U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Statistics Report

Wildfires of Portugal (2015)

Master's degree in Data Science & Engineering

Fundamentos de Ciência e Engenharia dos Dados

October 2022

Group 6

Henrique Ribeiro, Rojan Aslani, Sónia Ferreira

Professor: António Miguel Gomes

Contents

Introduction

Wildfires or forest fires can have significant impact on mortality and morbidity. Wildfire produces ash and smoke, which is a mixture of air pollutants, of which particulate matter is the principal public health threat. Hence, studying and managing the response to wildfires is important for maintaining resources, protecting people and ecosystems, and reducing air pollution [1]. However, to this day and to our knowledge there is not much information to understand how firefighters can be more effective on the response or even more efficient fighting fires, therefore research is needed to find trends in wildfires intensities.

This work presents an explanatory analysis for all the registered wildfires of 2015, in the districts of Braga and Santarém (Figure 1). The response time, extinction time and cause of the wildfires was explored and analyzed in each district, to understand each of these factors' influence on the total burned area. Knowing whether these factors have an influence on the yearly burned area, better management of resources could save thousands of kilometers of forests every year.

Figure 1. Map of Portugal with Braga and Santarém district colored in blue and red respectively. Braga has an area of 2,673 km² with 21 firefighter stations (1 per 0.12 km²) and Santarém has an area of 6,728 km² with 28 firefighter stations (1 per 0.23 km²).

To do this statistical evaluation, **Six-Steps Statistical Investigation Method** [2] was used. As the name suggests, this method is composed of 6 steps:

1. **Ask a research question**
2. **Design a study and collect data**
3. **Explore the data**
4. **Draw inferences**
5. **Formulate conclusions**
6. **Look back and ahead**

STEP 1 – Research Question

Compare response time, extinction time and cause of the wildfires, to understand their influence on the total burnt area for districts of Braga and Santarém.

Materials and Methods

To do the statistical evaluation and the preparation of the data, python programming language (www.python.org) was used.

STEP 2 – Data Collection

This study is based on the Rural Wildfires in Portugal in the year of 2015. The dataset was taken from *Instituto da Conservação da Natureza e das Florestas* (ICNF) portal. This dataset has 23175 rows of data and 38 columns (Figure 2). After an initial analysis of the complete dataset, since the main goal is to analyze the influence of response time, extinction time and cause of the wildfires on the burnt areas for 2 districts, it was determined that not all data present in the original dataset was necessary. To simplify the process, two causes were chosen to be evaluated, being cause 1 and 4 [3]. New data was derived from the original dataset. As is visible in Figure 2, several fields have null values as entries. After selecting the relevant data, the rows containing null values were removed from the dataset.

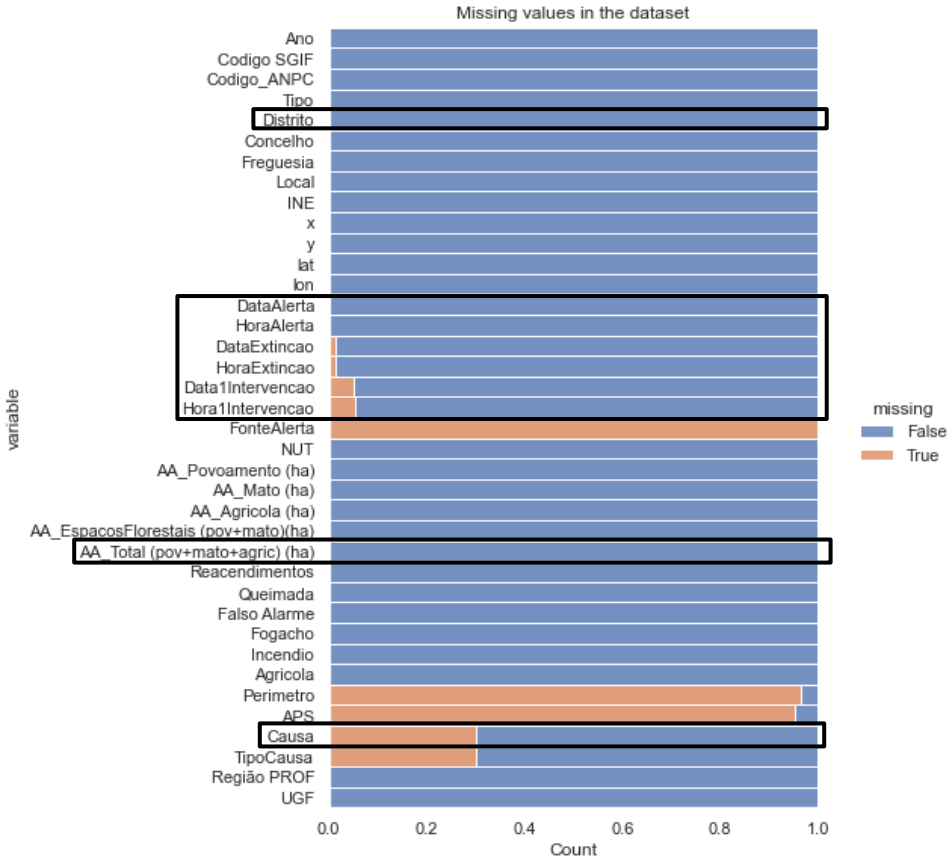

Figure 2 –Variables of the original dataset. The used variables are marked with a black square. The columns containing missing values are demonstrated with orange bars.

Based on the research question of this study, it was decided to create two datasets (one for each district), with the following data:

- District name (Categorical variable – Braga and Santarém)
- Code of cause of fire (Categorical variable – 1 and 4[1])
- Total Burned area (Numerical continuous variable)
- Intervention time (subtraction of Intervention time from Alert time; Numerical discrete variable)
- Extinction time (subtraction of Extinction time from intervention time; Numerical discrete variable)

For intervention and extinction time, all values under 2 minutes were considered input error and were removed, as this time of intervention is unrealistic.

STEP 3 – Data exploration:

The new datasets have 952 rows for Braga and 689 for Santarém. Statistical data regarding both these datasets are presented in Figure 3 and Figure 4. Regarding the intervention time, the districts have a similar mean, and the standard deviation (std) is a value relatively close to the mean. Therefore, we can assume that the data points cluster to the mean and that the values in the datasets are relatively consistent. On the other hand, for extinction time and total burned area, it's observed that the values of the standard deviation are higher, which signifies that the values spread out further from the mean, in both districts, and extreme values become more likely.

	Cause	InterventionTimeMin	ExtinctionTimeMin	TotalArea	District
count	952.000000	952.000000	952.000000	952.000000	952.0
mean	1.633403	13.153361	156.433824	3.341241	1.0
std	1.224984	8.890360	210.972516	11.873399	0.0
min	1.000000	2.000000	10.000000	0.000000	1.0
25%	1.000000	8.000000	70.000000	0.050000	1.0
50%	1.000000	12.000000	110.000000	0.450000	1.0
75%	1.000000	16.000000	175.000000	1.757500	1.0
max	4.000000	152.000000	3260.000000	143.000000	1.0

Figure 3 - Braga numerical summary and statistics.

	Cause	InterventionTimeMin	ExtinctionTimeMin	TotalArea	District
count	689.000000	689.000000	689.000000	689.000000	689.0
mean	2.258345	12.288824	93.483309	3.629433	2.0
std	1.481482	7.610382	122.317615	60.780020	0.0
min	1.000000	2.000000	7.000000	0.000100	2.0
25%	1.000000	7.000000	42.000000	0.015800	2.0
50%	1.000000	11.000000	68.000000	0.067000	2.0
75%	4.000000	16.000000	109.000000	0.313200	2.0
max	4.000000	44.000000	2096.000000	1580.000000	2.0

Figure 4 - Santarém numerical summary and statistics.

[1] Cause 1: Use of fire (codes 1-199)
Cause 4: Incendiarism (code 400-499)

The graphical representation of some of the data is presented in Figure 5, where it is observable in figure (a) and (b) that both intervention time and extinction time have a normal distribution and right skewed distribution, respectively.

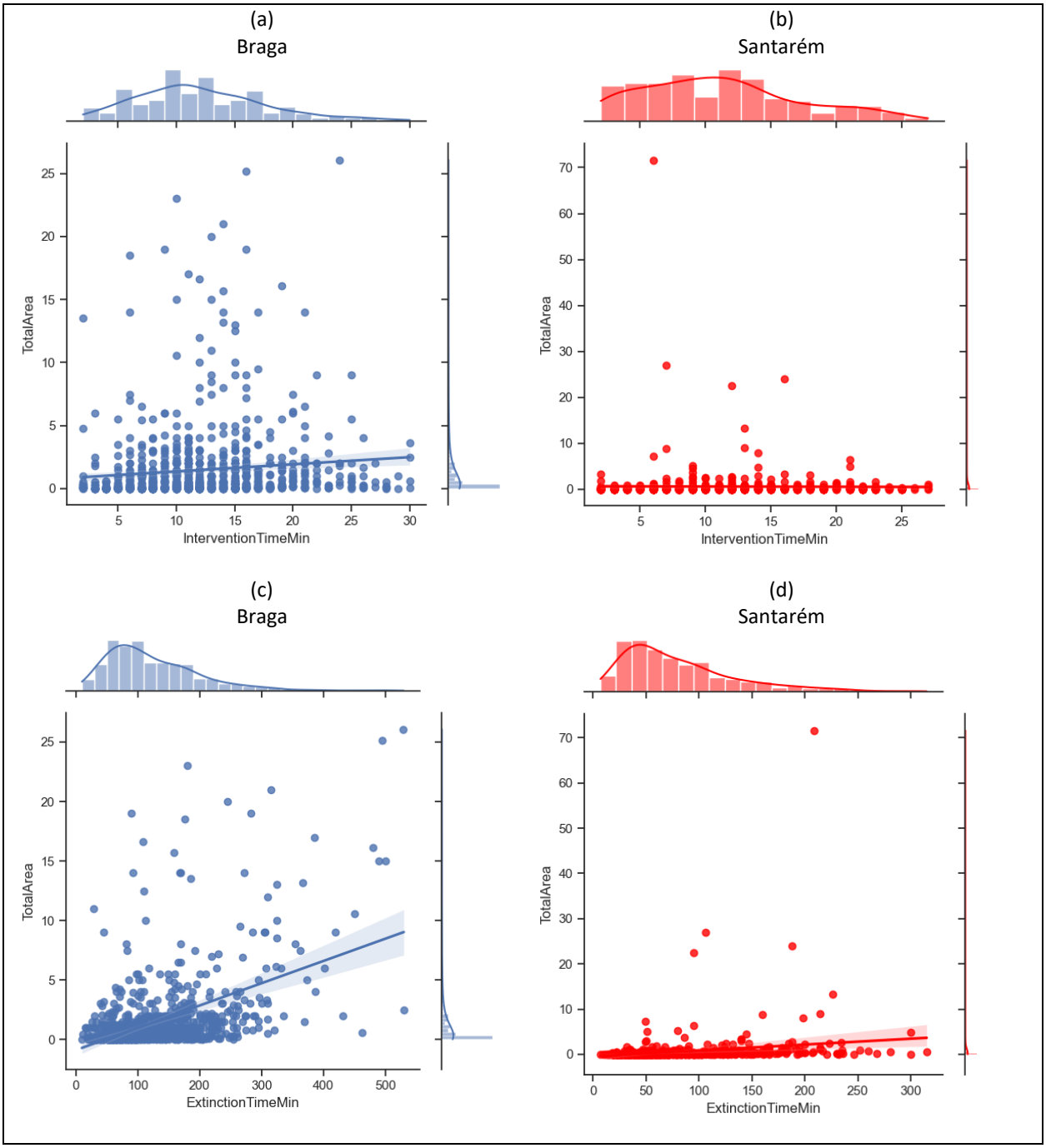

Figure 5. Graphical summary of the data for both districts, where Braga is represented by blue and Santarém by red. (a) Burnt area vs. intervention time in Braga. (b) Burnt area vs. intervention time in Santarém . (c) Burnt area vs. extinction time in Braga. (d) Burnt area vs. extinction time in Santarém.

Removing the outliers

In both datasets, values that were far from the central tendency were observed (outliers), which were interfering with the statistical and graphical evaluations. Therefore, values which exceeded the data by 2 standard deviations above, and 1 standard deviation below, were removed from the datasets (about 7% of the data was removed), as demonstrated in Figure 6.

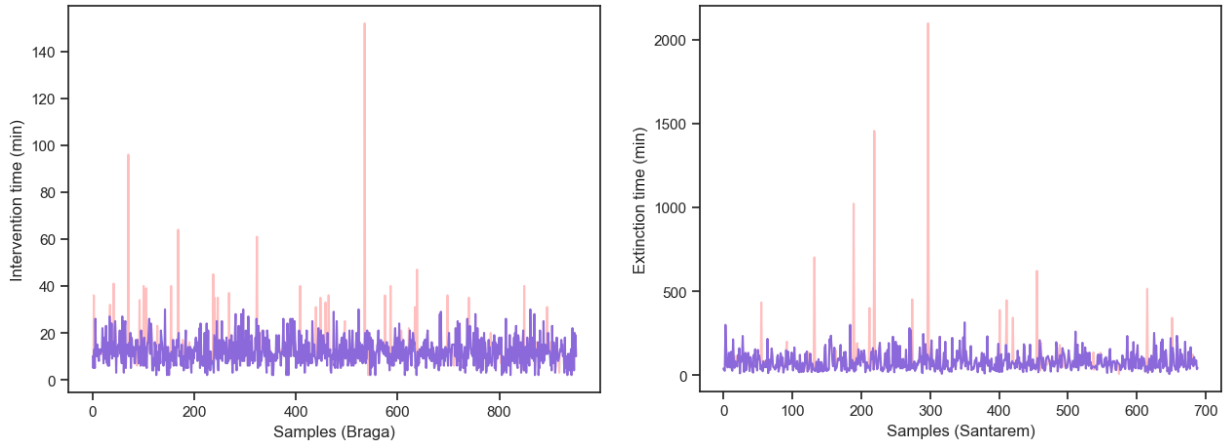

Figure 6. Graphical demonstration of the removal of outliers. Full data is plotted with pink and the new data after the removal of outliers is plotted in purple.

Results and Discussion

After preparing and exploring the data, inferences of the dataset were made. In this step, 3 inferences were made and evaluated.

STEP 4 – Inferences:

Question 1. Does the cause of fire impact the burnt area?

According to Figure 7 a question was raised whether the cause of fire has an association with the average burned area in each district. The 'T-Test' was the chosen statistical hypothesis test for this study, therefore we created two hypothesis – the null hypothesis (H_0) and the alternative hypothesis (H_A).

H_0 = There is no association between mean burnt area and the cause of fire.

H_a = there is an association between mean burnt area and the cause of fire.

We used a significance level of 0.05 (5%) therefore if the p-value is lower than 0.05 we can reject the null hypothesis (H_0); if the p-value is higher we fail to reject H_0.

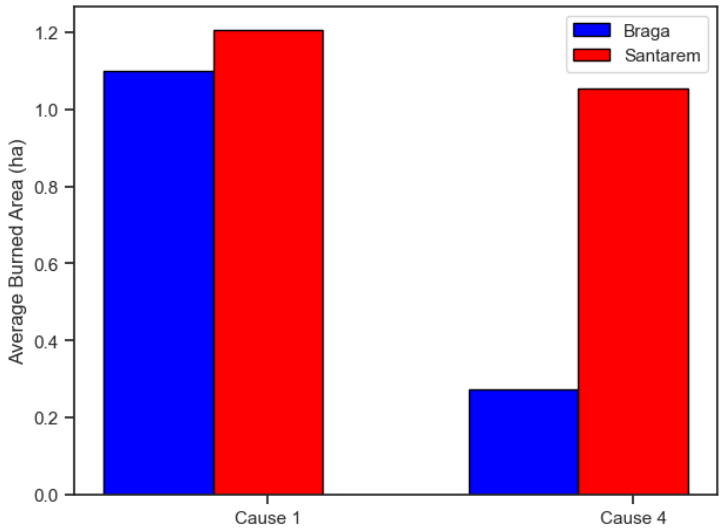

Figure 7. Graphical representation of average burned area for each district, divided by cause of fire.

Looking at our results (Figure 8), we can analyze that for Braga the p-value is 0.1, so we fail to reject the null hypothesis, meaning that there might not be an association between the mean of the burnt area and the cause of fire, hence we cannot make any conclusions regarding the association between the cause of fire and burned area in Braga.

According to the results the p-value for the whole dataset between cause 1 and 4 within Santarém is approximately 0 (less than 5% significance level), which means that H_0 is rejected. Hence, the results are in favor of H_a - there might be an association between burnt area and cause of fire in Santarém.

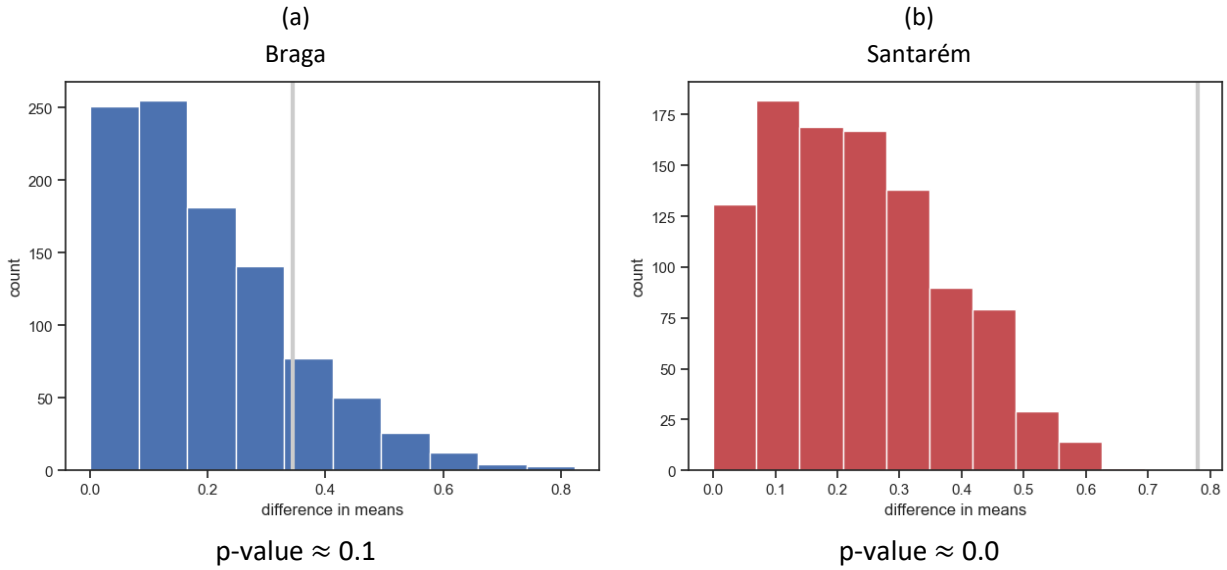

Figure 8. Results of p-value (difference in mean burned area) between causes 1 and 4 for each district. the actual p-value of the entire population is represented with a grey line and the colored data is the representation of the distribution of the values acquired for p-value when repeated for 1000 times on random samples.

Question 2. Does the intervention time impact the burnt area of the district?

The second question was designed to understand whether there is a relationship between the intervention time and the burnt area of the district – does it matter if the firefighters take a longer time to reach the wildfire?

To answer this question, covariance and Pearson's correlation coefficient were used. **Covariance** represents the relationship between two variables when one variable varies. If the value is high, there is a strong relationship between both variables, whereas a low covariance value means a weak relationship. **Pearson's coefficient (r)** is the ratio between the covariance of two covariances and the product of their standard deviations and is used to measure linear correlation. Its value is standardized between -1 and 1. Looking at our covariance levels we can conclude that there might not be a strong correlation due to the smallness of our values (1.8 and - 0.2 for Braga and Santarém, respectively).

If we look at our Pearson's correlation coefficient it will be easier to draw conclusions. Braga has a Pearson's coefficient of 0.1 which indicates a minimal, extremely minute, positive correlation between Intervention Time and Burnt Area. Pearson's correlation coefficient for Santarém is so low (at -0.01s) that we can assume that there is no correlation between Intervention time and Burnt Area (see Figure 5 b and d).

The values for the Pearson's correlation coefficient were done with different sample sizes and the results were consistent.

Question 3. Does the time impact the burnt area of the district?

In the third and final question, we aim to understand the impact that the extinction time has on the burnt area of each district (see Figure 5 a and c). For this, we used the same methods as Question 2: Covariance and Pearson's correlation coefficient.

This time the covariance levels are higher. Braga has a covariance level between total burnt area and extinction time of 122, whereas Santarém's covariance is 38. It is expected that the Pearson's correlation coefficients will also rise comparing to Question 2.

Braga has a Pearson correlation coefficient of 0.49 which indicates a moderately positive correlation between Extinction Time and Burnt Area – we have evidence that the extinction time might impact the Burnt Area.

Santarém has a Pearson correlation coefficient of 0.16 which indicates a low positive correlation between Extinction Time and Burnt Area – we have evidence that the extinction time might impact the burnt area.

A p-value test for the results of r value for Santarém was also done to confirm the results.

H_0 = *There is no association between mean burnt area and* extinction time.

H_a = *there is an association between mean burnt area and* extinction time.

The results of which are represented in Figure 9. It is clear that all the random sample variables are below the actual p-value, hence the p-value is equal to zero, rejecting the null hypothesis. This indicates a possible association between mean burnt area and extinction time in Santarém, confirming the previous results.

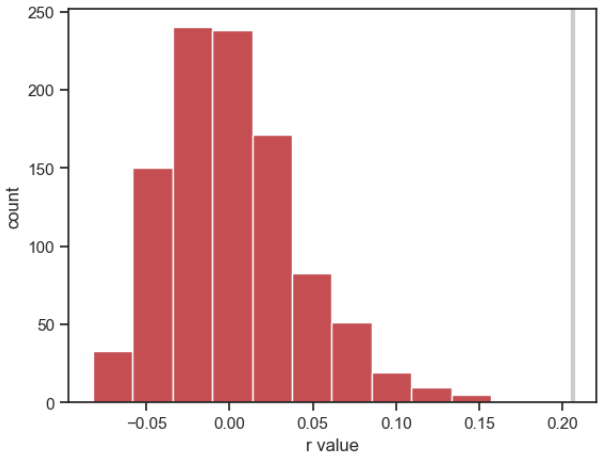

Figure 9. Results of p-value (difference in mean burned area) for extinction time in Santarém. The actual p-value of the entire population is represented with a grey line and the colored data is the representation of the distribution of the values acquired for p-value when repeated for 1000 times on random samples.

STEP 5 - Formulate conclusions

Based on the analysis of the datasets for the research question of this study, the authors concluded that:

Burnt area vs. cause of fire:

- There is no evidence that the cause of the fire impacted the burnt area in Braga.
- There is evidence that the cause of the fire might have an impact on the burnt area in Santarém.

Burnt area vs. intervention time:

- There is no evidence that the intervention time impacted the burnt area in Braga or Santarém.

Burnt area vs. extinction time:

- There is evidence that the extinction time might impact the burnt area in Braga.
- There is evidence that the extinction time might impact the burned area in Santarém.

Although there is strong evidence that the causes of fire that we analyzed have impacted the burnt area in Santarém, since at the time of fighting the wildfires this information is not known, there is no way to adapt the resources dispatched to the site accordingly. Nonetheless, we couldn't reach the same conclusion for Braga.

According to this data, it is not clear that a change on the intervention time will have a direct impact the total of the burnt area of the wildfire. On the other hand, there is evidence in our results that there might be an association between extinction time and burned area in both district, Braga more than Santarém.

Conclusions and Future works

The objective of this work being a statistical analysis of the wildfires of Portugal in 2015 and finding associations between the burned area, cause of fire, extinction time, and response time, we have concluded that there is strong evidence that there might be a relationship between extinction time and burned area. This result is not shocking, as it is assumed that the longer it takes to extinguish a fire, the more area gets burned. On the other hand, no evidence was found to support that the longer firefighters take to start the intervention of fire, the more area gets burned. These results might be incorrect because the extreme values were removed from the database before this analysis, hence, the extreme values of burned area and response time are removed. Therefore, deeper evaluation of the dataset should be done to confirm these results.

STEP 6 – Look back and ahead

It was concluded that the data inconsistency (due to manual entries) or the lack of detailed data can have a direct impact in the assessment of the study of the wildfires.

It would be interesting to have more information regarding, the number of teams that are sent for each wildfire and which technical equipment is sent. This information can directly impact the extinction time and the total burnt area of the wildfire.

Being a preliminary study, focusing the analysis of wildfires of only one year is a good start, but to have a global overview of the research question and provide a clearer analysis, the study could be extended to more years.

Wildfires can have a different number of occurrences during the year, depending on several variables like the temperature values, wind, humidity, and annual seasonality, therefore it appears more relevant to compare the data across the past years and months than across 1 year. Because of climate change, ideally the used data should be from the recent years.

Bibliography

[1] WHO, "Wildfires," 2022. [Online]. Available: https://www.who.int/health-topics/wildfires#tab=tab_2.

[2] N. T. e. al., Introduction to Statistical Investigations, Wiley, 2016.

[3] J. B. d. Carvalho, "CODIFICAÇÃO E DEFINIÇÃO DAS CATEGORIAS DAS CAUSAS," DGRF, Lisboa, 2003.

Listing of dataset

make_database_by_district.py	Cleans original dataset and creates .xlsx files of the districts
Group6_statistics.ipynb	Does all the statistical analysis of Braga and Santarém datasets
OriginalDatasetExplore.ipynb	Exploration of the original dataset
Braga.xlsx	File containing wildfires of Braga for cause 1 and 4 in 2015
Santarem .xlsx	File containing wildfires of Santarém for cause 1 and 4 in 2015
ListaIncendios_2015.xlsx	Dataset of all wildfires of Portugal in 2015
Group6_presentation.pdf	Presentation slides of the work