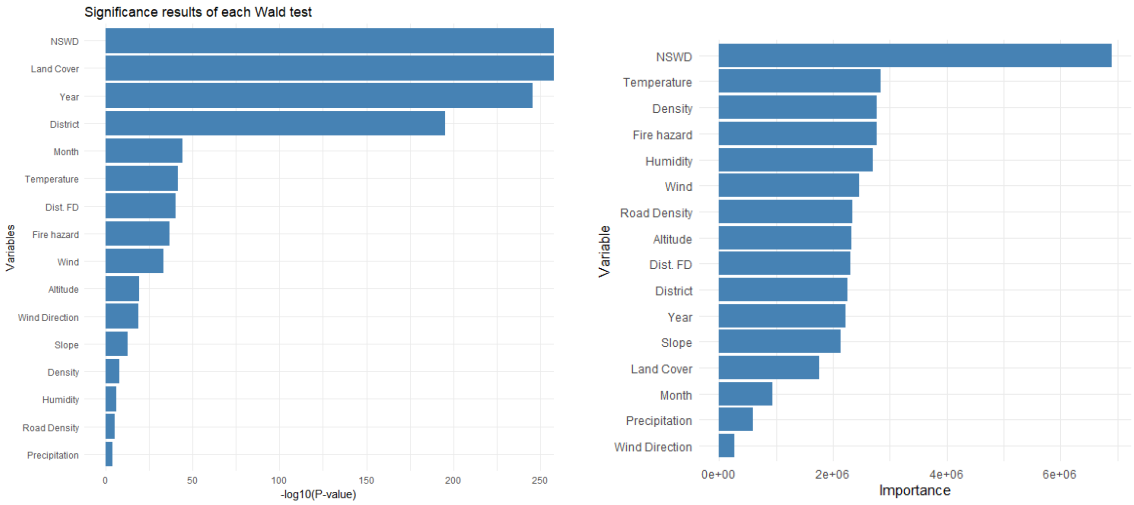
- We expect the severity of a wildfire to increase with the distance from where the ignition occurred to the nearest fire department (*Dist. FD*). In fact, it was proven that if the point of the ignition is located in less than 1.9 km (1st level), the burned area is expected to be 25% lower than if it is in a point further than 6.7 km (4th level).

As for the Wald tests to assess the significance of the predictor levels of regression (10), Table 4 displays on the rightmost column the p values of the significance tests of each estimated coefficient. Since none of them exceeds 0.05, we conclude that all the coefficients are significant at the level of 5% [1]. The referred tests were also performed with each explanatory variable, in order to compare them based on significance. The results are displayed in Figure 6(a), whereby the higher the transformation performed on the p-value, the more important is the respective variable. *Land Cover* and *NSWD* were identified as the predictors with the strongest influence on the expected burned area per ignition, with *Year* and *District* also playing an important role. *Precipitation*

[1] for the significance codes, see Table 15 in Appendix A

is the least important feature of the regression model, since in the majority of the observations there was no precipitation.

As for the RF model, an importance measure was computed, namely the one described in Section 3.2. The results are displayed in Figure 6(b). Similarly to the GLM, *NSWD* was pointed as the most significant predictor, but now with a considerable margin with respect to the remaining variables. The results highlighted by both models may reflect the strong marginal effect *NSWD* has on the expected severity of an ignition. In fact, the fire combat capacity is expected to be insufficient when there is a significant increase in the number of simultaneous active fires. As for the other predictors in the RF, the variability among them is not so extreme, regarding this measure. A similar evidence with respect to *Precipitation* was provided, with *Wind Direction* being the least important feature of the RF. On the other hand, some disparities between the results of the GLM and RF were also obtained. The most noticeable one concerns the variable *Land Cover*. In contrast with what happens in the regression model, the RF identified it as one of the least significant features.



(a) Significance results of the Wald tests - Severity GLM

(b) Variable importance - Severity RF

Figure 6: Comparison of the variable importance from the severity models

4.1.3 Validation

In order to assess the predictive ability of the models employed, we calculated the following metrics: root mean squared error (RMSE), mean absolute error (MAE) and the correlation between the predicted and observed severity per ignition. The results are displayed in Table 5.

Therefore, we must first use the regression and the RF calibrated to the training data, in order to predict the burned area per ignition in the test set. By modelling the logarithm in regression models, we face the problem of transforming the estimated expected *log(Burned Area)* back to expected *Burned Area* per ignition. To surpass this issue, the Duan's smearing factor (D_{Smear}) is estimated, using the residuals of the regressions ([14]). Thus, the mean response predictions of regression (10) is performed by applying the formula shown in Equation 13 below:

$$E(y|x) = e^{E(log(y|x))} D_{Smear} \tag{13}$$

	Regression (10)	RF model
Correlation	0.4553842	0.5892691
RMSE	18.96258	17.24226
MAE	9.7268	9.113688

Table 5: Performance metrics for each model

Table 2 displays the Pearson coefficient of correlation, as well as the RMSE and MAE between the predicted versus observed severity for each of the fitted models. Basing the analysis on the RMSE, we can see that the regression does not perform as good as the RF model. It presents a RMSE of 18.96258, which is 1.45% higher than the one of the RF. If we base the analysis in the other displayed measures, the conclusions are similar, as we observe a better performance of the RF model. As referred in Section 3.3, no prior categorizations in the explanatory variables were considered in the RF. This procedure offers the RF much more freedom in the variables treatment, allowing for some optimization automatically performed by the algorithm. The better predictive results of the RF in relation to regression (10) may be justified by this. In fact, in a first approach, the explanatory variables were subject to the exact same treatment as in the GLM implementation, while this final decision resulted in a significant improvement compared to the referred approach. Moreover, it is worth to mention that the results obtained gave strength to the decision of using the RF as a predictive model instead of the GLM, as we will further see in Chapter 5.

To offer a graphical visualization of the model performance, Figure 7 displays the observed versus predicted average burned area per ignition by each level of the variable *Month*. Thereby, we can assess the temporal patterns of the results obtained.

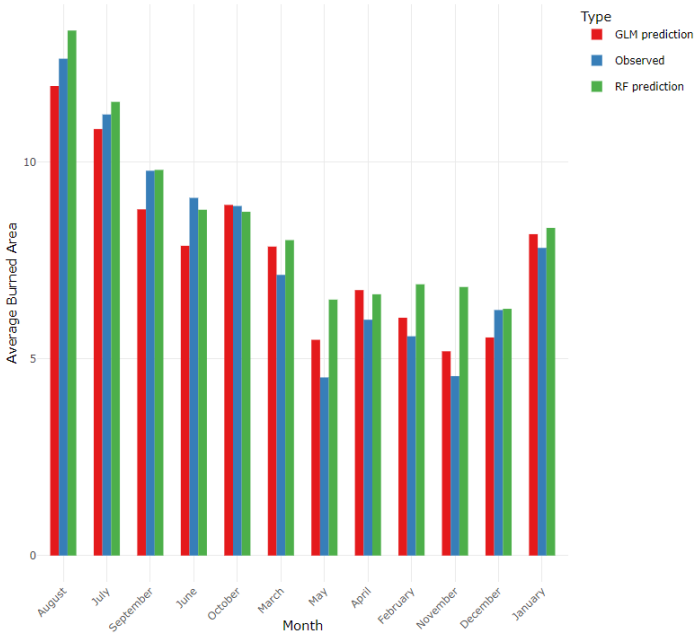


Figure 7: Observed versus Predicted average burned area by Month

By analysing Figure 7, we confirm some of the prior knowledge we had on the seasonal distribution of the wildfires severity in Mainland Portugal. July and August are the most concerning months, which is mainly due to the greater temperatures and the scarce levels of humidity and precipitation in the Portuguese summer. Our models

were able to capture that feature: the analysis of the graphic allows us to see a slightly underestimation with the GLM, in relation to what was observed in the test set.

On the other hand, the RF model overestimated in nearly all the months, the exceptions being October and June. In addition, January showed an identical behaviour to what was previously explained in the coefficient analysis, namely the unexpected high results for the predicted and observed severity.

As we are interested in assessing the spatial distribution of the wildfires severity, Figure 22 in Appendix A allows us to do an identical analysis regarding the average burned area per ignition in each of the 18 districts of Mainland Portugal.

Noting that the levels of the x-axis are in descendent order of ignitions number per district in the test set, special attention should be given to a fact identified in previous studies ([8] and [17]), namely that the areas of the country with the highest number of ignitions may not coincide with those where larger fires occur. This can be seen in Figure 22 (Appendix A), where we see that Braga and Porto are the districts with more ignitions, but are not even close of being the most concerning ones in terms of wildfire severity. Castelo Branco has the highest average burned area per ignition (15.2 ha). In fact, the model predictions for this district gave us interesting results: the RF model performed quite well, resulting in a slightly overestimation (15.4 ha), whereas the GLM resulted in a considerable underestimation (12.4 ha). The district of Coimbra is also known to have a concerning record, which can be related to the average burned area observed in the test set (14.9 ha). However, for this region, both models showed a relatively low prediction compared to what was observed: 11.4 ha with the GLM and 13.9 with the RF model. This can be partially explained by some severe ignitions that were not included in the training set and, consequently, were not used in the model calibration, namely the wildfires of 2017 in the municipalities of Góis and Figueira da Foz. In addition, we observe a tendency of higher predictions for the majority of the districts when using a RF model *versus* a GLM. Possible reasons for this may be related to the way each of the models treats the variables and their interactions. In fact, a GLM is a linearizable model and may not perform well in capturing complex or non linear relationships between a predictor and the response. On the other hand, a RF is a powerful technique in this field, since each of the created decision trees tries to capture different aspects of the data. Hence, it has a better ability to capture the heterogeneity and local variations in the data than the GLM. This may indeed contribute to greater predictions compared to the observed test set in districts where the GLM underestimates, as it is the case of Braga, Vila Real, Viseu, Santarém, Castelo Branco and Évora. Nevertheless, the GLM and the RF model are effective tools to explain and predict (respectively) the spatial patterns of the ignitions at the national level with good accuracy, which can be useful in decision-making for wildfire management.

4.2 Modelling the Probability of an Ignition reaching 100 ha

4.2.1 Justification of the metric

The probability of an ignition reaching 100 ha is a critical metric for understanding the likelihood of severe wildfires. In fact, they have disproportionate impacts on the environment and public safety, besides representing a huge risk to insurance companies.

Also, and as said in Section 4.1, the ignitions with a burned area greater than 100 ha were censored, as a way of controlling the influence of extreme values, but also to avoid computational issues in the model fitting of the wildfire severity. To elucidate the reader, we must mention that the 75% quantile of the burned area variable is only 5 ha, while the maximum is 53 619 ha (Lousã, October of 2017). Therefore, the current aim of this chapter is presenting the main details of our modelling approach, regarding the probability of an ignition reaching the threshold set by ICNF. Two main events must be considered in case of wildfire: the success (the ignition reaches 100 ha) and the failure (the ignition does not reach 100 ha). Similarly to the severity modelling, a GLM and a RF will be employed. The former contains a greater explanatory component, while the latter one offers a powerful predictive ability.

4.2.2 Results

As in the severity modelling, we start with the implementation of a GLM with an explanatory purpose, in order to highlight the contribution of each predictor to the likelihood of a severe occurrence. Some insights regarding variable importance from the RF model will follow in this section, in order to assess disparities between it and the GLM.

Since we are interested in implementing, through the GLM approach, a functional relation between the predicted probabilities of a severe ignition and the characteristics of the respective ignition points, a logistic regression will be employed. For calibrating purposes, it is evident that we must now use all the ignitions of our dataset with a burned area not smaller than 1 ha. Regression (14) has been considered as possible candidate, where p represents the probability of an ignition reaching 100 ha. Thus, we proceeded to the calibration of it alongside the RF model.

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = {} & \beta_0 + \beta_1 NSWD + \beta_2 Density + \beta_3 Year + \beta_4 District + \beta_5 Fire\ hazard \\ & + \beta_6 Month + \beta_7 Altitude + \beta_8 Land\ Cover + \beta_9 Slope \\ & + \beta_{10} Road\ Density + \beta_{11} Humidity + \beta_{12} Wind\ Direction \\ & + \beta_{13} Precipitation + \beta_{14} Temperature + \beta_{15} Wind + \beta_{16} Dist.\ FD \quad (14)\end{aligned}$$

Once the logistic regression is calibrated, we proceed to the interpretation of the estimated coefficients, as well as the assessment of the significance of each predictor. Table 6 refers to regression (14) calibrated on the whole dataset, as the more data we use, the more likely it is to generalise well (for all the coefficients, refer to Table 17 in Appendix A).

First, we checked that the signs of the coefficients were compatible with the ones estimated with the severity GLM. If we consider a logistic regression with dependent variable Y and a vector of explanatory variables X, then the marginal effect of any predictor on the odds of a severe ignition occur can be analyzed, by simply using the formula defined in Equation (15).

$$\frac{odds(Y = 1|X,\ X_k = 1)}{odds(Y = 1|X,\ X_k = 0)} = e^{\beta_k} \quad (15)$$

The formula displayed allows us to compute the marginal effect of the categorical variable X_k on the $odds(Y = 1|X,\ X_k = 1)$ against the $odds(Y = 1|X,\ X_k = 0)$. By

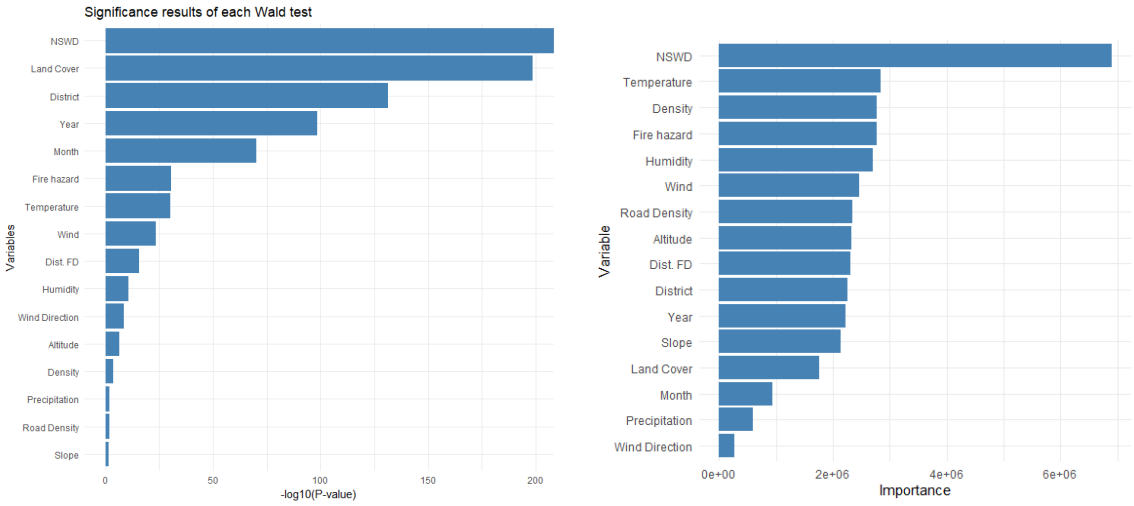
Coefficients	**Estimate**	**Std. Error**	**z value**	**p value**	**Sign. code**
(Intercept)	-7.0536	0.12501	-56.42428	0	***
Month November	-1.5137	0.25091	-6.0328	1.61e-09	***
Density LEV1	0.22784	0.06059	3.760	0.00017	***
Dist. FD LEV1	-0.53357	0.09069	-5.883	4.02e-09	***
Dist. FD LEV2	-0.44355	0.0559	-7.934	2.11e-15	***
Dist. FD LEV3	-0.22835	0.05324	-4.289	1.79e-05	***
Slope LEV2	0.11268	0.04993	2.257	0.02402	*
Road Density LEV2	-0.12523	0.05419	-2.311	0.02085	*
Humidity LEV1	0.32559	0.04846	6.718	1.84e-11	***
Altitude LEV1	-0.41191	0.07751	-5.314	1.07e-07	***
Altitude LEV2	-0.21353	0.05452	-3.917	8.98e-05	***
Wind LEV1	-0.25393	0.04975	-5.104	3.32e-07	***
Wind LEV3	0.39189	0.05218	7.511	5.89e-14	***

Table 6: Some of the estimated coefficients for regression (14)

applying the expression in Equation (15) to the results of Table 6, we can see that the odds of a severe ignition occurring in November is 78% lower than in the base level, July and August. Another interesting result is that the odds of a severe ignition is 26% higher for the points located in the area defined by the first level of variable *Density*, with respect to those located in the second level of the same variable. This means that, in Mainland Portugal, municipalities with a population density greater or equal than 75 individuals per km^2 have 26% lower odds than having a severe ignition compared to those with a population density lower than 75. Also, the coefficients estimated for the levels of the variable *Dist. FD* tend to increase from the first level to the fourth, meaning that the propensity to a severe ignition increases by moving away from the nearest fire department. As for variables *Slope* and *Road Density*, we obtained opposite effects: the former contributes positively to the odds of a severe ignition, as the zones in the higher slope level (above 16%) are 12% riskier than the zones in the lower one (under 16%). Evidence of a negative effect of the increase in the road density level was also proven, with respect to the propensity to a severe ignition. As for *Wind*, one may conclude a positive influence of it: between the 1st level and the 2nd one, we expect an increase of 22% on the odds of a severe ignition. On the other hand, an increase of 48% is expected between the 2nd and 3rd levels of the same variable.

Once again, the coefficients estimated for regression (14) are all statistically significant at a significance level of 0.05. Wald tests were performed in each predictor, in order to rank all the explanatory variables in terms of significance. As for the RF model, the importance measure described in Section 3.2 was calculated. Both results are displayed in Figures 8(a) and 8(b). In both models *NSWD* was identified as the strongest predictor. The most noticeable disparity relates to the variable *Land Cover*, since it plays a powerful role in the logistic model, but is one of the least important features of the RF. As in the severity GLM, *Year* and *District* also play a significant role in the regression (14). Another interesting result is the fact that *Temperature*, *Humidity* and *Wind* seem to have greater importance in the RF model compared to the logistic regression. In fact, it is the former one that will be used in the future

predictions of Section 5.2, where projections to the referred climate variables will be incorporated.



(a) Significance results of the Wald tests - Logistic regression

(b) Variable importance from RF

Figure 8: Comparison of variable importance - Probability of a severe ignition

Finally, the overall significance of regression (14) was assessed through the Hosmer and Lemeshow goodness-of-fit test, which is a measure of how well the model performs. If the significance of the test is small (i.e, less than 0.05), then the model does not adequately fit the data. By dividing the data into 10 groups (a standard practice in statistics as in [23]), we conclude that the model fits the data well ($\chi^2 = 5.7949, p\ value = 0.6702$).

4.2.3 Validation

In a first approach, an oversampling technique is applied, in order to deal with the imbalance of the data. Furthermore, risk classes will be created, in order to have a graphical assessment of the models performance. We must emphasize that the models have a regression purpose when used to predict the probability of a severe wildfire and a classification one when used to predict a severe or non severe ignition based on the probability estimated.

4.2.3.1 Using an oversampling technique

To validate the adjustment of the models, we will perform a 5-fold stratified cross-validation (CV) method (see [1]). We must take into consideration the fact that our dataset presents imbalanced proportions, since only 3.6% of the ignitions are severe, corresponding to 3193 out of 89 839 observations. Each of the 5 resulting models is calibrated in 80% of the dataset, such that all the observations are used once to test the model. Instead of a random sampling in the training and test split of the data, a stratified sampling was employed, allowing us to keep the same proportions of real data (3.6% of 1 and 96.4% of 0) in each of the 5 training sets. Otherwise, we would face the risk of having too few severe ignitions in the training set of some of the folds. In addition, we also used an oversampling technique in this CV process,

namely the Synthetic Minority Oversampling Technique (SMOTE) [2]. Basically, the SMOTE method creates synthetic data of the minority class (i.e, the class of severe ignitions), such that the training set reaches the 50% equilibrium for both classes of the dependent variable. As it is reported in [2], this technique uses the calculation of KNN (k-nearest neighbors) algorithm, which allows us to preserve the linear tendency of the original predictors, as well as the structure and the relationships present in the original data. Expression (16) displays the underlying formula for this technique:

$$s_i = x_i + (x_{z_i} - x_i) * \lambda \ , \tag{16}$$

where:

- s_i represents the new generated synthetic instance.
- x_i is the original instance of the minority class.
- x_{z_i} is a randomly selected neighbour instance of x_i.
- λ represents a value between 0 and 1 that controls the quantity of variation introduced in the creation of the synthetic instance.

In a k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. In each fold the model is trained in $k-1$ subsamples and the remaining one is used as test set. The process is then repeated k times, such that each subsample is used exactly once as validation data. To assess the performance of the final models resulting from this experiment, we display in Table 7 the mean and the variance of the sensitivity and specificity indexes across the 5 folds, since they measure the ability of the models in predicting severe and non severe ignitions, respectively. The sensitivity (also denoted true positive rate) and specificity (also denoted true negative rate) can be computed according to the formulas defined in Equations (17) and (18) [8].

$$Sensitivity = \frac{Number\ of\ true\ positives}{Number\ of\ true\ positives + Number\ of\ false\ negatives} \tag{17}$$

$$Specificty = \frac{Number\ of\ true\ negatives}{Number\ of\ true\ negatives + Number\ of\ false\ positives} \tag{18}$$

	Regression (14)	**RF**
Sensitivity (average)	0.8083	0.98997
Specificity (average)	0.7917	0.2841
SD(Sensitivity)	0.0043	0.0005
SD(Specificity)	0.0204	0.02134

Table 7: Mean and standard deviation of the sensitivity and specificity indexes across 5 different folds

We can see that the overall performance of the logistic regression is good, since we obtain an average of 81% of true positive rate and an average of 79% of true negative rate. As for the RF model, the results were quite interesting, since it performs almost perfectly in predicting 1 (average sensitivity of 99%), but the ability of predicting 0

was not as good (average specificity of 28%). This bias of the RF towards predicting the class of severe ignitions may be a result of how the model handles the synthetic samples generated by SMOTE. Although Random Forests are known for having a powerful predictive ability, when combined with SMOTE, they may become more prone to overfitting on these synthetic samples ([2]). These samples may not be a good representation of the minority class, leading to high sensitivity and low specificity.

Moreover, we may also perform an analysis based on receiver operating characteristics (ROC) to evaluate how well a presence-absence model is parameterized and calibrated, allowing to assess the model performance in a threshold independent fashion ([8]). The ROC curve can be obtained by plotting the Specificity *versus* Sensitivity for varying probability thresholds. Therefore, we were able to compute the mean area under the curve (AUC) across the 5 models, which was approximately 0.88 for the logistic regression and 0.91 for the RF. This result indicates good overall model performance in both cases, and reflects a ROC curve relatively close to the top left corner. As in ([8]), good model performance is characterized by a curve that maximizes sensitivity for low values of specificity (i.e, large areas under the curve). In fact, it is interesting to see that the performance analysis through ROC favours the RF model, despite the disparity in the results regarding sensitivity and specificity.

In addition, the information displayed in Table 8 provides the mean and standard deviation of three metrics computed across the 5 folds, in order to contribute to the quantification of the predictive ability of the models employed, namely the RMSE, the MAE and the correlation (Corr) between predicted and observed severe ignitions in the test set of each fold. Once again, the performance assessment of the models employed leaded to better results with the RF model. This conclusion follows from the lower values for the expected RMSE and MAE and a higher one in the correlation.

	Regression (14)	**RF**
RMSE (average)	0.4385	0.1874
MAE (average)	0.1923	0.0351
Corr (average)	0.2714	0.3643
SD(RMSE)	0.0041	0.0023
SD(MAE)	0.0036	0.0009
SD(Corr)	0.0064	0.0218

Table 8: Mean and standard deviation of performance metrics across 5 different folds

4.2.3.2 Using an approach with risk classes

Each of the 5 models of the CV process is calibrated in a different training set. Therefore, we extracted the respective index sets and trained a logistic regression in each of the 5 samples, in order to compare them in terms of significance. Since the model calibrated in the training set of fold 2 resulted in significance in all the levels of the explanatory variables, we selected the training and test samples of the referred fold for the next developments. In fact, the incoming approach requires a test set not used in the calibration, which motivated the fold selection.

The logistic regression and the RF model were both calibrated on the referred training set with the proportions previously referred: 3.6% of severe ignitions and 96.4% of non severe ones. Therefore, we are now using a training sample with no

synthetic data. In order to transform the estimated probabilities into binary outcomes we employed two different methods.

Method 1 consists in setting the cutoff value that maximizes the proximity between specificity and sensitivity. In the case of the logistic regression, Figure 9 displays the ROC curve for varying probability thresholds, whereas the cutoff value that meets our objective is 0.0423371, corresponding to the following combination situated on the top left corner of the curve:

- Sensitivity = 0.80721
- Specificity = 0.8222633

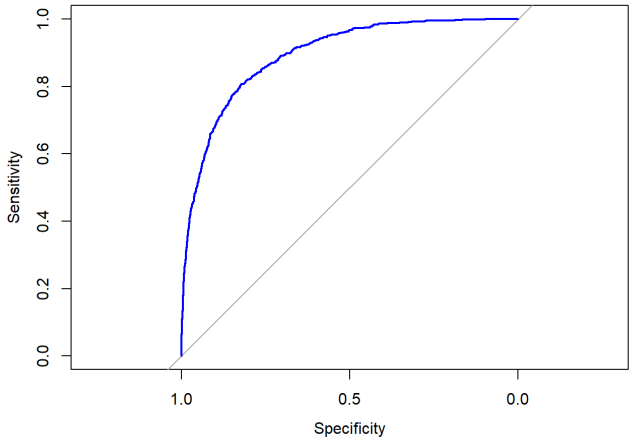


Figure 9: ROC curve for model predictions in the test set of Fold 2 - logistic regression

Method 2 consists in a Bernoulli sampling. Basically, we sample outcomes 0 or 1 for each record of the test set, from a Bernoulli distribution using the probabilities estimated through the implemented models.

In order to offer a visual comparison of the two methods and models, we decided to create risk classes through the intersection of the levels of the following explanatory variables: *Altitude*, *Wind Direction*, *Land Cover*, *Wind*, *NSWD*, *Year*, *Density*, *Road Density*, *Slope* and *Temperature*. This decision resulted in a total of 3229 classes with at least one observation in the test set. Table 9 gives an example of how a risk class is made and displays the predicted versus observed number of severe ignitions in the test set for that class.

Wind direction	**Year**	**...**	**Observed**	**Predicted**
East	$\geq$ 2017	...	2	4

Table 9: Example of a risk class

Figures 10, 11, 12 and 13 display graphical visualizations of the ability of the models to predict in the test set the number of severe ignitions for each risk class, using each of the methods of prediction.

From Figures 10 and 11, we can see that with method 1 the majority of the classes are situated on the right side of the bisector of odd quadrants for both models, indicating a tendency of overestimation. Indeed, only 74 and 35 out of the 3229 classes had more observed severe ignitions than the predicted ones in the regression and RF

models, respectively. We also observed that the prediction was 100% accurate in 1991 and 1737 classes (regression and RF, respectively), as the number of predicted severe ignitions coincided with the number of observed ones. It is worth to refer that, in the regression case, 93% of the 1991 classes had no severe ignitions (both predicted and observed).

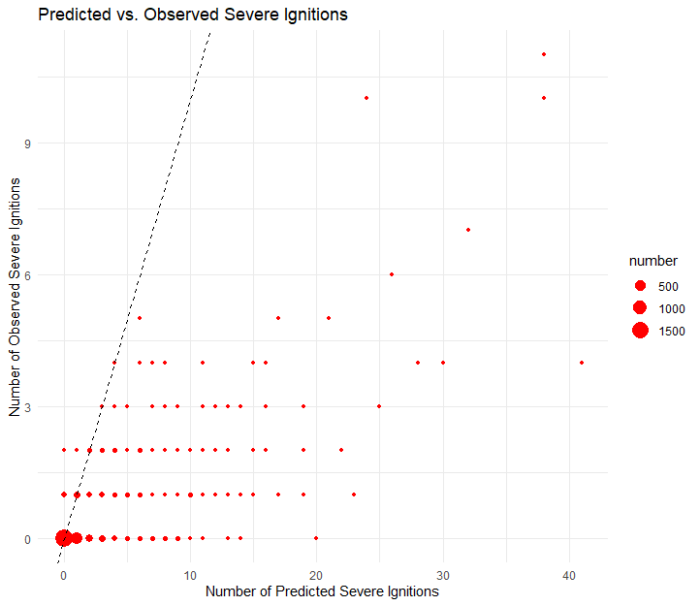


Figure 10: Predicted versus observed severe ignitions for different risk classes - Logistic regression and method 1

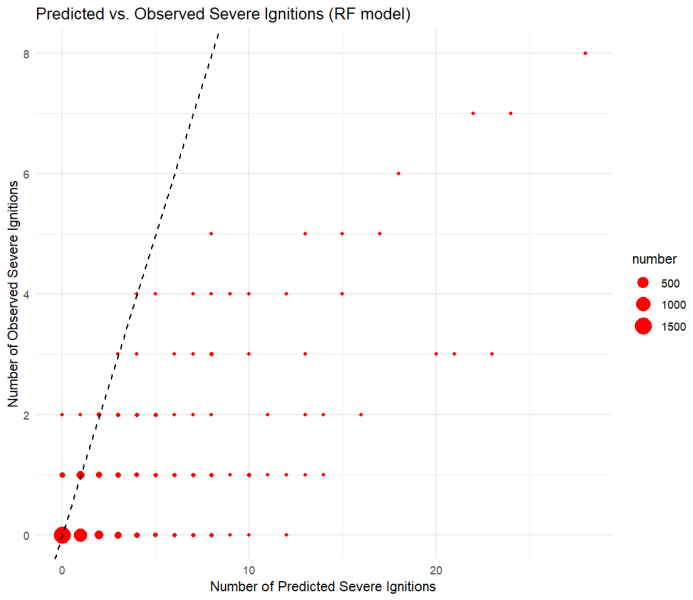


Figure 11: Predicted versus observed severe ignitions for different risk classes - RF model and method 1

Tables 10 and 11 display the confusion matrices of Method 1, which shows the relationship between predicted and observed 1 and 0. By applying the formula in expression (19), Method 1 results in an accuracy of 82.2% and 83.15% in the regression and the RF models, respectively.

$$Accuracy = \frac{Number\ of\ true\ positives + Number\ of\ true\ negatives}{Total\ Number\ of\ ignitions\ in\ the\ test\ set} \tag{19}$$

	Observed	
Predicted	0	1
0	14249	123
1	3080	515

Table 10: Confusion matrix of the predictions of Method 1 - Logistic model

	Observed	
Predicted	0	1
0	14405	103
1	2924	536

Table 11: Confusion matrix of the predictions of Method 1 - RF model

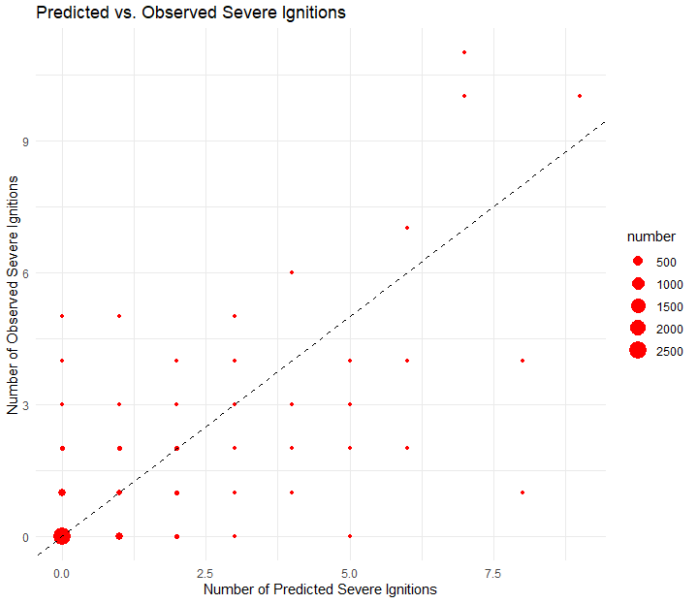


Figure 12: Predicted versus observed severe ignitions for different risk classes - Logistic regression and method 2

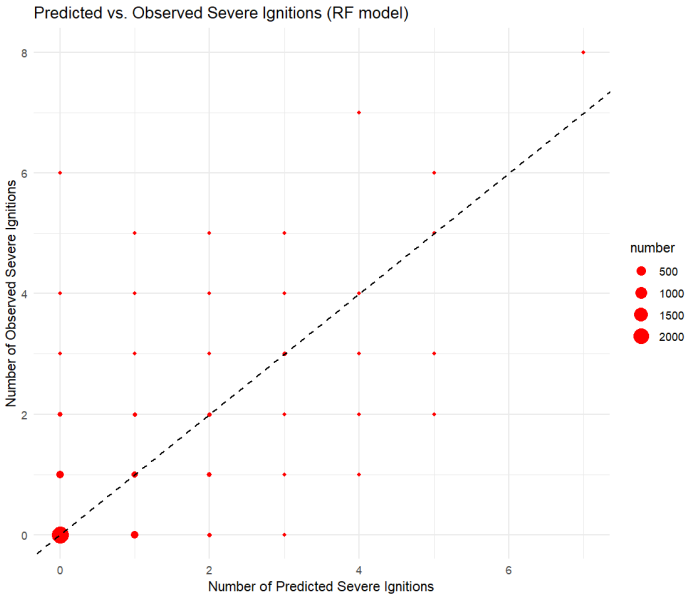


Figure 13: Predicted versus observed severe ignitions for different risk classes - RF model and method 2

Figures 12 and 13 show a greater predictive ability of Method 2, as the distribution of classes seems to have a better equilibrium among the two sides of the bisector of odd quadrants. To have a numeric insight on this, 306 and 288 out of the 3229 classes had more observed than predicted severe ignitions, whereas the opposite was verified for 300 and 297 risk classes (regression and RF, respectively). As for the 100% accurate predictions, the results were quite similar for both models: 81% and 82% of the total number of classes are situated in the bisector of odd quadrants, in the regression and RF models, respectively. In addition, Tables 12 and 13 display the confusion matrices of Method 2, from which we deduce an accuracy of 94.58% and 95% in the regression and the RF models, respectively.

	Observed	
Predicted	0	1
0	16848	492
1	481	146

Table 12: Confusion matrix of the predictions of Method 2 - Logistic model

	Observed	
Predicted	0	1
0	16880	466
1	449	173

Table 13: Confusion matrix of the predictions of Method 2 - RF model

These results confirm the graphical insights from Figures 10, 11, 12 and 13, not only by analysing the accuracy results but also the RMSE which is greater for Method 1 compared to Method 2 in both models: 0.422 versus 0.233 in the logistic regression and 0.41 versus 0.226 in the RF model.

In addition, in order to strengthen the conclusions taken from the previous approach, the analysis of the two models and the two cut-off methods on the test set was also performed at the risk class level, by assessing two metrics: i) the RMSE between the predicted and observed number of severe ignitions in each risk class, and ii) the weighted correlation (W.Corr) between predicted and observed values in each risk class, weighted by the number of ignitions belonging to the class. The current approach uses all of the 5 folds of the cross validation process. Hence, the displayed metrics in Table 14 result from averages across the 5 models. The lower E(RMSE) and the higher E(W.Corr) of the RF model with both Methods suggest a better predictive ability compared to the logistic regression. Methods 1 and 2 were denoted by M1 and M2, respectively.

	M1 - Reg. (14)	**M2 - Reg. (14)**	**M1 - RF**	**M2 - RF**
RMSE (average)	0.3375	0.1259	0.3056	0.1207
SD(RMSE)	0.4265	0.2766	0.3925	0.2685
W.Corr (average)	0.7691	0.6975	0.8079	0.7032
SD(W.Corr)	0.0376	0.0557	0.0274	0.0386

Table 14: Mean and standard deviation of the RMSE and W.Corr indexes of Regression (Reg.) (14) and RF model and cut-off methods M1 and M2 on 5 different folds

Similarly to the severity modelling, the validation process provided evidence of a greater performance of the RF model, compared to the logistic regression, in what concerns predictive ability of the probability of an ignition reaching 100 ha. In fact, the aim of the latest stage of the project is to produce and map future predictions of the spatial patterns of the ignitions, under a climate scenario. Therefore, the RF will once again be used, where we intend to predict in each municipality of Mainland Portugal the likelihood of an ignition developing into a severe wildfire.

5 Extensions of the RF Models

In this chapter, risk maps for Mainland Portugal will be produced through the predictions obtained for each municipality. The RF models are used for this purpose. In fact, and as said in Section 3.2, when the purpose of a model is to predict rather than explain, a machine learning approach is preferred over a regression based one ([24]). Moreover, a RF is less prone to overfitting than other machine learning techniques, particularly boosting ones ([24]).

5.1 Sensitivity Analysis of the Climate Variables

As a first step is necessary to assess the marginal behaviour of the variables *Wind*, *Temperature* and *Humidity* through a *ceteris paribus* approach. Our aim is to analyse the behaviour of these predictors in the RF models, as well as of having a prior insight of the impact of different climate scenarios on the future spatial predictions.

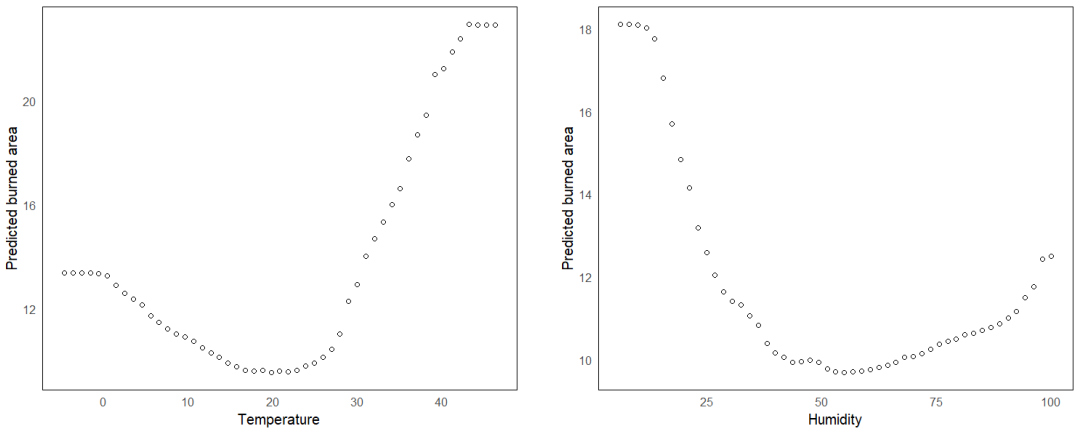
The partial dependence plots (PDP) show the marginal effect the selected variables have on the predicted outcome of a machine learning model ([13]). Hence, we may be able to check if the relationship between a feature and the target variable is linear, monotonic or more complex. When applied to a linear regression model, partial dependence plots always show a linear relationship.

For a classification problem where the model outputs probabilities, PDP displays the probability for a certain class given different values for the features of interest ([22]). The biggest issue with partial dependence plots is the assumption of independence, since it is considered that there is no correlation between the predictors of interest and the other ones. This may lead to biased results and misinterpretations. In addition, heterogeneous effects may not be captured by the partial dependence estimations, since PDP only show the average marginal effects. For further details regarding partial dependence, refer to Appendix B.

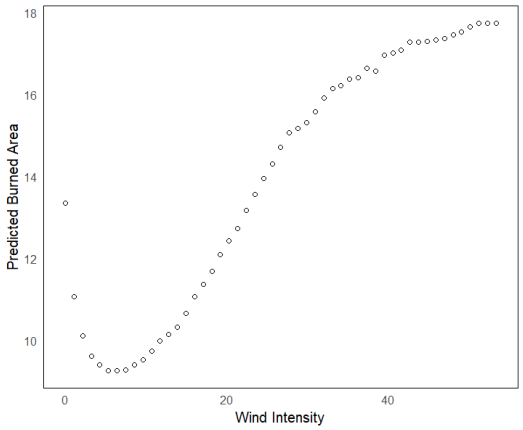
5.1.1 Burned area analysis

Under these guidelines, we will perform an analysis of the sensitivity of the burned area predictions to *Temperature*, *Humidity* and *Wind*. The PDP for *Temperature* in Figure 14(a) does not show a strictly monotonic relationship. In fact, it is of our knowledge that this variable may not be so significant for the lowest values as for the highest ones. We can explain this by the lack of data in the ranges where the RF model could probably not make a meaningful prediction. On the other hand, a temperature increase seems to be more significant at the highest ranges, namely above $29°C$. This was quite expected, since an increase of $1°C$ in the higher ranges has a much greater impact on the expected burned area, in case of an ignition.

For *Humidity*, the negative marginal effect that one may expect is observed in the most critical range, namely under 45% (see Figure 14(b)). Similarly to temperature, the most significant impacts of humidity increases in the expected burned area can be observed in a particular range of the distribution, namely between 12% and 27%. Above certain threshold, namely 55%, we observe a positive marginal effect on the expected burned area. In fact, this range only concentrates 36% of the training data and may contain the values where the relative humidity is not so significant. This may be the reason that prevent us from obtaining a strictly monotonic relationship between *Humidity* and the expected burned area due to a wildfire ignition.



(a) *Temperature* PDP (b) *Humidity* PDP



(c) *Wind* PDP

Figure 14: Severity PDPs

The variable *Wind* shows an unexpected result in the downward part of the curve, namely in the interval (0, 2.17) km/h (see Figure 14(c)). In fact, the strongest marginal impact of this feature is situated in this range. As it was previously referred, these results must be analysed with special attention, since the interval above refers to only 2.6% of the training set. The main purpose of this partial dependence analysis is to understand the behaviour of a feature in the most "significant part" of its distribution, i.e, where the majority of the ignitions is concentrated. In fact, above 9.26 km/h, the wind intensity has a positive marginal impact as it would be expected from experience and the GLM analysis.

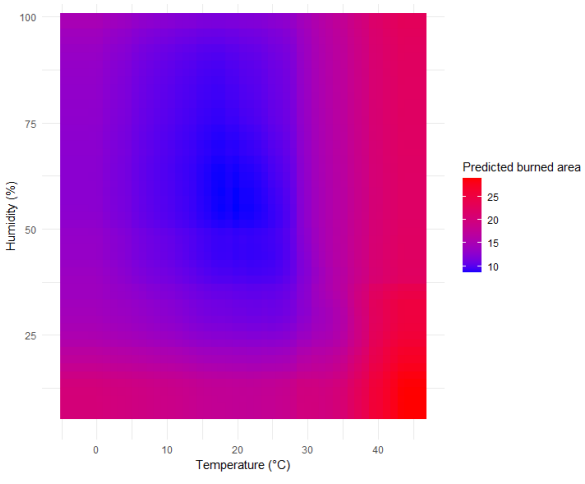


Figure 15: PDP for *Temperature* and *Humidity*

In addition, the partial dependence of interactions between climate variables was also analysed. The heat maps produced (Figure 15 and Figures 23(a) and 23(b) in Appendix A) display the average prediction for each selected vector of values

for the two predictors to consider. The results for the 3 pairs of variables were compatible with our prior expectations: in general, the visual inspection of the graphic for *Temperature* and *Humidity* (Figure 15) shows the expected effect that each of those has on the predicted burned area. The most critical scenario is located in the high ranges of *Temperature* and low ranges of *Humidity*. The results for high ranges of both variables are slightly more concerning compared to the low ranges scenarios. This may reflect the fact that, based on the importance measure previously introduced, *Temperature* is a stronger predictor than *Humidity*. As for *Temperature* and *Wind* (Figure 23(a) in Appendix A), both variables tend to have a positive marginal effect in their most significant ranges, i.e, where most of the ignitions occurred. An identical analysis can also be done with the heat map of *Wind* and *Humidity* (Figure 23(b) in Appendix A). In this case, the most critical scenario is located in the high ranges of the wind intensity and low ranges of the relative humidity.

An alternative approach to partial dependence was also carried out (see Appendix C).

5.1.2 Analysis of the probability of an ignition reaching 100 ha

The partial dependence analysis was also performed in the RF model for the probability of a severe ignition. Figures 24(a), 24(b) and 24(c) in Appendix A show an identical tendency in what regards the marginal effect of the climate variables. As for *Temperature*, the downward part of the curve is located in a similar range as in the burned area model, namely under $20°C$. However, the marginal behaviour in this interval seems to be much more attenuated than for the severity, as the variations in the predictions are much less significant. As for the upward part of the curve, visual differences are not so immediate, although the predominant positive marginal effect in the range above $20°C$ was patent in both models. For *Wind* it may be worth noting that both models display the strongest marginal impact in a similar range, namely (0, 2.17) km/h. Once again, one may consider it an unexpected result, since the referred range is located in the downward part of the curve. In what regards the upward part of the curve, no significant differences between the two models were noticed. Similarly, the partial dependencies of *Humidity* in both models show identical patterns in what concerns the ranges of positive and negative marginal effects on the average predictions.

5.2 Mapping the Predictions

The predictive component of this study was framed in the context of climate scenarios. Several approaches were considered, followed by the assessment of the impact of each of those in the resulting predictions. Thereby, we start this section by describing the one that had the strongest impact, in particular the three main steps necessary to implement the climate scenario.

As a first step, is necessary to model the climate variables through regression methods. Instead of considering a general approach for the entire Mainland Portugal, we proceeded to an analysis at the district and time period level. Regression models were considered with the district and part of the day (6:00 - 11:00, 11:00 - 16:00, 16:00 - 21:00, 21:00 - 6:00) as explanatory variables. After inspection of the density

functions, linear regressions (Normal distribution) were considered for *Humidity* and *Temperature* and a GLM (Gamma distribution) for *Wind*, as defined in Equations (20), (21) and (22).

$$Temperature = \beta_0 + \beta_1 Hour + \beta_2 District \tag{20}$$

$$Humidity = \beta_0 + \beta_1 Hour + \beta_2 District \tag{21}$$

$$\frac{1}{Wind} = \beta_0 + \beta_1 Hour + \beta_2 District \tag{22}$$

Thereby, the models are used to predict the climate variables, according to the district and the hour associated to each observation of our dataset. For each variable, 72 different predictions are obtained, each corresponding to a combination of time period and district. Hence, we complete the first step with preliminary estimations of the climate variables.

The second step can be summed up by the simulation of the variations to be applied in the 89 839 estimated values. Thereby, those variations are treated as random variables. As it was in our interest to consider more critical scenarios in relation to the preliminary estimates obtained, we carried out the following simulations of the differentials to be applied in the referred estimates:

- Average increase of $2°C$ in the temperature.
- Average decrease of 15% in the relative humidity.
- Average increase of 5 km/h in the wind intensity.

To give a clearer insight of this procedure, the differentials to apply in the temperature and relative humidity were generated according to a Normal distribution with means of $2°C$ and -15%, respectively. As for the wind, a Gamma distribution is used with a mean of 5 km/h. Since there is no variation on the preliminary estimations inside each pair (*Hour*, *District*) (as they are all equal), the standard deviations to be considered in those simulations must be the ones calculated with the real data for each set. Afterwards, the prediction set is now available before proceeding to the final step.

In the third stage, the predictions for the burned area and likelihood of a severe ignition must be performed using the obtained prediction set. Afterwards, we introduce a "penalty" component to be applied directly in the predictions, through multiplicative factors of 10%. This operation is only applied in the most critical ranges of each predictor:

- Temperature $> 38°C$
- Humidity $< 15\%$
- Wind > 33 km/h

Thereby, we are imposing a critical profile to the ranges where the RF may not be able to reflect a marginal behaviour as strong as one would expect, due to the lack of observations.

5.2.1 Risk maps

In order to be able to map the geographical predictions, we opted to aggregate them by municipality and produce an average prediction for each one. Therefore, in case that a future ignition occurs in a certain municipality, we were able to predict the expected burned area caused by it and the probability of that ignition being a severe one.

The maps displayed in Figures 16 and 17 contain the geographical predictions obtained for the burned area and the probability of a severe ignition, respectively. First we needed to make sure that both maps followed identical spatial patterns. In fact, the municipalities with the most expected burned area due to a wildfire ignition tend to be the most likely ones to be affected by a severe wildfire, conditioned to the ignition occurrence.

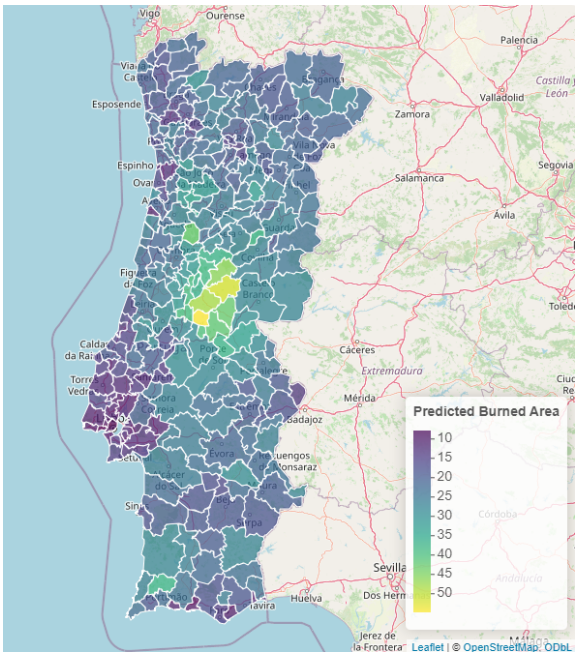


Figure 16: Predicted severity per ignition by municipality

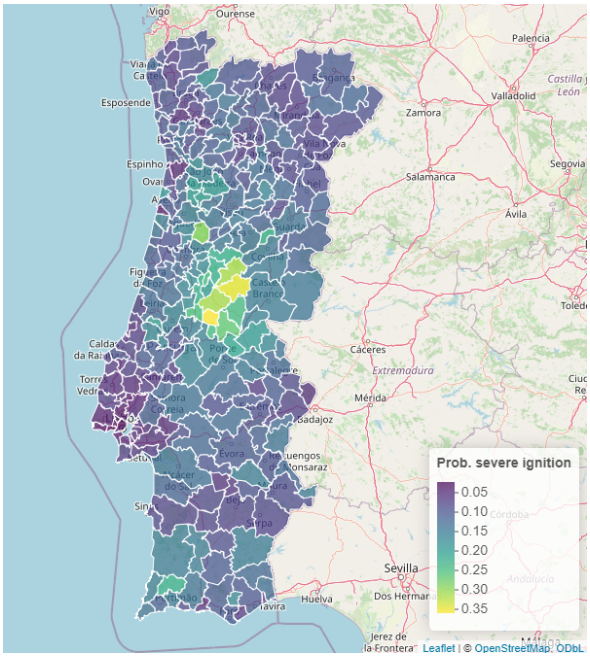


Figure 17: Predicted probability of a severe ignition by municipality

The riskiest municipalities seem to be centrally located in our study area, as it is the case of Vila de Rei, Oleiros, Sertã (Castelo Branco) and Pampilhosa da Serra (Coimbra). In the first one, we expect that 36% of the future ignitions will be severe ones. This value sustains the fact that this may be one of the most critical areas of the country due to the predominance of forest cover, although the wildfire historic also plays a heavy role in this result. In addition, an ignition occurrence in this region is expected to result in 55 ha of burned area. Special attention must be given to the fact that severe ignitions were censored at 100 ha in the severity model. Although the spatial patterns may not be highly affected, the predicted values for the severity must be interpreted having this in consideration. Moreover, in the validation stage of the RF model, the predictions for the test set gave us the highest results for the districts of Castelo Branco and Coimbra, which is also related to the results obtained for the municipalities above. In fact, the aforementioned municipalities may be considered

"re-incident" ones, since they have been registering multiple severe ignitions over the years. For example, the municipality of Oleiros was affected by severe wildfires in nine different years (2002-2005, 2011, 2012, 2015, 2017 and 2020) over the time period of this study (2001-2022).

The Algarve mountain ranges are also known for having a noticeable wildfire historic. In fact, a future ignition in the municipality of Monchique is expected to burn an area of 37 ha. As for the predicted probability of a future ignition reaching 100 ha of burned area, a value of 23% was estimated. This municipality has been considerably affected by some severe ignitions: a total of 10 wildfires reaching 100 ha have been registered in the time period of our dataset, namely in the years 2001, 2003, 2004, 2015, 2016, 2018 and 2021. One of the main factors contributing to this tendency is the predominance of forest cover. The high temperatures and the scarce levels of humidity and precipitation typical of the summer in this region are also related to these results. In addition, all of the 10 mentioned ignitions occurred in steep slope zones, which is indeed a relevant factor that favours the wildfire propagation.

On the other hand, the Lisbon Metropolitan area was considered to be the least critical area of the country, namely in the municipalities of Cascais, Oeiras and Lisboa. In fact, the latest one had the lowest predicted burned area per ignition (8 ha) and it is also the one with the lowest probability of a future ignition developing into a severe one (0.03). In fact, Lisboa is one of the municipalities with the highest population density, which was proven to have a negative effect on the expected severity and the probability of a severe ignition in the GLMs employed, but the lack of burning matter and the near fire departments may also be factors that contributed to this result, as it is one of the least affected municipalities in terms of wildfire historic.

Moreover, one may be interested in assessing the impact of the climate scenario not only on the expected severity of a wildfire, but also in the probability of it reaching 100 ha. Thereby, besides the predictions obtained under the projection scenarios previously described, it was also of our interest to consider predictions with no climate scenarios. In the latest, only real data from the ICNF dataset is considered. Afterwards, we consider, for each geographical coordinate, the future prediction under the climate scenario and the one with no scenario. The difference is then calculated, in order to take into account the variation imposed by the scenario.

Figures 18 and 19 display the average increase by municipality, with respect to the expected burned area of a future ignition and the propensity to a severe wildfire, respectively. Both maps show some heterogeneity in the spatial patterns of the predicted impacts per municipality. In fact, the North districts, such as Viana do Castelo, Vila Real or Bragança tend to be less affected than other ones, namely in the Alentejo region. It is quite interesting to realise that this behaviour is in line with some insights observed in the sensitivity analysis approached in Chapter 5.1. We proved that the most significant marginal effects occur in the most representative ranges of the predictors of interest, i.e, the ones that contain higher numbers of observations.

For the predictor *Temperature*, Figure 14(a) and Figure 24(a) in Appendix A showed that the strongest impact of a temperature increase occurs in the range above $29°C$. In fact, the Alentejo region is one of the areas of the country most affected by periods of high temperatures, which relates to the stronger impact obtained with the climate scenarios. The scarce levels of humidity in this region also contribute to the patterns observed, since the marginal effect of this variable behaves in an opposite way

according to Figure 14(b) and Figure 24(b) in Appendix A. The climate scenarios considered had the strongest predicted impact in the municipality of Gavião (Portalegre), in terms of expected increases in wildfire severity and proportion of severe ignitions, respectively. It is expected, on average, an increase of 14 ha in burned area. As for the proportion of severe wildfires, an increase of 10% is predicted for that municipality.

Although outside of the main scope of this thesis, the burned area of severe wildfires was also addressed. For a brief development on this topic, refer to Appendix D.

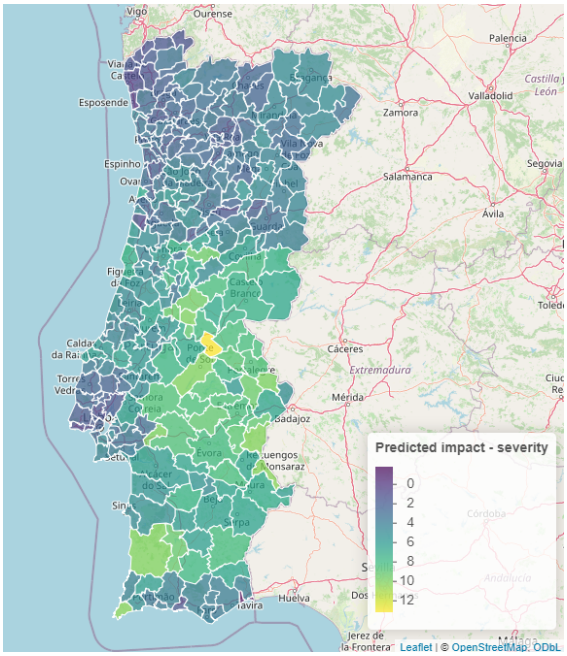


Figure 18: Predicted impact by municipality of the climate scenario with respect to the burned area of a future ignition

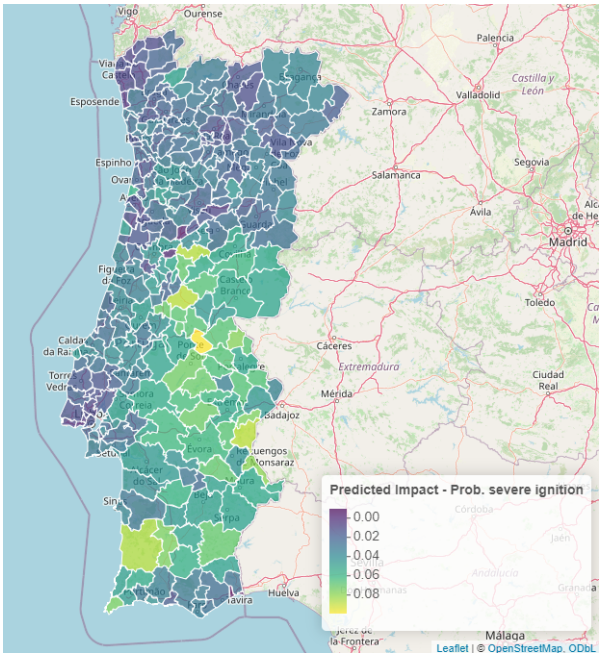


Figure 19: Predicted impact by municipality of the climate scenario with respect to the probability of a severe ignition

6 Conclusions

The nature of the ignitions dataset provided by ICNF makes the estimation of the ignition risk a difficult challenge: generating geographical coordinates for non-ignition points would be feasible; extracting real data for the relevant features with respect to these points would not.

In fact, previous studies ([8], [9], [7], [27], [20], [12], [6], [5] and [10]) have addressed this topic through a variety of different methods. For this work we proceeded to an alternative approach that aims two modelling strands not addressed by the referenced studies: we provided methodologies to estimate the severity of a wildfire and the probability of an ignition reaching 100 ha. The results obtained from the analysis of the coefficients of the regressions employed gave interesting conclusions regarding the impact of each variable in both modelling strands. The variable *Land Cover* showed a strong influence in the GLMs employed for both expected severity and risk of a severe ignition. Among the predictors with a positive estimated effect, *NSWD* highlighted a strong influence, with considerably high values in the estimated coefficients. On the other hand, a negative influence of the variable *Density* was found, meaning higher probability of a severe wildfire and expected burned area per ignition in the less populated areas where firefighting may not be so favourable.

Moreover, Random Forests were employed in both modelling strands alongside partial dependence tools, in order to assess the marginal behaviour of the climate variables. RF are characterized by their ability to capture non-linearity and interactions in the model inputs, which confer them a good predictive accuracy. Hence, we used them to predict the spatial patterns of the ignitions with a climate scenario. As one may expect, the resulting maps with the predicted severity and proportion of severe ignitions showed identical patterns. Some of the most critical municipalities are centrally located in Mainland Portugal (Oleiros, Vila de Rei, Pampilhosa da Serra), in line with the abundance of forest cover in those areas. As for the impact of the scenario considered in this report, the Alentejo region is expected to be the most affected one. In fact, some heterogeneity was deduced regarding the spatial patterns obtained, whereas the insights observed in the PD analysis were compatible with these results. The highest predicted increases in expected burned area per wildfire and proportion of severe ignitions were observed in the municipality of Gavião.

To sum up, the results from the implemented models showed an adequate adherence to the real world. The explanatory models employed revealed behaviours for each predictor that were compatible with our prior expectations. As for the predictive component of this study, we were able to achieve significant climate scenario impacts, meeting the interests of the company.

References

[1] Acervo Lima. *Validação Cruzada Estratificada k-Fold*. 2024. URL: https://acervolima.com/validacao-cruzada-estratificada-k-fold/ (visited on 09/29/2024).

[2] Bernardo Alemar. *Técnicas para Dados Desbalanceados (SMOTE e ADASYN)*. 2023. URL: https://medium.com/@balemar/t%C3%A9cnicas-para-dados-desbalanceados-smote-e-adasyn-f891f9c46c6e (visited on 05/15/2024).

[3] Ernesto Alvarado, David V Sandberg, and Pickford Stewart G. "Modeling Large Forest Fires as Extreme Event". In: *Northwest Science,* Vol. 72 (Jan. 1998), pp. 66–75.

[4] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. *A Practitioner's Guide to Generalized Linear Models*. 2007.

[5] Roba Bairakdar, Mathieu Boudreault, and Melina Mailhot. "Random Forests for Wildfire Insurance Applications". In: *Variance,* Vol. 16.(2) (2023).

[6] Joao Carreiras and José Pereira. "An inductive fire risk map for Portugal". In: *Forest Ecology and Management,* Vol. 234 (Nov. 2006), S56. DOI: 10.1016/j.foreco.2006.08.077.

[7] Pedro Miguel de Castro Alves. "Probabilidade de Ignição e Suscetibilidade de incêndios Florestais". MA thesis. Faculdade de Letras - Universidade do Porto, 2012.

[8] Filipe Catry, Francisco Rego, Fernando Bação, and Francisco Moreira. "Modeling and mapping wildfire ignition risk in Portugal". In: *International Journal of Wildland Fire,* Vol. 18 (Jan. 2009), pp. 921–31.

[9] Alexandra da Costa Ricardo. "Modelação da Probabilidade de ocorrência de incêndio em povoamentos florestais de Portugal Continental". MA thesis. Instituto Superior de Agronomia - Universidade Técnica de Lisboa, 2010.

[10] Marina D., Antonio G., Mario E., Raffaella L., Vincenzo G., Giuseppina S., Giuseppe C., Raffaele L., and Giovanni S. "Modeling fire ignition probability and frequency using Hurdle models: a cross-regional study in Southern Europe". In: *Ecol Process,* Vol. 5.(54) (2020). DOI: https://doi.org/10.1186/s13717-020-00263-4.

[11] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. 4th. Texts in Statistical Science. Boca Raton, FL: CRC Press, 2018.

[12] Luiza Cintra Fernandes. "Modelagem de risco de incêndios florestais utilizando redes neurais artificiais aplicada às regiões metropolitanas". Programa de Pós-Graduação em Modelagem e Análise de Sistemas Ambientais. Universidade Federal de Minas Gerais, Instituto de Geociências, 2019.

[13] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[14] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[15] T. Hastie, R. Tibshirani, and J. Friedman. "Random Forests". In: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009. Chap. 15, pp. 587–604. DOI: https://doi.org/10.1007/978-0-387-84858-7_15.

[16] ICNF. *8.º Relatório Provisório de Incêndios Rurais de 2023*. 2023. URL: https://www.icnf.pt/api/file/doc/058d65a2c60898dc (visited on 04/26/2024).

[17] INESC INOVAÇÃO and CEABN/ADISA. *Análise da Rede Nacional de Postos de Vigia em Portugal*. Tech. rep. Instituto Superior de Agronomia - Universidade Técnica de Lisboa, Dec. 2024. URL: https://www.isa.ulisboa.pt/ceabn/uploads/docs/projectos/postos_vigia/B7_Analise_da_RNPV_Relatorio_Final.pdf (visited on 08/28/2024).

[18] Insurance Information Institute. *Facts + Statistics: Global catastrophes*. 2024. URL: https://www.iii.org/fact-statistic/facts-statistics-global-catastrophes (visited on 03/27/2024).

[19] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "Tree-Based Methods". In: *An Introduction to Statistical Learning with Applications in R*. Springer texts in Statistics. Springer, New York, NY, 2023. Chap. 8, pp. 327–366.

[20] S. Marques, J. G. Borges, J. Garcia-Gonzalo, F. Moreira, J. M. B. Carreiras, M. M. Oliveira, A. Cantarinha, B. Botequim, and J. M. C. Pereira. "Characterization of wildfires in Portugal". In: *European Journal of Forest Research,* Vol. 130 (2011), pp. 775–84.

[21] Joe McNorton. *Machine learning ignites wildfire forecasting*. 2024. URL: https://www.ecmwf.int/en/about/media-centre/science-blog/2024/machine-learning-ignites-wildfire-forecasting (visited on 04/19/2024).

[22] Christoph Molnar. *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. 2024. Chap. 8.

[23] Ramzi W. Nahhas. *Introduction to Regression Methods for Public Health Using R*. 2024. Chap. 6. URL: https://bookdown.org/rwnahhas/RMPH/blr.html.

[24] Andrew Nailman. *When to Use Regression in Machine Learning: A Comprehensive Guide*. 2024. URL: https://machinelearningmodels.org/when-to-use-regression-in-machine-learning-a-comprehensive-guide/ (visited on 09/26/2024).

[25] Direção Geral dos Recursos Florestais. *Incêndios Florestais: Relatório de 2005*. Tech. rep. Direcção Geral dos Recursos Florestais, 2006. URL: https://www.icnf.pt/api/file/doc/4431684502bcf220 (visited on 04/27/2024).

[26] RTP. *Incêndios em 2005 devastaram mais de 325 mil hectares*. 2006. URL: https://www.rtp.pt/noticias/pais/incendios-em-2005-devastaram-mais-de-325-mil-hectares_n24777 (visited on 04/22/2024).

[27] Hugo Manuel dos Santos Saturnino. "Modelação e Mapeamento da Probabilidade de Incêndio Florestal". MA thesis. Instituto Politécnico de Castelo Branco - Escola Superior Agrária, 2011.

[28] Giovani L. Silva, Paulo Soares, Susete Marques, M. Inês Dias, M. Manuela Oliveira, and José G. Borges. "A Bayesian Modelling of Wildfires in Portugal". In: *Dynamics, Games and Science*. CIM Series in Mathematical Sciences, Vol 1. Springer, Cham, 2015, pp. 723–33. DOI: https://doi.org/10.1007/978-3-319-16118-1_38.

[29] Wikipedia. *Incêndio florestal de Pedrógão Grande em 2017*. 2024. URL: https://pt.wikipedia.org/wiki/Inc%C3%AAndio_florestal_de_Pedr%C3%B3g%C3%A3o_Grande_em_2017 (visited on 04/28/2024).

A Figures and Tables

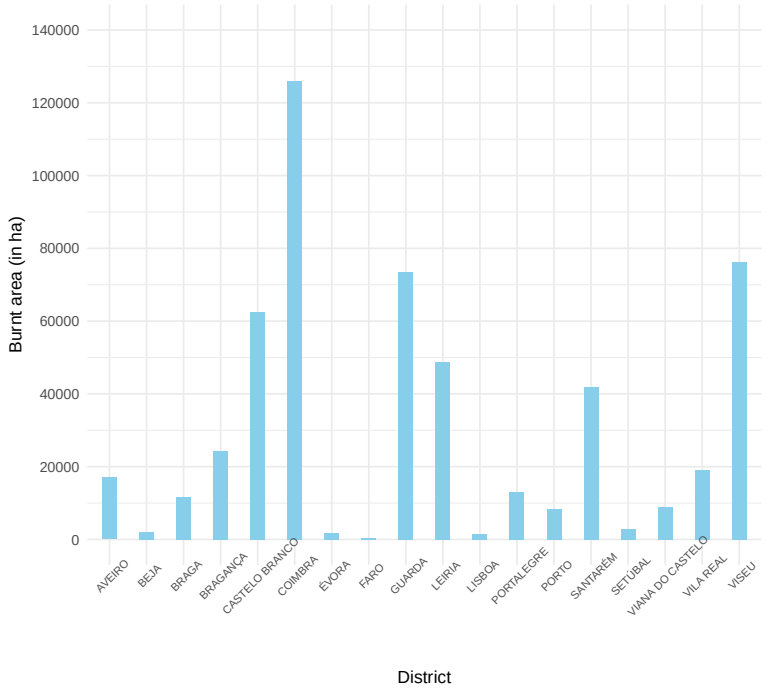
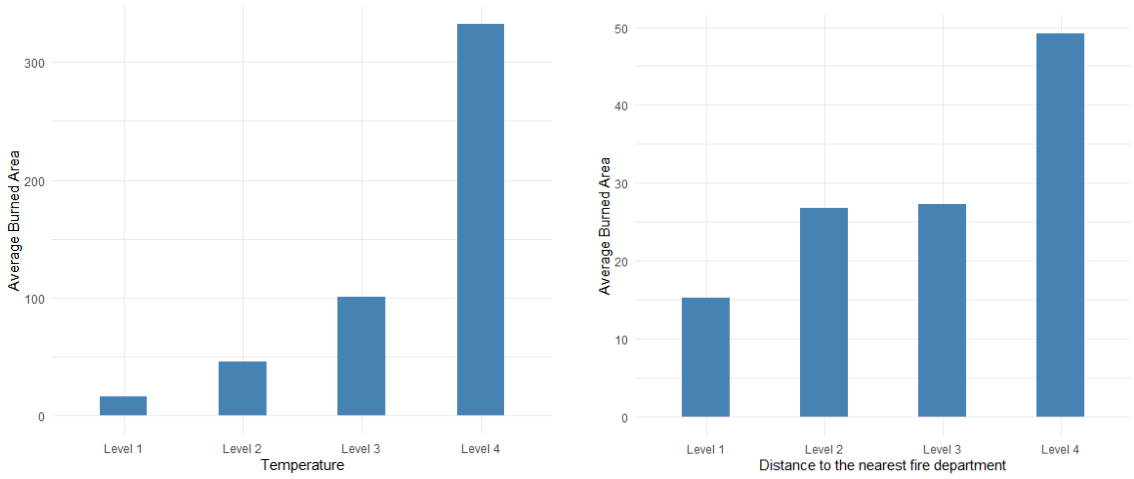


Figure 20: Total burned area per district in 2017



(a) Average burned area per ignition vs *Tempera-ture*

(b) Average burned area per ignition vs *Dist. FD*

Figure 21: Descriptive Plots for *Temperature* and *Dist. FD*

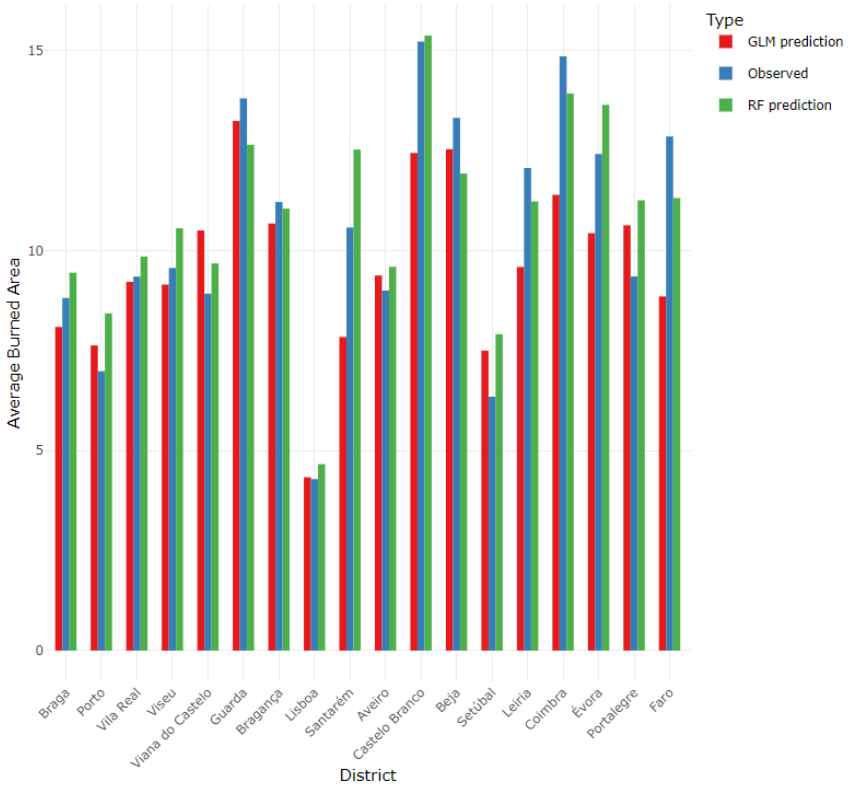
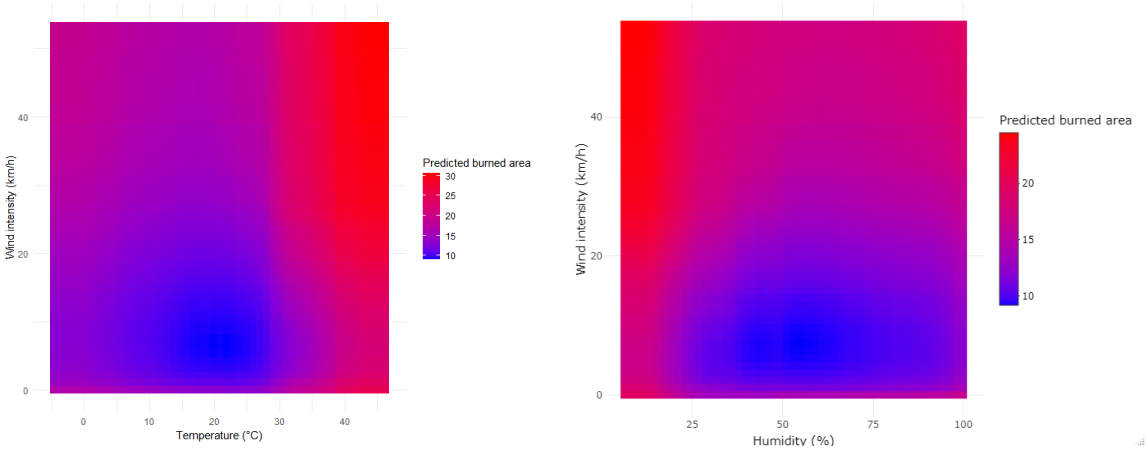


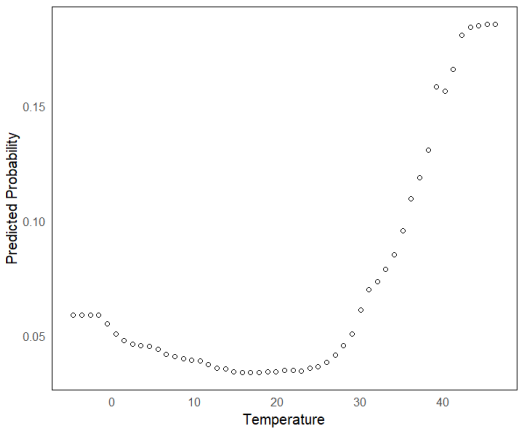
Figure 22: Observed versus Predicted average burned area by District



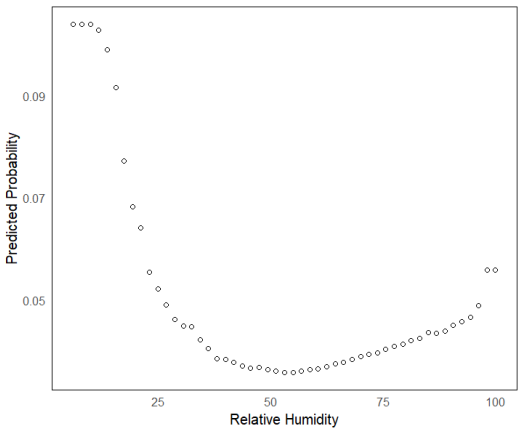
(a) PDP for *Temperature* and *Wind*

(b) PDP for *Humidity* and *Wind*

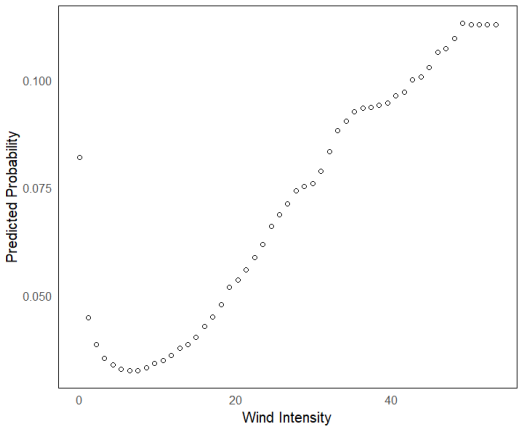
Figure 23: Heat Maps - Severity RF



(a) PDP for *Temperature*



(b) PDP for *Humidity*



(c) PDP for *Wind*

Figure 24: PDPs for the probability of a severe ignition RF

Significance code	**p-value**
***	[0, 0.001]
**	(0.001, 0.01]
*	(0.01, 0.05]
.	(0.05, 0.1]
	(0.1, 1]

Table 15: Significance codes

Coefficients	Estimate	Std. Error	t value	p value	Sign. code
(Intercept)	0.8894	0.03670	24.236	< 2e-16	***
Year LEV2	0.32166	0.01492	21.561	< 2e-16	***
Year LEV3	0.64654	0.02069	31.250	< 2e-16	***
NSWD LEV2	0.41192	0.01799	22.898	< 2e-16	***
NSWD LEV3	1.15567	0.01773	65.189	< 2e-16	***
Density LEV2	0.09627	0.02016	4.775	1.80e-06	***
Density LEV1	0.15173	0.02602	5.832	5.50e-09	***
District Lisboa and Santarém	0.18342	0.02850	6.436	1.24e-10	***
District Aveiro	0.37176	0.03898	9.537	< 2e-16	***
District Viseu	0.22964	0.02781	8.257	< 2e-16	***
District Leiria and Coimbra	0.46437	0.03995	11.624	< 2e-16	***
District Vila Real	0.23891	0.02936	8.136	4.13e-16	***
District Viana do Castelo	0.46173	0.02735	16.881	< 2e-16	***
District Évora and Setúbal	0.71395	0.04380	16.301	< 2e-16	***
District Guarda	0.76434	0.03460	22.091	< 2e-16	***
District Braganca	0.5993	0.03689	16.243	< 2e-16	***
District Castelo Branco	0.55513	0.04709	11.788	< 2e-16	***
District Faro	0.74772	0.06334	11.804	< 2e-16	***
District Beja	1.05279	0.05594	18.819	< 2e-16	***
District Portalegre	0.8507	0.06422	13.247	< 2e-16	***
Fire hazard LEV1	-0.21959	0.01711	-12.832	< 2e-16	***
Month September	-0.0857	0.0193	-4.440	9.01e-06	***
Month October	-0.15944	0.02521	-6.324	2.56e-10	***
Month March	-0.22714	0.02705	-8.397	< 2e-16	***
Month June	-0.08741	0.02565	-3.407	0.000657	***
Month April	-0.26705	0.03287	-8.125	4.54e-16	***
Month February	-0.32482	0.03328	-9.761	< 2e-16	***
Month May	-0.32950	0.03539	-9.311	< 2e-16	***
Month November	-0.31893	0.04025	-7.924	2.33e-15	***
Month January and December	-0.09946	0.0435	-2.286	0.022230	*
Altitude LEV 2	-0.15275	0.01876	-8.143	3.90e-16	***
Altitude LEV 1	-0.22411	0.02511	-8.926	< 2e-16	***
Land Cover Forest	0.79094	0.01408	56.181	< 2e-16	***
Land Cover Agriculture	0.24985	0.02558	9.768	< 2e-16	***
Slope LEV3	0.10273	0.01612	6.374	1.85e-10	***
Slope LEV1	-0.08228	0.02118	-3.884	0.000103	***
Road Density LEV2	-0.08234	0.01819	-4.528	5.96e-06	***
Humidity LEV3	-0.04876	0.01636	-2.981	0.002876	**
Humidity LEV1	0.14124	0.03173	4.451	8.55e-06	***
Wind Direction East	0.12998	0.01426	9.113	< 2e-16	***
Precipitation LEV2	-0.0827	0.02079	-3.979	6.94e-05	***
Temperature LEV2	0.15304	0.01911	8.008	1.18e-15	***
Temperature LEV3	0.42854	0.03777	11.345	< 2e-16	***
Temperature LEV4	0.63192	0.05837	10.826	< 2e-16	***
Wind LEV1	-0.12135	0.01738	-6.984	2.89e-12	***
Wind LEV2	-0.07199	0.01822	-3.952	7.76e-05	***
Wind LEV4	0.12745	0.01939	6.571	5.02e-11	***
Dist. FD LEV3	-0.12084	0.01821	-6.636	3.24e-11	***
Dist. FD LEV2	-0.21245	0.01777	-11.956	< 2e-16	***
Dist. FD LEV1	-0.28712	0.0248	-11.579	< 2e-16	***

Table 16: Estimated coefficients for regression (10) - Severity GLM

Coefficients	**Estimate**	**Std. Error**	**z value**	**p value**	**Sign. code**
(Intercept)	-7.0536	0.12501	-56.424	< 2e-16	***
Year LEV2	0.51564	0.04690	10.993	< 2e-16	***
Year LEV3	1.32544	0.06311	21.001	< 2e-16	***
NSWD LEV2	1.12292	0.07281	15.422	< 2e-16	***
NSWD LEV3	2.7402	0.06384	42.925	< 2e-16	***
Density LEV1	0.22784	0.06059	3.760	0.00017	***
District Lisboa and Santarém	0.96962	0.10631	9.121	< 2e-16	***
District Aveiro	0.88219	0.12042	7.326	2.37e-13	***
District Viseu	0.98665	0.08074	12.22	< 2e-16	***
District Leiria and Coimbra	1.82449	0.10196	17.894	< 2e-16	***
District Vila Real	0.89149	0.09059	9.841	< 2e-16	***
District Viana do Castelo	0.83806	0.08926	9.389	< 2e-16	***
District Évora and Setúbal	1.70674	0.14604	11.687	< 2e-16	***
District Guarda	1.81701	0.09686	18.759	< 2e-16	***
District Bragança	1.63624	0.10146	16.128	< 2e-16	***
District Castelo Branco	1.71367	0.11763	14.569	< 2e-16	***
District Faro	2.70811	0.16850	16.072	< 2e-16	***
District Beja	2.32374	0.15652	14.846	< 2e-16	***
District Portalegre	2.07107	0.17383	11.914	< 2e-16	***
Fire hazard LEV2	-0.50920	0.05225	-9.745	< 2e-16	***
Fire hazard LEV1	-0.76421	0.07622	-10.026	< 2e-16	***
Month September	-0.281	0.05717	-4.915	8.87e-07	***
Month October	-0.65372	0.08030	-8.141	3.92e-16	***
Month March	-1.35348	0.113	-11.978	< 2e-16	***
Month June	-0.24350	0.08333	-2.922	0.00348	**
Month April	-1.53818	0.17785	-8.649	< 2e-16	***
Month February	-1.90246	0.19817	-9.6	< 2e-16	***
Month May	-1.29663	0.20812	-6.23	4.66e-10	***
Month November	-1.51370	0.25091	-6.033	1.61e-09	***
Month January and December	-0.81527	0.2023	-4.03	5.58e-05	***
Altitude LEV2	-0.21353	0.05452	-3.917	8.98e-05	***
Altitude LEV1	-0.41191	0.07751	-5.314	1.07e-07	***
Land Cover Forest	1.78835	0.05974	29.938	< 2e-16	***
Land Cover Agriculture	1.1626	0.09113	12.758	< 2e-16	***
Slope LEV2	0.11268	0.04993	2.257	0.02402	*
Road Density LEV2	-0.12523	0.05419	-2.311	0.02085	*
Humidity LEV1	0.32559	0.04846	6.718	1.84e-11	***
Wind Direction East	0.25525	0.04245	6.013	1.83e-09	***
Precipitation LEV2	-0.19003	0.07514	-2.529	0.01144	*
Temperature LEV2	0.45934	0.05948	7.722	1.15e-14	***
Temperature LEV3	1.07854	0.09769	11.04	< 2e-16	***
Wind LEV1	-0.25393	0.04975	-5.104	3.32e-07	***
Wind LEV3	0.39189	0.05218	7.511	5.89e-14	***
Dist. FD LEV3	-0.22835	0.05324	-4.289	1.79e-05	***
Dist. FD LEV2	-0.44355	0.05590	-7.934	2.11e-15	***
Dist. FD LEV1	-0.53357	0.09069	-5.883	4.02e-09	***

Table 17: Estimated coefficients for the logistic regression - Regression (14)

B Theoretical Description of Partial Dependence

As ([22]) reports, the partial dependence function for regression is defined as:

$$\hat{f}_S(x_S) = E_{X_C}[\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C) \tag{23}$$

The x_S represent the variables of the model for which we are interested in plotting the partial dependence function and X_C are the other variables used in the RF model $\hat{f}$, which must be treated as random variables. Usually, the set S contains only one or two features, being the ones we want to assess the effect on the prediction. Hence, the total feature space is composed of the feature vectors x_S and x_C. The main aim can be summed up by the averaging of the RF model output over the distribution of the features in set C, in order to get a function that depends only on features in S, while interactions with other features are included.

The x-axis of the graphical representations reflect the values of the variable for which the partial dependence function should be plotted. Thereby, the plotted function is estimated by calculating expected values in sets based on the training set. The variables we are interested in assume fictitious values, while the remaining ones are kept unchanged. We are then able to obtain the relationship between the predictors of interest and the predicted outcome.

To be able to provide a discrete analysis through the partial dependence approach, the partial function $\hat{f}_S$ must be estimated through Monte Carlo method by calculating averages in the training data ([22]):

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{24}$$

The $x_C^{(i)}$ represent the actual feature values from the dataset for the predictors in which we are not interested, and n is the number of instances in the dataset.

C Severity RF using an alternative approach

Prior to the partial dependence analysis, a different approach was implemented in order to study the sensitivity of the severity model to the climate variables. Although it may not be so effective as the partial dependence analysis, we will sum up briefly the main conclusions obtained. The marginal behaviour of the climate variables was assessed through predictions for the test set in different scenarios. By manipulating the samples in the test set of those variables we were able to perform a *ceteris paribus* analysis for each of the three isolated predictors, and also for the whole of them. We will only display the main results for the latter one, as the marginal effect of each single variable (and interactions of two) was already assessed through partial dependence and the results of both approaches were similar.

First, in each of the 3 samples considered for each of the 3 features, the observations were all set at the same value. For *Temperature*, samples of $15°C$, $25°C$ and $35°C$ were considered. As for *Humidity*, the samples were set at 25%, 50% and 75%. For *Wind*, we considered the values of 5, 15 and 25 km/h. With this approach we were able to achieve

monotony on the test set predictions for each of the three scenarios, denoted as good, intermediate and critical. This conclusion follows from Figure 25. The good scenario considers the least critical samples of each predictor: $15°C$ for temperature, relative humidity of 75% and a wind intensity of 5 km/h. The intermediate one considers $25°C$, 50% and 15 km/h for the temperature, humidity and wind intensity, respectively. As for the critical scenario the values of $35°C$, 25% and 25 km/h were set. Although the marginal impact of the intermediate scenario compared to the good one is positive, one may think it was not so significant as the one caused by the critical scenario. In fact, this behaviour is compatible with what was proven by the partial dependence analysis of each isolated predictor. The marginal impact of variations in the climate features show evidence of being stronger in the most representative ranges of the predictors. As for the comparison between the intermediate and good scenarios, one could say that the degree of risk does not suffer a huge increase. The observed bar of the graphic corresponds to the original test set prediction with no sample manipulation. In fact, one may deduce through visual inspection of Figure 25 some similarity between this one and the intermediate scenario output.

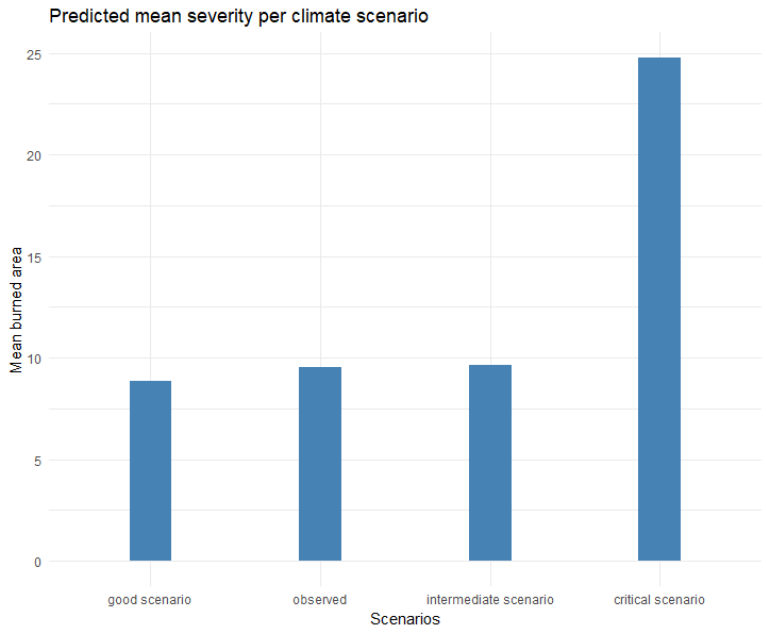


Figure 25: Test set predictions on different scenarios

D Burned Area of Severe Wildfires

A third modelling strand has been left out of the main scope of this project, namely the expected burned area of severe wildfires. In fact, the severity model does not fully discriminates the burned area of those ignitions that were censored at 100 ha, as reported in Section 4.1. Although the dataset only contains 3193 severe observations, developments on this topic were produced through similar methods to those used in the modelling strands previously approached.

First and foremost, insights regarding the GLM that was implemented will follow. As one could expect, some of the 16 predictors were not significant and had to be excluded as it was the case of *Wind*, *Road Density*, *Altitude* and *Precipitation*. The

influences obtained for each predictor exhibit an identical behaviour to that observed in the other models, except the month of October: due to the 103 severe wildfires occurring in this month in 2017, the expected severity of a large ignition in either July or August was estimated to be 23% inferior to one occurring in October.

Although the lack of data may represent an adversity for implementing a machine learning model, a RF was also employed (for the predictive ability of the two models, refer to Table 18). The spatial patterns of the RF predictions were then assessed alongside the climate projections previously considered (see Figures 26 and 27). The most critical municipalities are similar to the ones previously analysed, being Pampilhosa da Serra the one with the highest predicted burned area: 4106 ha are expected to burn, in case a severe ignition occurs. As for the climate scenario, the strongest impact is verified in the municipality of Vila de Rei with a predicted increase of 699 ha.

In light of the findings presented in this section, we consider that future research in this field should be encouraged to deepen our understanding. Despite the lack of observations, we obtained reliable results and compatibility with the models presented in this study.

	GLM	**RF**
RMSE	1449.938	1390.235
Correlation	0.323	0.444

Table 18: Performance metrics for each model - Burned area of severe ignitions

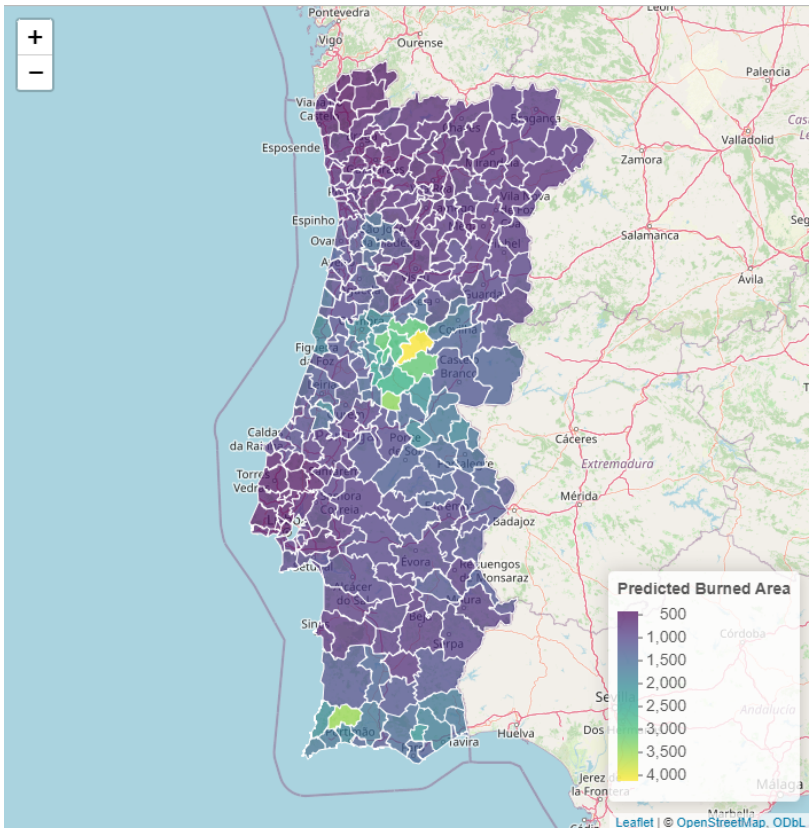


Figure 26: Predicted burned area of a severe ignition by municipality

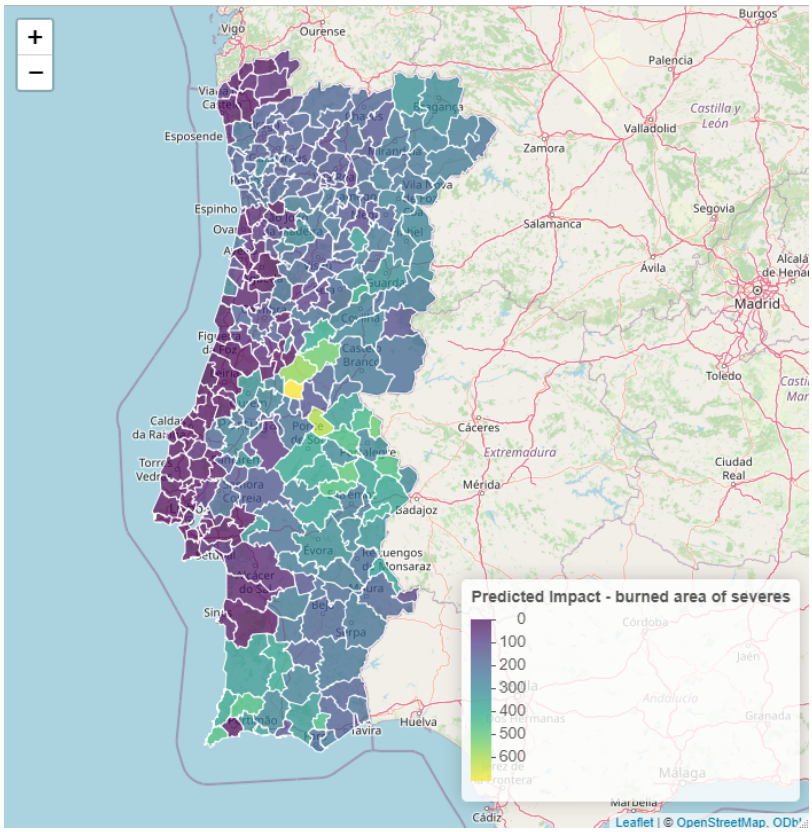


Figure 27: Predicted impact by municipality of the climate scenario with respect to the burned area of a severe ignition