Wildfire Susceptibility Mapping
Case Study of the French Pyrénées Region
CS 433: Machine Learning
Project 2

Omar El Malki (310545), Hamza Hassoune (314506), Nael Ouerghemi (310435)

***Abstract*—Wildfires are a major environmental and economic threat in many regions of the world, and effective wildfire susceptibility mapping can help to mitigate their impact. In this study, we propose a machine learning-based approach for wildfire susceptibility mapping in the French regions of Hautes-Pyrénées and Pyrénées Atlantiques. Our approach involves the use of satellite imagery, geographical data to train and evaluate machine learning models for predicting the likelihood of wildfire occurrence in these regions. We were able to successfully improve the current susceptibility map's resolution and provide more explainability to support decision making of authorities.**

I Introduction

Wildfires can have significant impacts on local communities and the environment, including loss of life, damage to infrastructure, natural resources, and air pollution. Effective wildfire susceptibility mapping can help to mitigate these impacts by providing information about the likelihood of wildfire occurrence in different regions, which can inform prevention and management efforts. The objective of the study is to improve the current susceptibility maps that are at the municipality level, estimate correctly wildfire probabilities using supervised models and provide explainability to support decision making.

II Data Acquisition and Initial Analysis

II-A Area of Interest

Located in the southern part of France, the Hautes-Pyrénées and Pyrénées Atlantiques regions border Spain and are home to a mountainous landscape, including the Pyrenees mountain range. With a total area of 4,464km^2, Hautes-Pyrénées is slightly smaller than Pyrénées-Atlantiques, which spans 7,645 km^2. These regions are known for their complex topography, with slopes reaching steep inclines in many areas. The diverse vegetation in these regions is depicted in Figure 1, showing the dense vegetation life of the surface area. The figure also shows artificial surfaces and burnt areas present in the region.

The dataset provided contains fire reports in the region between 2001 and 2022. As seen in Figure 2, there are more reports for recent years. This can be explained by multiple reasons : the data collection process was more comprehensive in the last few years; improved book-keeping of the reports; increasing number of wildfires year after year. This last possibility seems unlikely in our analysis after comparing with reports [1] and [2] in other regions that do not follow this trend.

Figure 3 displays the pixels for which there was a fire report. These reports will be our reference as labels in our Machine Learning Methodology (section III).

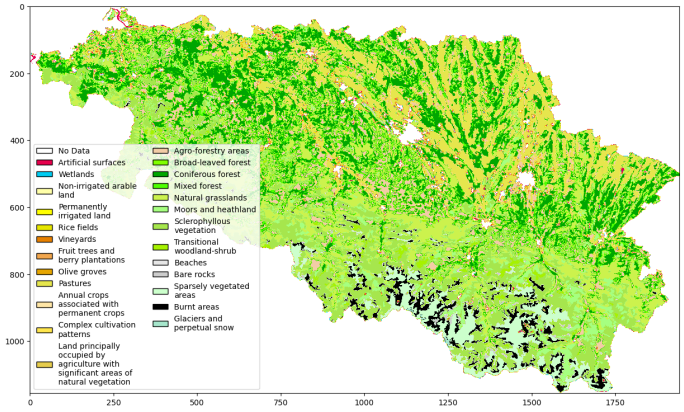

Fig. 1: Area of Interest - Hautes-Pyrénées and Pyrénées Atlantiques in 2018.

This figure contains the vegetation type, artificial surfaces and burnt areas of the region.

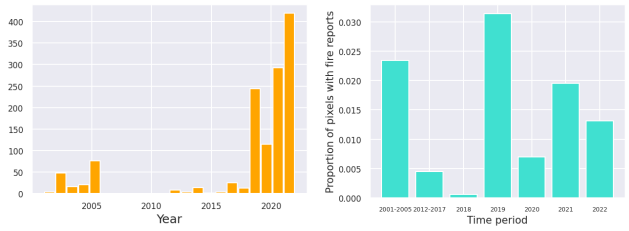

Fig. 2: Number of fires per year in the Area (left) and Pixels with fire reports proportion distribution (right)

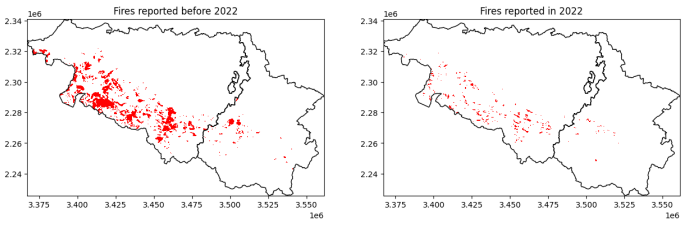

Fig. 3: Spatial Fire Distribution for training and testing periods.

The axes of the graphs represent the coordinates in the ESPG:3035 system.

II-B Preprocessing and Feature engineering

Data for twenty-two different geographical features have been collected from publicly available repositories [3] [4] [5] [6] [7] [8]. Further details on the data collection and engineering are provided in Appendix A and B. The code

to reproduce the analyses is available on our GitHub repository.

After merging and reprojecting the data from the different sources, the original land cover data have been aggregated in 19 classes from the original 37, combining together all the urban areas and water lands together. Such areas are not prone to fires and have been therefore excluded from the model validity domain.

The original spatial data, available as raster and vector files, have been rearranged into tabular data, with each pixel of the raster corresponding to one row, and the different raster values forming the columns. Having raster data for different years, we decided to only keep the values corresponding to the most recent fire year, if a fire has ever been reported in the pixel, in the case where no fire was ever reported, we simply take the most recent values in the study period.

In this study, we performed feature derivation on our geographical data by calculating three derived variables: the distance to the nearest artificial surface, the most frequent raster value in the square of the 3x3 neighbours of a pixel and the percentage of forest and scrub, also in the square of the 3x3 neighbours of a cell. These features were obtained using the QGIS pre-processing tool [9].

As part of our feature engineering process, we aggregated our data to calculate the total number of fires that occurred for a given pixel in the past. Since our tabular dataset was constructed from only one year for each pixel, this step allows us to capture the repeatability of fires.

Table I reports on the features included in the final dataset.

Feature Name	Variable Type	Range
Slope	Continuous	[0, 255]
DEM	Continuous	[-23.5, 3243]
Protected area	Categorical	0 or 1
Aspect	Continuous	[0, 255]
Fires value for different years	Categorical	0 or 1
Fire presence	Categorical	0 or 1
Land Cover	Categorical	19 classes
Urban distance	Continuous	[0, 202]
Amplitude	Continuous	[-3277, 3000]
Min PPI* value	Continuous	[-3277, 2668]
Max PPI value	Continuous	[-3277, 3000]
Right slope	Continuous	[-3277, 1491]
Left slope	Continuous	[-3277, 1237]
Emissivity	Continuous	[0, 250]
Land Surface Temperature	Continuous	[0,1450]
Vegetation Index (VI)	Continuous	[0, 5169]
Normalized Difference (VI)	Continuous	[-1882, 9917]
Enhanced VI	Continuous	[-937, 6854]
is_fire	Categorical	0 or 1
fire_count	Discrete	[0, 5]

TABLE I: Features and their characteristics.
*Plant Phenology Index

III Methodology

Analyses were performed using Random Forest and Light gradient-boosting machine (LightGBM) that we benchmarked with reference to the prediction of forest fire susceptibility. The major focus in the training of the models was the application of a spatial cross validation procedure to reduce the level of (auto)correlation within the input data. The trained models are then calibrated (III-B), as the focus is not the classification per se, but rather assessing correctly the probability of having fires, as this is the relevant information for decision-makers to design their prevention policies. To this aim, interpretability of the results is important which is why we provided an interpretability model using SHAP to identify the most contribuiting features.

Random Forest

Random forests is an ensemble learning method for classification, regression and other tasks, that are based on decision trees. The number of decision trees and the maximum depth of the trees to be used in the model have been selected during the model training with cross-validation. In general, The number of estimators should be large enough so that every input gets predicted a few times. The maximum depth shouldn't be too low so that the model doesn't underfit the data and shouldn't be too high so that the model doesn't overfit.

LightGBM

LightGBM is a gradient boosting machine learning algorithm that is used for classification and regression tasks. It is a variant of the gradient boosting algorithm and is specifically designed to be more efficient and faster than traditional gradient boosting algorithms like Random Forests.

In training, the number of decision trees, the maximum depth of the trees and the minimum number of leaves have been selected with the same procedure used in the case of Random Forest.

III-A Model Validation

1) Cross Validation

In machine learning, it is common practice to divide a dataset into training, validation, and testing sets to assess the performance of a model. However, when analyzing a spatial environmental phenomenon, randomly selecting the points to be included in the different set can lead to the inclusion of spatially autocorrelated samples in a same set, as observations that are close to each other have similar characteristics. This can result in an overestimation of model performances.

To cope with this issue, we designed a spatial k-fold cross validation method [10]. First, the spatial data is divided into blocks of 129 x 101 pixels covering the region of interest. Then, the dataset is divided into k-folds ensuring that all data from a block are used into a same fold; note that this is analogous to a grouped k-fold, where groups are represented by the spatial blocks. Finally, the model is trained on k-1 folds and validated on the last fold. The process is repeated for every fold as the validation fold and the score is the average of all the scores.

We used Grid Search with this cross validation technique to select the best hyperparameters for every model. See Appendix E for more details on the hyperparameter values considered.

2) Train-Test Split

The goal of this susceptibility model is to be used to predict wildfires on a yearly basis. Thus, the only indicator of performance is its performance on the year 2022. Hence, we

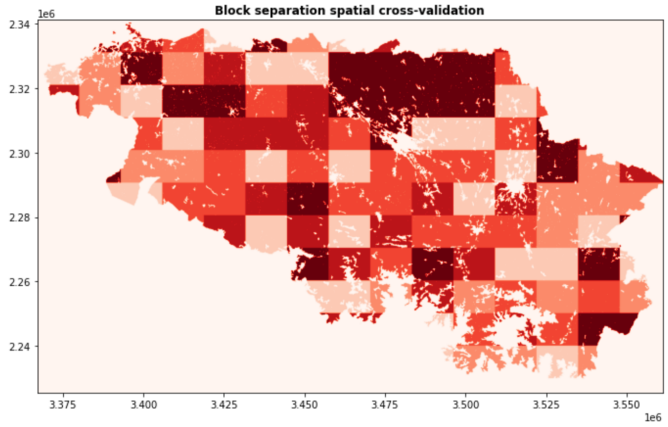

Fig. 4: Spatial Cross validation (5-fold)

This figure displays in different colors the different folds used for spatial cross-validation. The color of each pixel indicates its fold.

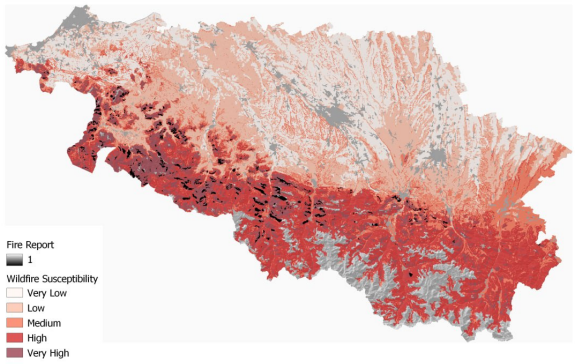

Fig. 5: Wildfire Susceptibility Map - Random Forest

This figure displays the wildfire susceptibility in probabilities between 0 and 1 in different classes as seen in the legend. Black pixels indicate pixels for which there was a fire report in 2022.

used data up to 2021 for training and validation, and the last year being 2022 of the dataset as testing set.

III-B Calibration

Probability calibration [11] is the process of adjusting the predicted probabilities output by a machine learning model such that they accurately reflect the true underlying probabilities of the classes. In other words, it involves adjusting the model's predictions so that they are more closely aligned with the actual outcomes of the events being predicted.

We experimented with the isotonic regression probability calibration method, which involves fitting a step function to the predicted probabilities, such that the probabilities are monotonically increasing.

IV Results

See results in Table II.

IV-A Performance Interpretation

1) Metric Interpretation and Choice

In our problem, the goal is to provide a wildfire susceptibility map of the region. Thus, we want to provide a high susceptibility for pixels prone to wildfires and a lower one for pixels less prone to wildfires.

Since we opted for a classification task predicting if each pixel actually has a fire report in the testing set year range, a very high Precision is a not necessarily a good indicator of success. In fact, even though a lot of pixels may be very susceptible to a wildfire, there is no guarantee that they actually happen during the temporal range of the prediction. The authorities of the region should still be aware of this result. False positives are normal and required by the constraint of the problem. However, Precision is still useful to help the model be less likely to mistakenly identify low-risk areas as high-risk, which can help to avoid unnecessary resources being spent on prevention and management efforts in those areas.

On the other hand, Recall is important since the authorities should be able to properly allocate resources to all areas susceptible to wildfires. False negatives should be limited.

However, we should pay attention to the susceptibility map to make sure our model is not too generous when predicting fires and stays coherent.

Thus, to take into account both nuances, we chose to optimize on the F1-score. We display F1-Score, LogLoss, Recall and Precision for a complete overview of our approaches.

2) Grid-Search, Hyperparameter tuning Interpretation

The best performing hyperparameters for Random Forest were :

- Maximum Tree Depth : 2

From a tree depth of 3, the model starts overfitting.

- Number of Trees (estimators) : 200

The performance kept improving until the 200 estimators and converged. Our computational resources did not allow us to explore more than 1000 estimators.

The best performing hyperparameters for LightGBM were :

- Maximum Tree Depth : 2

Similarly, larger max tree depths lead to overfitting, and lower depths had worse performance.

- Number of Trees (estimators) : 12

The model heavily overfit after 12 trees.

- Number of Leaves (estimators) : 4

There was no performance change after tuning the number of leaves with the values we selected.

IV-B Explainability

The calibrated model feature importance will be calculated using SHAP library. **SHAP** (SHapley Additive exPlanations) [12] is a method for explaining the output of machine learning models. It provides an explanation of the predictions made

Model	F1-score		LogLoss	Precision	Recall
	Train	Test	-	-	-
Baseline (0)	0.0	0.0	-	0.0	0.0
Baseline (1)	0.11	0.02	-	0.013	1.0
Best Random Forest Model	0.39	0.36	0.09	0.26	0.61
Best LightGBM Model	0.49	0.31	0.10	0.20	0.73

TABLE II: Results

These results are for the best hyperparameters as described in IV-A2. Baseline 0 and 1 correspond to only predicting 0s and 1s respectively.

by a model by attributing the prediction to the features that contributed to it, and quantifying the contribution of each feature. SHAP works by approximating the Shapley values, a concept from cooperative game theory, for each feature. The Shapley value measures the average marginal contribution of a feature to the model's output, taking into account the interactions between the features.

As we can see in the SHAP summary plot 6 of our Random Forest model, the slope is the feature with the most important contribution to wildfires. This is explained by the fact that the steepness of a slope has an incremental effect on fire behaviour and its speed. Fire intensity and rate of spread normally reduces when fires are burning downslope. This result converges with the susceptibility analysis made by the administrative region of the area of interest as it describes the slope as a "factor, which concerns all municipalities with a relief accident, [that] aggravates the hazard since it promotes the spread of a fire but also because it reduces the effectiveness of the means of control."[13]

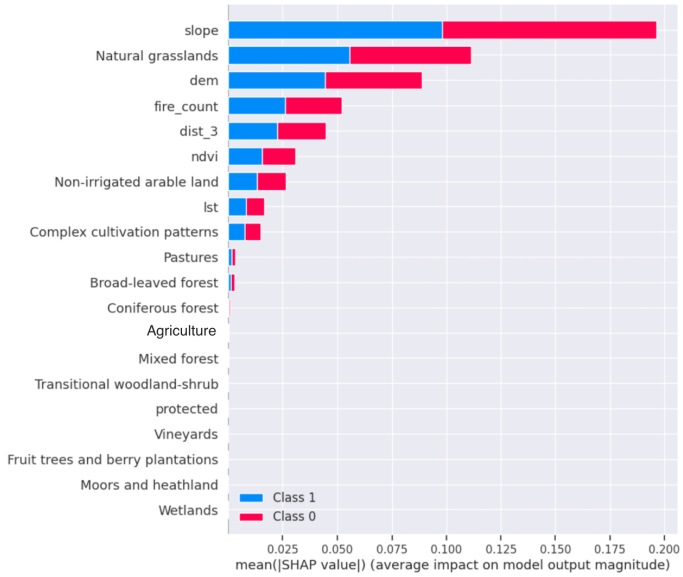

Fig. 6: Shap Summary Plot - Random Forest

V Ablation study

Even though the primary goal of the study, which was to obtain a susceptibility map with higher precision than the current one, was executed successfully, we notice that the results could be improved.
This could come from the data collection and choice of features. Thus, we performed an ablation study removing multiple features as an attempt to identify the origin of this issue.

See in Tables III and IV the results of the ablation study.

Model	F1-score	LogLoss
Full model	0.36	0.09
Model without Land Cover	0.31	0.10
Model without Fire Count	0.31	0.09
Model without Distance to Artificial Surfaces	0.36	0.09

TABLE III: Ablation study for the best Random Forest model

Model	F1-score	LogLoss
Full model	0.31	0.10
Model without Land Cover	0.30	0.12
Model without Fire Count	0.01	0.09
Model without Distance to Artificial Surfaces	0.28	0.13

TABLE IV: Ablation study for the Best LightGBM model

VI Conclusion

Forest fire prevention is and will remain as a very important challenge in Europe and the entire world, as no country seems exempt. But as data becomes increasingly more available, thanks to plateforms such as Copernicus [5] or the Goddard Space Flight Center [7], this will be a long-term process that will eventually provide accurate results to assist fire managers. Our project work can be summarized into the following points :

- Data processing and engineering to obtain features that would be significative with regards to the forest fire probability prediction.
- Developing three supervised models : Random forest, LightGBM and Logistic regression, to provide explainability to support decision making for the regions Pyrénées Atlantique and Hautes-Pyrénées.
- Providing an improved susceptibility map with regards to the current ones at the municipality level. - Providing an explainability model using SHAP to proide better insight and support decision-making of authorities.

Further work should include improving the ablation study with more features as well as more importantly review the data collection process to cope with the target data imbalance.

References

[1] Haifeng Lin Shuwen Xu Yanyan Sun, Fuquan Zhang. A forest fire susceptibility modeling approach based on light gradient boosting machine algorithm. *mdpi.*, 2022.

[2] Guido Biondi Silvia Degli Esposti Andrea Trucchia Paolo Fiorucci Marj Tonini, Mirko D'Andrea. A machine learning-based approach for wildfire susceptibility mapping. the case study of the liguria region in italy. *Geosciences.*, 2020.

[3] Gouvernement.fr. Contours des départements français issus d'openstreetmap. https://www.data.gouv.fr/fr/datasets/contours-des-departements-francais-issus-d-openstreetmap/, 2018.

[4] Copernicus. Eu digital elevation models. https://land.copernicus.eu/imagery-in-situ/eu-dem, 2016.

[5] Copernicus. Total burnt areas. https://effis.jrc.ec.europa.eu/applications/data-and-services, 2022.

[6] Copernicus. Eu land cover. https://land.copernicus.eu/pan-european/corine-land-cover, 2021.

[7] NASA. Land surface temperature emissivity. https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/science-domain/land-surface-temperature-and-emissivity/, 2007.

[8] FR Government. Finpn - données du programme 'espaces protégés'. https://www.data.gouv.fr/fr/datasets/inpn-donnees-du-programme-espaces-proteges/, 2022.

[9] QGIS Development Team. Qgis software. https://www.qgis.org/fr/site/, 2022.

[10] Simone Ciuti Mark S. Boyce Jane Elith Gurutzeta Guillera-Arroita Severin Hauenstein José J. Lahoz-Monfort Boris Schröder Wilfried Thuiller David I. Warton Brendan A. Wintle Florian Hartig Carsten F. Dormann David R. Roberts, Volker Bahn. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography.*, 2016.

[11] Probability calibration. https://scikit-learn.org/stable/modules/calibration.html.

[12] Shapley additive explanations. https://github.com/slundberg/shap.

[13] Région Pyrénées-Atlantiques. Forest protection plan of wildfire. https://www.pyrenees-atlantiques.gouv.fr/Actions-de-l-Etat/Agriculture-foret-et-developpement-rural/Forets/Plan-departemental-de-protection-des-forets-contre-les-incendies-PDPFCI/Plan-departemental-de-protection-des-forets-contre-les-incendies-PDPFCI, 2020.

[14] Jannes Muenchow Robin Lovelace, Jakub Nowosad. *Geocomputation with R*. CRC Press, 2022.

[15] Jakub Nowosad Robin Lovelace Michael Dorman, Anita Graser. Geocomputation with python. https://geocompr.github.io/py/, 2022.

[16] Even Rouault Frank Warmerdam et al. Gdal documentation. https://gdal.org/api/python_bindings.html, 2022.

[17] Even Rouault Frank Warmerdam et al. Probability calibration. https://scikit-learn.org/stable/modules/calibration.html, 2022.

[18] Shap tree explainer. https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html.

[19] Sklearn python library. https://scikit-learn.org/stable/.

[20] Rasterio python library. https://rasterio.readthedocs.io/en/latest/.

Appendix

A Data collection

The data collection was one of the important challenges that we faced, especially for the vegetation. Most of the files were available on Capernicus and Nasa websites [put references]. However the vegetation data was only reachable through the Wekeo API [put reference]. Hence we had to get used to it, doing the data collection, filtering and pre-processing (as described here [put link]). We provided a python notebook to reproduce this part.

B Data pre-processing

The pre-processing of the data was mostly divided in 3 distinct parts : Changing the resolution of the rasters, changing the coordinate system to a shared one, and cropping the data to our wanted region.

For all the raster files, we chose to select as a resolution the highest resolution amond the files, which was 100x100. We then applied this to all the other layers. Furthermore, regarding the coordinate system of the rasters we chose ESPG:3035, which corresponds well to our region (EUROPE LAEA). Finally, the cropping was done using the french departments regions in vector format.

Most of the data pre-processing was done using QGIS, which is an open-source geographic information system application that supports viewing, editing, printing, and analysis of geospatial data.

C Features

The used features can be divided in 4 categories : Topographic factors, Vegetation factors, Climatic factors and human activity factors. We will here briefly describe each of them.

Land topography is considered to be one of the most important fire factors, hence we tried to include as much of them as possible. We used the slope, digital elevation models, aspect, hillshade and land covers.

The vegetation type is also an influencing factor, as it is critical in the spreading of the fire and can be a fuel to it. As for the chosen features, we used : The amplitude, the minimum and the maximum plan phenology index.

Furthermore, the climatic were considered not only because they are obviously an important factor of forest fires, but also because they have a great influence on the combustion conditions of the vegetation. Finally, we also considered human activity factors, as we have a big influence on fires (campfires left unattended, the burning of debris, dropped cigarettes ends, or even intentional acts). Hence we added for each pixel the distance to urban areas.

D Hyperparameters tuning

We briefly discuss the hyperparameters that we considered for our models in this section. For Random Forest, the following parameters were considered alongside with the values : the number of boosted trees to fit $\{100, 200, 300\}$, the maximum tree depth for base learners $\{2, 3, 4, 5, 7, 10, 12, 15, 20\}$. For LightGBM, we tuned the same parameters but with respective values $\{1, 3, 7, 12, 25, 50, 100, 200, 300\}$ and $\{2, 4, 7, 12, 20\}$ as well as the number of leaves with values $\{4, 16, 31\}$

E Current advances in wildfire susceptibility

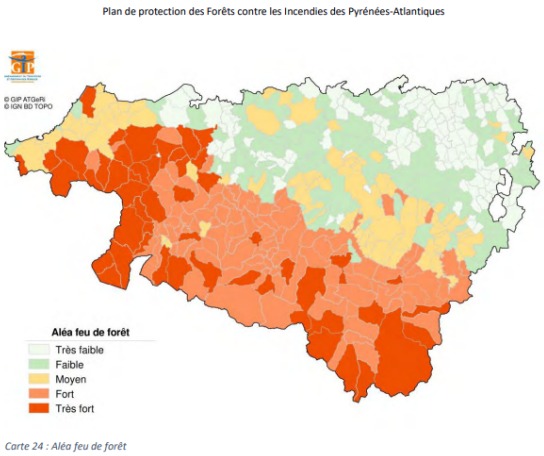

Fig. 7: Wildfire susceptibility map - Forest Protection Plan of Wildfire Pyrénées-Atlantiques

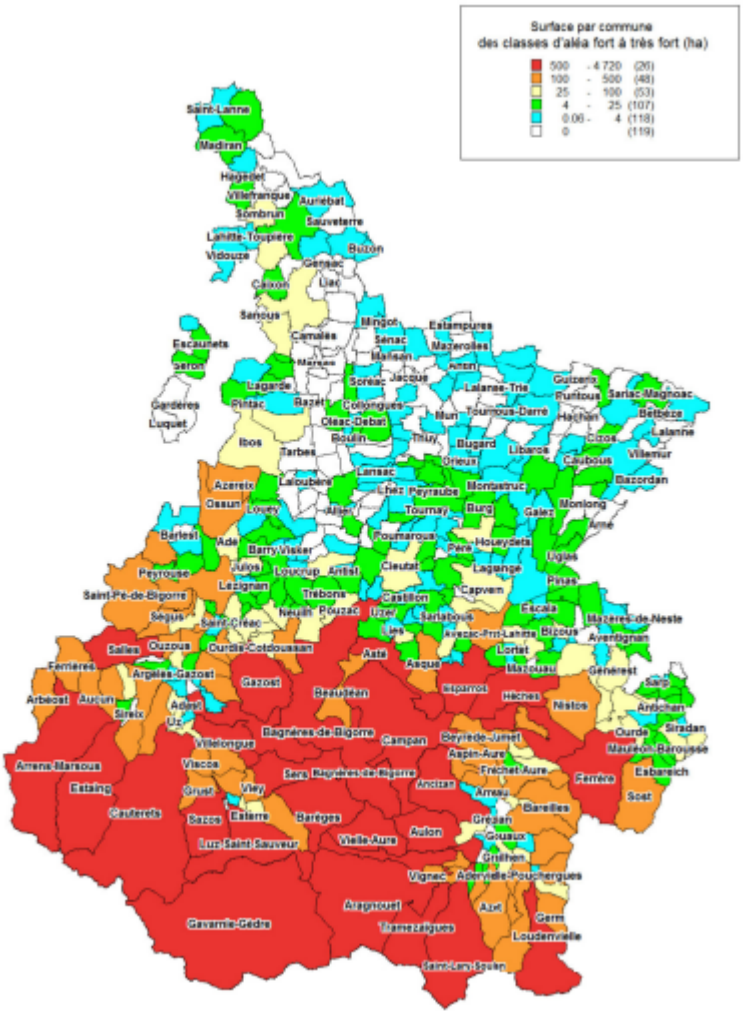

Fig. 8: Wildfire susceptibility map - Forest Protection Plan of Wildfire Hautes-Pyrénées