Wildfire Risk Mapping Based on Multi-source Data and Machine Learning

Ghinevra Comiti(✉), Paul-Antoine Bisgambiglia, and Paul Bisgambiglia

University of Corsica, CNRS UMR SPE, Corte, Corse, France
{COMITI_G,BISGAMBIGLIA_PA}@univ-corse.fr

Abstract. The management and prevention of forest fires are crucial in fire-prone regions such as Corsica, a French island in the Mediterranean. In this study, an approach to mapping wildfire vulnerability is presented using different data sources, including meteorological, temporal, geographical and economic datasets. These heterogeneous datasets are seamlessly integrated to produce a comprehensive forest fire vulnerability map for Corsica. The methodology involves the collection and pre-processing of a variety of data, such as historical forest fire events, meteorological variables, land cover data, socio-economic indicators and temporal factors. Machine learning models are used to visualise the complex relationships between these variables and predict wildfire susceptibility. Finally, we were able to create a daily fire susceptibility map for the island of Corsica.

Keywords: Machine Learning · Wildfire · Map

1 Introduction

Forest fires have a catastrophic impact on the environment, the economy and even people's lives. In recent decades, these fires have become more intense, more frequent and more deadly under the influence of climate change. Therefore, understanding and preventing wildfires has become an important scientific topic and many papers have addressed this issue [2].

The island of Corsica, our study island, is particularly affected by this problem: More than 75,000 hectares of forest burned in 2021 alone. The Mediterranean climate (cold winters with little precipitation and very hot and dry summers) favours the development of forest fires. For this reason, some researchers at the University of Corsica had the idea of launching the GOLIAT project, a multidisciplinary project aimed at better understanding forest fires in order to prevent them. The acronym GOLIAT stands for "Group of Tools for Fire Fighting and Regional Planning". It is a multidisciplinary project involving researchers from various fields: Physics, Biology, Economics, History, Computer Science, etc.

Supported by "Collectivité de Corse" CdC through the GOLIAT project.

J.-L. Guisado-Lizar et al. (Eds.): SIMUtools 2023, LNICST 519, pp. 110–119, 2024.
https://doi.org/10.1007/978-3-031-57523-5_9

The aim is to put scientific knowledge at the service of professionals in the field (firefighters, foresters, etc.) by creating tools that these professionals can use on a daily basis. This is why many employees are also members of the GOLIAT project.

In this context, our goal was to create a forest fire risk map: a map that would show us the fire risk for every place on the island for every day. But first we had to answer a (seemingly) simple question: How can we define "risk"?

Academically, the concept of "risk" refers to a state of uncertainty in which certain possibilities include the occurrence of a loss, disaster or other undesirable outcome [19]. It is an integral part of various areas such as finance, business, insurance and everyday life. Risk consists of two key components: probability and impact.

Probability refers to the likelihood or chance that a certain event or outcome will occur. Impact refers to the extent or severity of the possible consequences of that event or outcome. This definition is very broad and general. If we apply it to the prevention of forest fires, the "probability" component would represent the likelihood of a fire occurring and the "impact" component would represent the vulnerability of a particular place: the presence of people, infrastructures that should be protected, etc.

Although this is a two-component problem, our current software focuses mainly on predicting the probability of fire occurrence. However, taking into account the vulnerability of the location could be an interesting improvement for the future of the method.

To create our forest fire risk map, we used a 2×2 km resolution grid to divide our area. Each grid cell is assigned a forest fire risk between 0 and 1, with 0 representing a very low risk and 1 representing a very high risk. Each grid cell is also linked to the data of the respective day and shows whether a forest fire has occurred or not. This data is fed into a machine learning model for classification. This model is able to tell us, based on the data we provide it with, how likely it is that a fire will break out on a particular date and location. Of course, these models can make mistakes. Therefore, we want to try out different models and choose the best one in terms of certain precision metrics, which we will discuss later in this document.

Once we have an optimal machine learning model, we select a day and run the model for each cell in our grid for that day. Each cell is associated with a forest fire risk between 0 and 1, which we use to create a choropleth map. We now have a tool that allows us to create a wildfire risk map for any given day, assuming the data is available.

In the following sections, we will first look at the background work. Then we will present our solution in detail, starting with an explanation of the methodology used and ending with the applications of our method.

2 Background

The quantity and nature of the data to be processed are essential parameters in selecting a machine learning model. Related works have shown a wide variety of data to be useful [6,13,15]. The relevant data can vary significantly from one study to another and appear to depend on the specific characteristics of each location (geography, climate, population, etc.). Therefore, we relied on studies conducted in regions similar to Corsica (high altitude variation, Mediterranean climate, low population density) and identified the relevant variables [6,9,15,16,18]. They are listed below:

- The coordinates of the fire: the fire is identified either by a point or by its position on a grid. Often, the position on a grid dividing the relevant territory is used when we are trying to create a map, as is the case in our situation.
- Temperature: most commonly, the average temperature of the given day is used.
- Wind strength: here, as well, the average wind strength over the day is taken into consideration.
- Wetness: Wetness plays a great role on ignition and spreading probability of the fire. Indeed, wet fuel will ignite less easily and more slowly.
- Topography: altitude affects the vegetation cover and influences the available fuel and consequently the potential for a fire to ignite and spread [12,15]. In addition, the degree of slope also plays an important role, as a fire will spread more quickly if it follows a slope.
- Land Use: this is about categorizing the type of land on the island [9,15,16]. Certain types of land, such as urban or cultivated areas, are less prone to fires than others, such as bushes.
- Day of the week: in Europe, more than 80% of fires are of human origin [3,6,18]. This means that human activity has a significant impact on the risk of fire. The day of the week has a major influence on human activity. Towards the end of the week, for example, people spend more time in nature, where their activities can lead to the ignition of fires. The day of the week is therefore likely to be an important criterion.
- Unemployment rate: in certain studies, it is assumed that the unemployment rate is a factor that contributes to the ignition of forest fires [12]. This could be because unemployed people are less employed and therefore have more time to fuel conflicts with their neighbours, e.g. by starting fires on their properties.
- Road density: some studies have also shown the effects of road density [9,15]. Roads concentrate human activities, especially in summer, in touristic places like Corsica. Therefore, this seems to be an interesting parameter to consider.
- Vegetation: vegetation serves as the medium for fire propagation [9]. Thus, it has a direct impact on the probability of ignition as well as the spread of a potential forest fire.

We were able to obtain this data from the relevant authorities. Then we had to use it to train the model. But first we had to select model. There is indeed

a wide range of machine learning models. Since our goal here is to classify cell grids, we focused on classification methods. Also for efficiency and transparency reasons, we focused on supervised machine learning. After reviewing related work and the literature, a few potential models emerged:

- Random Forest: a Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to improve predictive accuracy and reduce overfitting [14,17]. Random Forest creates multiple random subsets of the dataset through a process called bootstrapping. For each subset, it builds a decision tree using a random subset of features at each node of the tree. When making predictions, each tree in the forest independently predicts the output, via majority voting. Finally, Random Forest combines the predictions from all the individual trees to produce a more robust and accurate prediction.
- Adaboost: AdaBoost, short for Adaptive Boosting, is another ensemble machine learning algorithm used for binary classification [10]. It works a little bit like Random Forest, except Adaboost trains its learners sequentially and assigns weight to the data, paying more attention to often misclassified data. It makes Adaboost more powerful, but also more prone to overfitting.
- Gradient Boosting: Gradient boosting is a machine learning technique that builds a predictive model in a stage-wise fashion by combining the predictions of multiple weak models, typically decision trees [11]. Like the two previous models, it combines the predictions of several individual models to improve overall predictive performance.

The workflow is recapitulated in Fig. 1. We made the choice to try all of these models and chose the more performant one to use in our application. Our definition of "performant" will be detailed afterward.

3 Method

In this section, we will give further details on the method we used to create our map.

As mentioned in previous sections, we selected and gathered a wide range of data. Those data were obtained from relevant authorities: Copernicus [1] for weather data, the French governmental opensource data website "data.gouv" [5] for road, topography and land use, French forest database [4] for vegetation and the National Institute of Statistics and Economical Studies [7] for unemployment rate.

To locate fire and relevant data, we decided to divide the map to a 2×2 km grid. We chose this particular grid format because it's the one that is already used to locate fires in France. To each grid cell are assigned data of the area. We consider that each cell grid, each day, is a data entry [8]. We assign to this entry all the relevant data of the grid cell, and a boolean "fire", set to "True" if a fire happened at this time and location and to "False" instead.

Using this method, we found ourselves with a tremendous amount of data where "fire" is set to "False", because obviously, there are more days without fire

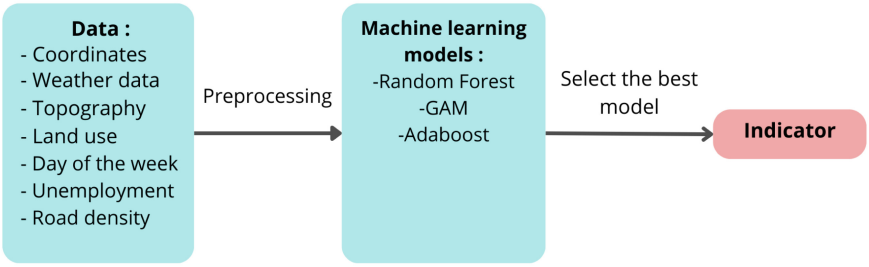

Fig. 1. Strategy used to answer our problematic.

than with fire. If one wants to do machine learning, this huge data imbalance can be problematic. Indeed, it could introduce a bias, and the model might be bad at generalization. Thus, it is essential to have a pretty well-balanced dataset. To this end, we deleted part of our "no fire" data. We now have a balanced dataset with 27 500 examples of fires and 27 500 examples of "no fire", that will help us to train our model. Now, the next step would be to choose a model.

Once the data was prepared, we had to train the models and see which one performed best. We used the scikit-learn library. The first step was dividing the data into two sets, a training set and a validation set. Then, we created instances of the models we're interested in. To capable the parameters, we used the GridSearchCV function of the scikit-learn library. This function tests a model with different sets of parameters, to determine which parameters give the best results. We ensure that all our models have optimal parameter settings, enabling us to easily compare them.

To compare the models, we decided to use their AUC. AUC stands for "Area Under the (Receiver Operating Characteristic) Curve." It is a commonly used metric in machine learning and statistics to evaluate the performance of a binary classification model. The Receiver Operating Characteristic (ROC) curve is a graphical representation of a model's ability to distinguish between the two classes it is trying to classify, typically the positive class (e.g., presence of a fire) and the negative class (e.g., absence of a fire). The AUC varies between 0 and 1, with 0.5 being the performance achieved by random choice.

After a careful comparison of the models, the Random Forest appeared to be the one with the best AUC. It was around 0.80, which is considered quite good. Gradient Boosting was behind with around 0.70, and Adaboost was pretty close with a score of about 0.75.

Once we had selected our model, we also tried to optimize its recall. Recall is a performance metric that allows you to evaluate the effectiveness of a binary classification model, especially when one of the classes is of greater importance or when you want to minimize the number of false negatives. This is our case: it would indeed be dramatic to "miss" a fire by underestimating the risk in a given area. On the other hand, it would also be problematic to overestimate the risk, as this could lead to a fragmentation of forces. We have therefore tried to find a kind of balance between recall and another metric, precision, which focuses on minimising false positives, while still prioritizing recall. A target that has been chosen in similar articles and seems quite reasonable to us (and which we have therefore tried to replicate) is to achieve a recall of over 0.9 and a precision of over 0.6, which is already an honourable value in this field [8].

4 Application

We then generated the risk map. We edited it for a day for which we had data, the 8th of July 2009. Then, we compared it with the fires that really happened this day. The result is represented in Fig. 2, with the blue stars being the location of the actual fires happening that day.

As we can observe, the predictions for this day seem pretty accurate. Most of the fires are located in places where the fire risk is above 0.7, or 0.6 for three of them. Only one is located in a place where the fire risk is below that (about 0.4 to 0.5). No fire happened in locations where the fire risk is above 0.8, but we can argue that those places are already identified by the capable authorities as being very prone to fire, and are thus very closely monitored.

The prefecture of the island already publishes its own forest fire risk map. It is published every summer day on the prefecture's website. This map, based exclusively on meteorological data, is shown in the Fig. 3. As we can see, it is far less accurate than our map, which makes it less useful for the staff on the ground. This is actually the main reason why this project was launched in the first place.

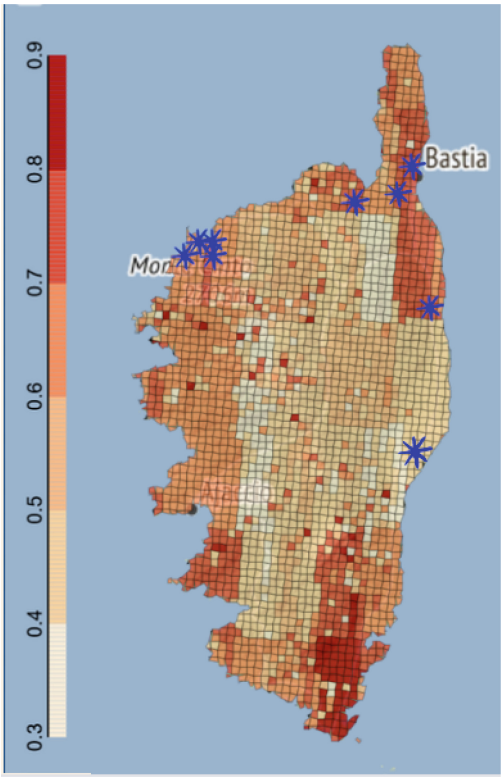

Fig. 2. The fire risk map edited for the 8th of July 2009. The blue stars represent the actual fires. (Color figure online)

We were also able to obtain from the model the contribution of each variable to the result. These data may be of crucial importance. Indeed, knowing which variables have an influence on fire susceptibility allows us to prevent fire more efficiently. These contributions are represented in Fig. 4.

As we can observe in Fig. 4, the contribution can be very different from a variable to another. Road density, time of the day, geographical and meteorological values seem to play a pretty important role, as well as altitude variables. On the other hand, day of the week, unemployment rate, vegetation and land use don't seem very meaningful. We can try to explain some of those results. For instance, the lack of importance of the "day of the week" variale can be explained because fires often happen during summer, when there are a lot of tourists on the island. The day of the week is way less meaningful for this population, who doesn't work at the moment and is thus more free. For the vegetation, we can argue that the Corsican vegetation is pretty homogenous. Furthermore, the vast majority of fires are forest fire (more than 90% of them) which can explain why land use isn't of great importance as well.

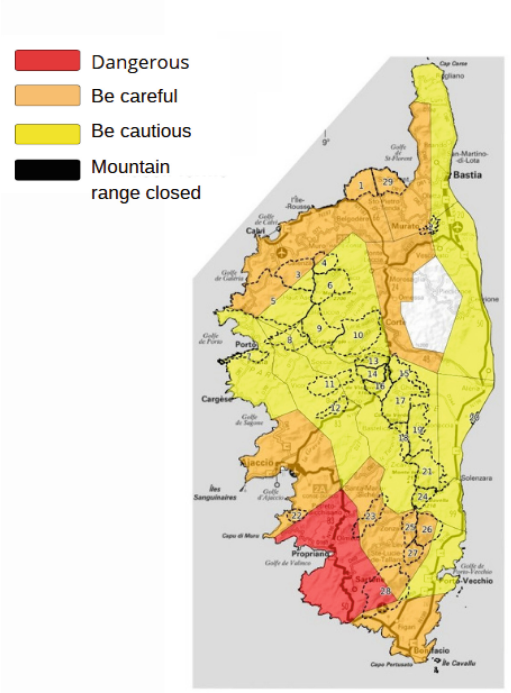

Fig. 3. The prefecture's map.

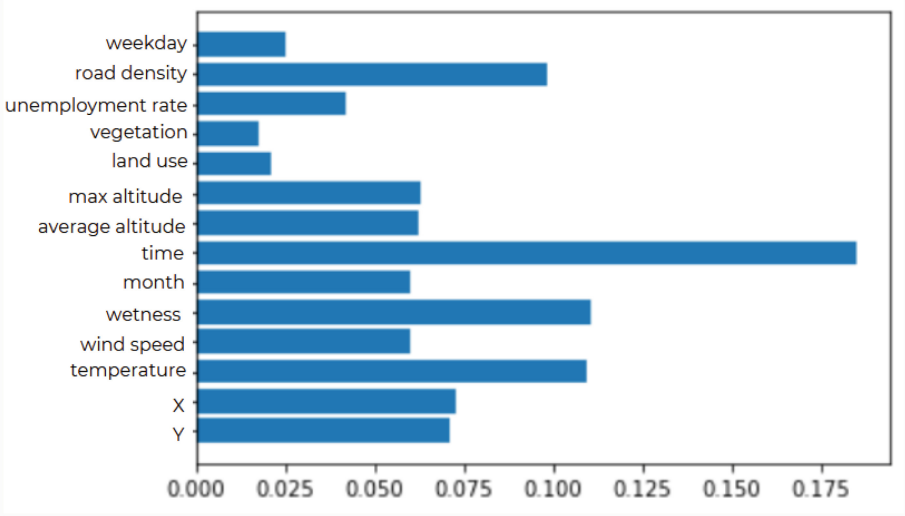

Fig. 4. The relative contribution of variables to our model.

We can conclude that our method can help us to prevent fires, and in addition, to better understand our territory and how to protect it.

5 Conclusion

To summarise, we have proposed a method to predict the probability of forest fires for a given day at different locations in Corsica. This method involves the collection and processing of a large dataset containing historical records of forest fires, meteorological parameters, land cover information, socio-economic indicators and temporal aspects. Using machine learning models, we calculate a forest fire risk between 0 and 1. This risk is calculated for each of our grid cells. We then use these numbers to create a choropleth map. The map created is a great tool to help decision makers make better decisions, and for the public to get information before going on a hike or other outdoor activity.

The prospects for this project include a phase of method validation and consolidation, in which we will assess the accuracy of the method in more detail. We will also take a closer look at the current map and the predictions and accuracy differences between the two maps. In addition, we would like to combine the map's hazard prediction with more traditional simulation techniques for the high fire risk areas identified in the map. This would allow us to predict fire risk more efficiently. Finally, before the map is made available to the public, it needs to be validated in practise.

References

1. Accueil copernicus. https://www.copernicus.eu/en
2. Evolving Risk of Wildfires in Europe - Thematic paper by the European Science & Technology Advisory Group (E-STAG). https://www.undrr.org/publication/evolving-risk-wildfires-europe-thematic-paper-european-science-technology-advisory
3. Fire — Free Full-Text — Forest Fire Susceptibility and Risk Mapping Using Social/Infrastructural Vulnerability and Environmental Variables. https://www.mdpi.com/2571-6255/2/3/50
4. French forest database, national institute of geograophy. https://geoservices.ign.fr/bdforet
5. French governemental data. https://www.data.gouv.fr/fr/
6. Full article: Modelling temporal variation of fire-occurrence towards the dynamic prediction of human wildfire ignition danger in northeast Spain. https://www.tandfonline.com/doi/full/10.1080/19475705.2018.1526219
7. National institute of statistics and economical studies. https://www.insee.fr/fr/accueil
8. Apostolakis, A., Girtsou, S., Giannopoulos, G., Bartsotas, N.S., Kontoes, C.: Estimating next day's forest fire risk via a complete machine learning methodology. Remote Sens. **14**(5), 1222 (2022). https://doi.org/10.3390/rs14051222. https://www.mdpi.com/2072-4292/14/5/1222

9. Carmona, A., González, M.E., Nahuelhual, L., Silva, J.: Spatio-temporal effects of human drivers on fire danger in Mediterranean Chile. Bosque **33**(3), 31–32 (2012). https://doi.org/10.4067/S0717-92002012000300016
10. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000). https://doi.org/10.1214/aos/1016218223. https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-2/Additive-logistic-regression--a-statistical-view-of-boosting-With/10.1214/aos/1016218223.full
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. (2001)
12. Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Aryal, J.: Forest fire susceptibility and risk mapping using social/infrastructural vulnerability and environmental variables. Fire **2**(3), 50 (2019). https://doi.org/10.3390/fire2030050. https://www.mdpi.com/2571-6255/2/3/50
13. Jain, P., Coogan, S.C., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D.: A review of machine learning applications in wildfire science and management. Environ. Rev. **28**(4), 478–505 (2020). https://doi.org/10.1139/er-2020-0019. https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019
14. Kam, H.T.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition (1995)
15. Pourtaghi, Z.S., Pourghasemi, H.R., Aretano, R., Semeraro, T.: Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. Ecol. Ind. **64**, 72–84 (2016). https://doi.org/10.1016/j.ecolind.2015.12.030. https://www.sciencedirect.com/science/article/pii/S1470160X15007359
16. Rodrigues, M., Jiménez, A., de la Riva, J.: Analysis of recent spatial-temporal evolution of human driving factors of wildfires in Spain. Nat. Hazards **84**(3), 2049–2070 (2016). https://doi.org/10.1007/s11069-016-2533-4
17. Safavian, S., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991). https://doi.org/10.1109/21.97458. http://ieeexplore.ieee.org/document/97458/
18. Vilar, L., Camia, A., San-Miguel-Ayanz, J., Martín, M.P.: Modeling temporal changes in human-caused wildfires in mediterranean europe based on land use-land cover interfaces. Forest Ecol. Manag. **378**, 68–78 (2016). https://doi.org/10.1016/j.foreco.2016.07.020. https://www.sciencedirect.com/science/article/pii/S0378112716303760
19. Šotić, A., Rajić, R.: The Review of the Definition of Risk, vol. 3, no. 3 (2015)