Applying Hierarchical Clustering to Wildfires in Spain

Núria Orgaz

***Abstract*—Better understanding on wildfires is crucial for prevention, detection, suppression and management activities. This paper intends to elucidate patterns on forest fires in Spain, extending from 2001 to 2015, as similar research has already been done implemented on Portuguese fires. We found worsening on fire extension after the 2008 recession although the total number went down. Additionally, we found significant clusters that showed differences between them. The results of this study will support effective fire management mitigation and suppression, including identifying preventive measures to reduce wildfires.**

***Index Terms*—Forest fires, Hierarchical Clustering, Unsupervised Learning, Wildfires**

I. Introduction

Fire is a crucial process in the planet and plays necessary roles in terrestrial, atmospheric, and aquatic systems.[1] However, it has serious repercussions on soil, vegetation structure and functioning, water quality and availability, air pollution, human health and economic losses.[1][2] In addition, the Iberian peninsula accounts for more than 50% of the fires registered in Europe, with the majority of them occurring during the period form May to October.[3] For these reasons, a better insight of fires is essential for fire prevention, detection, suppression and management activities.[4]

Similar research was applied on Portugal where they used clustering to show geographical differences.[5][6] We believe we can get similar results using hierarchical clustering and geographic analysis.

II. Methods

A. Dataset

Dataset elaborated by Civio [7] from data of the General Statistics on Forest Fires (EGIF), compiled by the Spanish Wildfire Information Coordination Centre (CCINIF) with information provided by each Autonomous Community yearly.

B. Data preparation

We began by exploring the data by analysing the percentage of missing values. We found that the percentage of missing values on four columns was over 50% of the total values and had to be deleted. Those represented the deaths, injured, expenses, and economic losses. Columns that represented categorical data with number had their values renamed to strings to facilitate both the analysis and further transformations. We also created a new categorical column using the burnt area values to classify the fires into three categories of fire by magnitude. These are outbreak (less than 1 ha), large forest fire (more than 500 ha), and forest fire (between 1 ha and 500 ha). This was later used to facilitate data visualizations.

C. Clustering

We performed unsupervised learning on two subsets of the data. One contained only the numerical data, which was burnt area, time to control the fire, time to extinguish the fire, personnel, and means. The other subset also contained the province column and the cause column, which were transformed into dummy variables using Pandas[8][9]. This allowed the clustering algorithm to use those values in the calculations.

Using the library Scikit-learn[10], we applied normalization and Principal Component Analysis (PCA)[11] to the numerical values only subset, and the subset that included the dummy variables. Number of components for PCA was selected accounting for a cumulative explained variance of 99% or more[12].

A hierarchical agglomerative clustering algorithm based on ward linkage was used, as it has been applied on similar data before[13]. We generated a dendrogram for each subset using scipy to select the number of clusters.

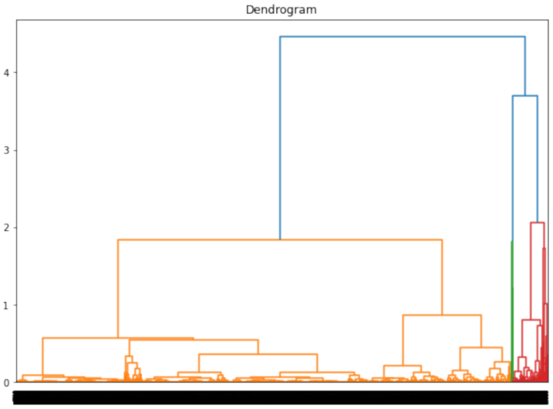

Fig. 1. Agglomerative dendrogram showing the formed clusters on the numerical only subset

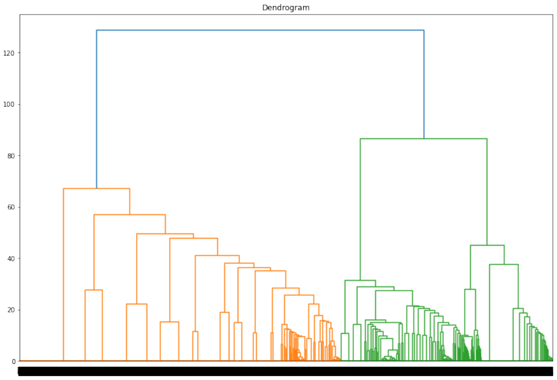

Fig. 2. Agglomerative dendrogram showing the formed clusters on the numerical and dummy subset

TABLE I
TOTAL BURNT AREA DIVIDED BY YEAR AND CAUSE

Year	Accidental	Arson	Lightning	Reproduced	Unknown
2001	22 593	44 267	8 886	730	10 055
2002	19 383	59 751	718	1 062	16 478
2003	34 373	57 595	22 770	1 336	15 837
2004	25 879	92 668	2 975	1 755	7 350
2005	43 671	111 857	12 602	2 270	14 439
2006	25 981	103 719	4 648	5 699	12 389
2007	19 721	56 510	2 541	280	3 620
2008	14 983	28 870	331	143	2 951
2009	34 193	54 227	18 149	1 743	6 296
2010	6 019	43 914	464	666	2 114
2011	12 095	72 375	1 380	8 834	4 931
2012	102 998	97 990	1 600	2 643	9 207
2013	19 012	35 155	1 597	579	3 698
2014	17 580	25 414	2 195	171	1 795
2015	23 295	63 123	12 328	5 765	3 171

Due to lack of better computing resources, we had to take a sample of 25% before applying the algorithm.

III. Results

A. Descriptive analysis

Upon analysing the data, we found out that there are very large discrepancies in personnel and equipment expenditure depending on the cause of the fire. The forest fires that require the most are the ones caused by lightning. On average, they use up more than double of personnel and almost three times the means than fires caused by arson or those where the cause is unknown. Those means used include aerial means such as specialised airplanes and helicopters that are used for the tasks of water and/or retardant release, coordination, observation and transportation of personnel. [14]

However, those kinds of fires are not the ones that cause the most harm. By far, arson fires burn the most hectares. Without taking into account large forest fires, with over 500 ha burnt, arson calcined over 600 000 ha, whereas the next most destructive cause, accidents and negligence, scorched a little over 400 000 ha in total, taking into account large forest fires too.

The total wildfire count has gone down, but the average fire size has increased. The average control time and average extinction time have also augmented. It is unclear nonetheless whether the burnt hectares have increased leading to higher control and extinction time, or conversely, the longer time taken to control and extinguish have caused the average burnt area to increase.

We can see on figure 6 that fire location, especially when taking into account the cause of the fire, is not random. There's a large gap where the provinces of Soria, Teruel, Cuenca and Albacete are. It is also remarkable the large concentration of reproduced fires on the north-western regions of Galicia and Asturias. Almost all of the spots regarding this type of fire are located in that region.

B. Hierarchical Clustering

The numerical and dummy subset formed two very balanced clusters, as shown on figure 4. However, it got a silhouette score[15] of 0.27. When plotting the clusters on map, they showed no pattern or difference across them either.

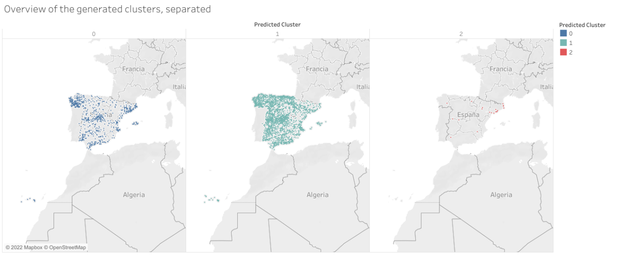
Overview of the generated clusters, separated
Predicted Cluster
0
1
2
Predicted Cluster
0
1
2
Francia
Italia
España
Algeria
© 2022 Mapbox © OpenStreetMap

Fig. 3. Overview of the generated clusters for the numerical only subset

The numerical only subset, however, formed three clusters of very different sizes, as shown on figure 3, and this division got a silhouette score of 0.79. When we plotted the results, the clusters showed a distinct distribution. The smallest cluster, cluster 2, is particularly concentrated on the Catalonian coast, on the provinces of Barcelona, Girona and Tarragona. This cluster also takes up about a third of the total burnt area on accidental fires, about 25 000 ha, and a similar area is taken up too on arson fires. Cluster 0 shows a remarkable pattern. We can see that it is concentrated near large population areas, such as Madrid and surrounding areas, the whole Mediterranean coast, but also Galicia, which is not a densely populated region but the descriptive analysis showed it has a high concentration of wildfires and shows differences in the proportion of fire causes compared to the other regions. Additionally, this cluster has a similar pattern to that of the forestry map reported by the Spanish Ministry for ecological transition. In this case, this cluster matches the soil labelled as "Arbolado" (woodland in Spanish) in figure X. This would explain much better the relation of the locations of the data. Lastly, cluster 1 is the cluster that holds the most data. Unfortunately, there is not such a defined pattern, mostly due to the size of this cluster compared to the others.

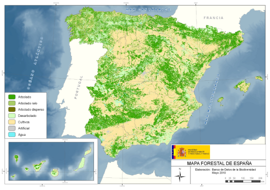

Fig. 4. Spain Forestry Map[16]

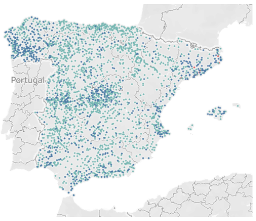

Fig. 5. All the clusters overlapped in which they show a similar pattern to that of the forestry map

IV. Discussion

The provinces where there were less forest fires, Soria, Teruel, Cuenca and Albacete, are not doing something different compared to the other provinces. They take part of what is called "España Vaciada", or Empty Spain. These regions are rural and have undergone great depopulation. The lack of population is the most likely cause of the small number of fires, not different fire prevention programs.

The economic recession of 2008 led to a series of economic cutbacks that affected the budgets of fire prevention and extinguishing.[17] This caused or at least affected the surge of

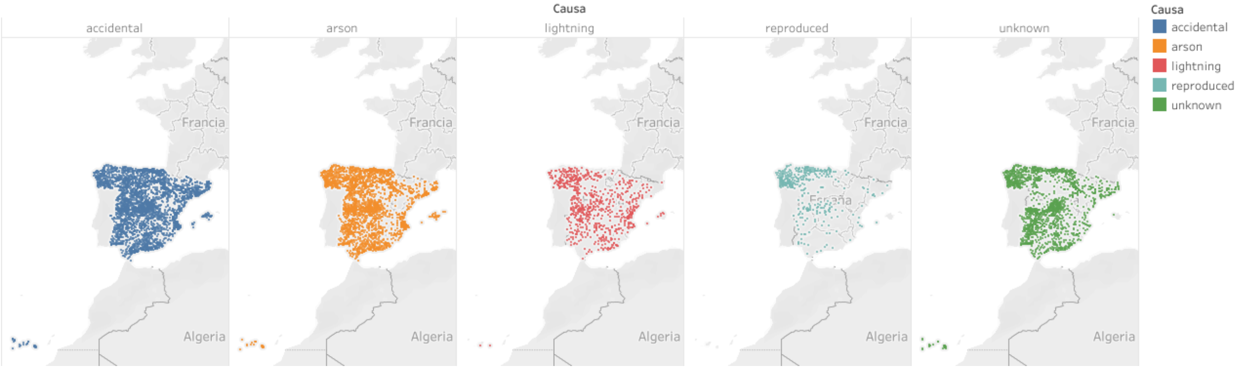

Fig. 6. Wildfire locations by cause

burnt hectares in the following years. The large concentration of reproduced fires in the north-western region is also most likely due to mismanagement, not geographical or climatic differences. This kind of fire happening means that another fire was thought to be controlled or extinguished but started again. This differences deserves an in-depth study done specifically in that region that would perhaps shed light into the causes.

The evidence shows an increase of time taken to control and extinguish, and this presumably produced larger fires. Considering this, there needs to be a decision to increase resources, both human and technological, allocated to prevent and mitigate wildfire harm.

Regarding the unsupervised learning, it would be interesting to perform clustering on more data, for which stronger computing resources would be needed. Likewise, different algorithms could be applied to show results that support our results and conclusions. Moreover, a deeper analysis using the common tree species that form the woodland, as well as climate and meteorological conditions, could add more to the clustering results obtained

References

[1] J. Bedia, S. Herrera, and J. M. Gutiérrez. "Assessing the predictability of fire occurrence and area burned across phytoclimatic regions in Spain". In: *Natural Hazards and Earth System Sciences* 14.1 (2014), pp. 53–66. DOI: 10.5194/nhess-14-53-2014.

[2] R.M.B. Santos, L.F. Sanches Fernandes, S.G.P. Varandas, M.G. Pereira, R. Sousa, A. Teixeira, M. Lopes-Lima, R.M.V. Cortes, and F.A.L. Pacheco. "Impacts of climate change and land-use scenarios on Margaritifera margaritifera, an environmental indicator and endangered species". In: *Science of The Total Environment* 511 (2015), pp. 477–488. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2014.12.090. URL: https://www.sciencedirect.com/science/article/pii/S0048969714018014.

[3] Mário G. Pereira, Ricardo M. Trigo, Carlos C. da Camara, José M.C. Pereira, and Solange M. Leite. "Synoptic patterns associated with large summer forest fires in Portugal". In: *Agricultural and Forest Meteorology* 129.1 (2005), pp. 11–25. ISSN: 0168-1923. DOI: https://doi.org/10.1016/j.agrformet.2004.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0168192305000043.

[4] Yves Bergeron, Alain Leduc, Brian Harvey, and Sylvie Gauthier. "Natural fire regime: A guide for sustainable management of the Canadian Boreal Forest". In: *Silva Fennica* 36.1 (Jan. 2002). DOI: 10.14214/sf.553.

[5] Ana C. Meira Castro, Adélia Nunes, António Sousa, and Luciano Lourenço. "Mapping the causes of forest fires in Portugal by Clustering Analysis". In: *Geosciences* 10.2 (2020), p. 53. DOI: 10.3390/geosciences10020053.

[6] Joana Parente, Mário G. Pereira, and Marj Tonini. "Space-time clustering analysis of wildfires: The influence of dataset characteristics, fire prevention policy decisions, weather and climate". In: *Science of The Total Environment* 559 (2016), pp. 151–165. DOI: 10.1016/j.scitotenv.2016.03.129.

[7] *Todos los incendios forestales*. URL: https://datos.civio.es/dataset/todos-los-incendios-forestales/.

[8] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[9] The pandas development team. *pandas-dev/pandas: Pandas*. Version 1.3.4. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[10] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] Ferath Kherif and Adeliya Latypova. "Principal component analysis". In: *Machine Learning* (2020), pp. 209–225. DOI: 10.1016/b978-0-12-815739-8.00012-2.

[12] Jake VanderPlas. *Python Data Science Handbook*. O'Reilly Media, 2016.

[13] Vijaya, Shweta Sharma, and Neha Batra. "Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering". In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (2019). DOI: 10.1109/comitcon.2019.8862232.

[14] *Los Medios Aéreos de Extinción de Incendios Forestales*. URL: https://www.miteco.gob.es/es/biodiversidad/temas/incendios-forestales/extincion/medios_aereos.aspx.

[15] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[16] URL: https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/mfe50.aspx.

[17] Enric Sopena Daganzo. “El fuego arrasa España”. In: (2012). URL: https://dialnet.unirioja.es/servlet/articulo?codigo=4044593.