HACKATHON GROUP 2

FINAL REPORT

The main drivers to Predict Wildfire's Rate of Spread from the PT-FireSprd Database

Diogo Gomes, nº 26843
Rafael Oliveira, nº 26606
Xavier Loreto, nº 28648

INSTITUTO SUPERIOR DE AGRONOMIA
Universidade de Lisboa
Hinc Patriam Sustinet

Index

1. Introduction

Wildfires have caused an unprecedented number of fatalities and injuries, along with the loss of countless homes and critical infrastructure. Yet, despite their growing impact, our understanding of these complex and dynamic events remains limited. Gaining deeper insight into fire behavior, particularly the processes that drive ignition and propagation in varied landscapes, is essential. Such knowledge strengthens the entire fire management cycle by enabling more informed, timely, and effective decision-making process.

To achieve these substantial improvements in fire management, the FIRE-HACK project aims to take action in the following key steps: i) to have a comprehensive and high-quality dataset of fire behavior, ii) to robustly characterize the relationships between fire behavior drivers and their impacts, and iii) to effectively disseminate this knowledge to the relevant decision-makers. These findings can have positive impacts in every part of the fire management cycle, can help to define policies to protect significant areas and reduce risk, to delineate adequate alert and safe evacuation plans, to predict active fire behavior in order to support the appropriate deployment of resources for wildfire combat, and to find windows of opportunity for the wildfire's combat, and even to dictate post-fire recovery actions.

This report focuses in a section of the FIRE-HACK project, namely: from the already built PT-FireSprd dataset, which contains the reconstruction of the spread of 155 large wildfires that occurred in Portugal between 2015 and 2025, and a detailed set of fire behavior descriptors, update it with environmental information and to define and characterize the relationships between these fire behavior drivers and a fire behavior variable, the Fire Rate of Spread. This work will culminate on the creation of an interface that can be used to predict the future Fire Rate of Spread of an wildfire (current or hypothetical) from the weather conditions and geographic information, and one of two built-from-the-ground-up fire behavior machine learning models, a more complex and accurate XGBoost Model and a simpler and easy-to-use Linear Regression Model. We have outlined these four phases as depicted in Figure 1. Each step was overseen by a designated individual, as seen in Figure 2. Rafael Oliveira serves as both the Project Coordinator, Domain Expert and Data Engineer, Diogo Gomes takes on the roles of Rapporteur and Model Developer and Xavier Loreto is assigned as the Data Scientist, aided by Rafael Oliveira.

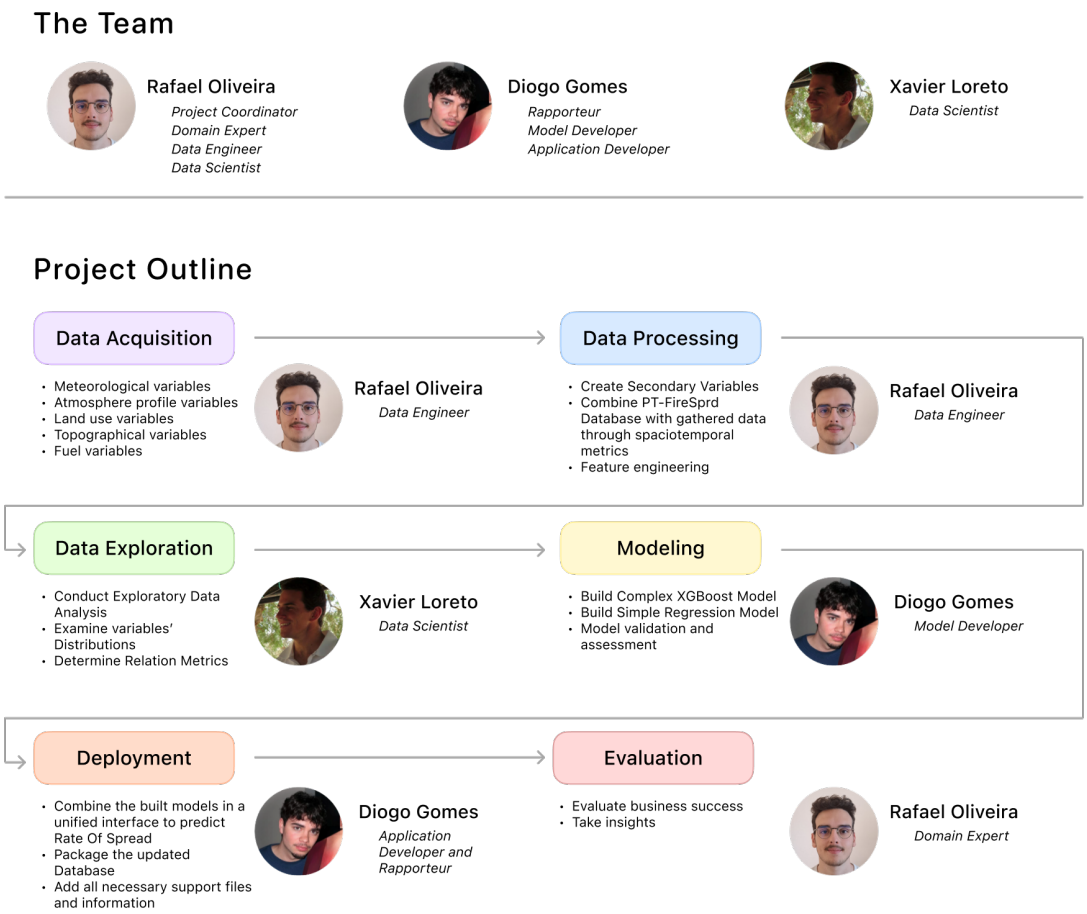


Figure 1 - Team and project outline diagrams.

Our research will be conducted using Python, Excel and QGIS on Windows 11 and macOS 26 operating systems.

2. State of the art

The interaction between meteorology, topography, and fuels is decisive for the behavior of a forest fire, as it acts directly on the three elements of the fire triangle, which represents the conditions essential for ignition and the spread of fire: heat, fuel, and oxygen. The variation of these elements, controlled by different environmental factors, determines the intensity, speed of spread, and predictability of the fire.

Topography, through steep slopes, intensifies heat transfer by radiation and convection, raising the temperature of fuels upstream and accelerating the spread of fire, while valleys and ridges modify wind patterns, increasing oxygen supply and turbulence. The characteristics of fuels, such as continuity, porosity, and load, determine the rate of energy released and the ability to sustain extensive fronts. Meteorology integrates and amplifies these processes by regulating the flammability of fuels and influencing the spread and intensity of fire through air temperature, relative humidity, wind, and atmospheric instability. Recent literature demonstrates that the weather influences wildfire behavior not only at the surface, but also through the vertical structure of the atmosphere, which plays a decisive role in accelerating the speed of propagation. Under conditions of atmospheric instability, warm air rises more easily, favoring the development of intense convective columns. The presence of steep

temperature gradients, low relative humidity values at altitude, or high levels of convective available energy (CAPE) contribute to sudden variations in fire intensity and ventilation of the propagation front.

In high-intensity fires, the energy released can affect his own environment, creating columns of smoke with strong convection, which can directly influence atmospheric flow. This phenomenon, known as pyroconvection, can alter the direction and intensity of the wind near the ground, massively elevate incandescent particles that generate secondary ignitions, and cause sudden accelerations in the speed of propagation. When the convective column begins to actively modify the local wind field, fire-atmosphere coupling occurs, a process documented in several large fires and associated with very high fireline Rate of Spread values.

The study of fire behavior has evolved significantly in recent decades, but the ability to accurately predict the spread of a fire based on its environmental contex remains limited. Despite recent advances, many operational systems currently in use continue to be based on classical empirical models, the most important being the *Rothermel* (1972) propagation model. This model forms the basis of several widely used simulators, such as *BEHAVE*, *BehavePlus*, *FARSITE*, *FlamMap*, and *Prometheus*. These systems describe the speed of propagation based on empirical relationships between wind, slope, and fine fuel characteristics, assuming relatively stable fronts and homogeneous fuels. However, these simplifications make them less adequate for representing fast, intense fires or those influenced by convective and other complex phenomena, which have been more common in recent years due to global warming and climate change.

In order to overcome these limitations, coupled atmosphere-fire physical models such as *FIRETEC* or *WRF-Fire* have emerged, which explicitly simulate turbulence, heat transport, and the interaction between fire and the atmosphere. These models more realistically represent processes such as convection, wind shear, variations in the vertical thermal profile, and extreme fire behavior. However, due to their high computational cost, they are not yet suitable for real-time operational forecasting. Thus, in the context of fire management, it is still necessary to resort to empirical or hybrid models, complemented by high-resolution meteorological and geographical information.

At the same time, the increased availability of high-resolution atmospheric and geospatial data has enabled the development of models based on machine learning techniques. Several studies show that methods such as Random Forest, Gradient Boosting, XGBoost, or neural networks can overcome some limitations of classical models by capturing complex and nonlinear relationships between meteorological, topographical, and fuel variables. However, their performance depends heavily on the quality and consistency of their training dataset, making consistent structured databases with high temporal and spatial resolution, that characterize completely the environmental context of the fire, such as PT-FireSprd, is essential.

In summary, although there is consolidated knowledge about the effect of wind, fuel, and topography, reliable prediction of propagation speed still requires work, namely in the joint integration of vertical atmospheric variables, convective processes, fuel characteristics, and topographic context. The approach followed in this project, enriching PT-FireSprd with detailed meteorological and geographical variables and combining them with machine learning models to define the main fire behavior drivers, is directly in line with this new generation of studies, addressing the limitations of traditional models and creating conditions for more realistic and operational predictions of fire spread rates while working to uncover the complex relations that drive the rate of spread of a wildfire.

3. Methodology

3.1 Data Acquisition

3.1.1 The PT-FireSprd Database

The Portuguese Large Wildfire Spread database (PT-FireSprd) is a spatiotemporal database that includes reconstructed progressions of more than 155 large wildfires that occurred in continental Portugal between 2015 and 2025, with area greater than 100 ha. This dataset also contains relevant fire behavior information such as Rate of Spread, Spread Direction, Fire Growth Rate and Fireline Intensity.

This database splits the data into 3 detail levels. The L1 level represents the temporal and spacial spread of each wildfire, not only defining ignition and active flame zones and fire progression polygons in time and space, but also establishing the relations between these polygons. Each progressions' limits were compiled and manually digitalized taking into account a multitude of data sources including airborne images taken from planes and helicopters amidst the combat, satellite imagery in the visible and the infra-red spectrum from the MODIS Terra and Aqua, VIIRS NPP and NOAA-20, Landsat 8/9, Sentinel 2 (S2) and 3, MSG-SEVIRI and PROBA_V satellites, images and videos captures by fire operatives and their georreferenced data, official fire data such as burned area, ignitions and fire logs, and also the Reports on the large wildfires of 2017, which led to the highest at the time recorded values of ROS and FGR. Where there was insufficient data to determine when an area burned, the spread polygon was flagged as uncertain.

Since this process was hand-made, despite using various data sources, and since it is very difficult to identify when an area has burned completely, in addition to the variations in data resolution and availability in the study's time frame, there will certainly be inaccuracies and inconsistencies that will affect the result of this report. This effect will unfortunately be stronger in the oldest and smallest reconstructed fires where the methodologies and available technology was not the most adequate. Other sources of error are stated in Benali et al. (2023).

The L2 levels adds information on fire behavior descriptors. The Spread Direction is defined by the longest of the shortest lines between the vertices of the fire front at the beginning of the polygon and the end of the same polygon. The Rate Of Spread is calculated by the ratio of this direction by the time interval at which the progression pertains to. The Fireline Intensity of a polygon was determined from Byram’s equation taking into account Rate Of Spread, fuel heat of combustion and fuel consumed in the active flaming front. All these metrics will have two variations each, ros_p, spdir_p and int_p calculated between two consecutive polygons, and ros_i, spdir_i and int_i calculated between the previous active flaming front and the fire progression. This database level also contains data on the Fire Growth Rate obtained from each polygons area and duration. The firebehavior information was not calculated where the polygon's start and end date was not known.

The L3 level simplifies this behavior data by averaging the attributes for each burning period, we consider that in each burning period the fire run is homogenous, allowing for advanced analysis of the data despite a level of loss of detail by making simplifications to the calculation of the fire behavior metrics.

The researcher Akli Benali has allowed us access to the latest version of the database, that already has the reconstructed progressions of some of the 2025 wildfires. This updated database is not publicly available yet but will soon be published. Since the modeling approach requires a lot of data to generate accurate results, we will be using the L2 level which contains 3367 progressions (we will only use a part of those due to some constrains that will later be explained).

We will complement the PT-FireSprd Database with environmental information such as topography (elevation, aspect, landform), land use, fuel characteristics (fuel load, fuel model, and fire history), and a comprehensive set of meteorological variables (soil moisture, temperature, humidity, wind, pressure, stability indices, and atmospheric profiles).

3.1.2 The Meteorological Data

The meteorological information was originally gathered by the *Copernicus Earth Observation Programme*, that is part of the *European Union's Space Programme*, based on satellite and in situ observations. It is divided into six thematic information services, from which we will use the *Copernicus Climate Change Service (C3S)*, that provides information about the past, present and future climate in Europe and the rest of the World, and the *Copernicus Emergency Management Service (Copernicus EMS)* that provides, among other data, information on the real-time and historical information on forest fires and forest fire regimes, through their *European Forest Fire Information System (EFFIS)*.

From the *Copernicus Climate Change Service (C3S)* and their *Climate Data Store (CDS)*, we will download 3 separate data sets through their API service, CDS API:

- The ERA5 hourly data on single levels from 1940 to present - This dataset provides hourly information throughly describing the state of the atmosphere close to the Earth's surface but with a lower spacial resolution (0.25°).
- The ERA5 Land hourly time-series data from 1950 to present - This dataset is a land-surface enhancement of ERA5, produced at higher spatial resolution (0.1°) and optimized specifically for representing the most important surface processes.
- The ERA5 hourly data on pressure levels from 1940 to present - This dataset provides a detailed 3D representation of the atmosphere, offering fields at standard pressure levels across the entire vertical atmospheric column, at the expense of lower spatial resolution (0.25°).

All of these ERA5 datasets are global reanalyses, which means they consist of historical reconstructions of the weather and climate of the atmosphere, land and ocean conditions, made by merging historical observations from various sources and modern numerical weather prediction models, using data assimilation systems to continuously add observations and correct the model. However, the use of reanalysis products also has important limitations, since they are model-based reconstructions and may not faithfully represent localized or short-lived extreme events. They also are strongly depend on the underlying numerical model, so systematic biases can affect multiple variables and may exhibit artificial temporal inconsistencies due to changes in the observing system.

From the *Copernicus EMS* and their *CEMS Early Warning Data Store*, we will download the Fire danger indices historical data from the *Copernicus Emergency Management Service* dataset, through the same as before API service, CDS API. This dataset provides a global, daily, gridded reconstruction of historical fire-danger conditions based on the ECMWF ERA5 reanalysis. This method combines 3-hourly forecast outputs from different times to create a snapshot of atmospheric conditions around local noon. Being model-based, it, much like the other datasets, inherits biases from the underlying numerical weather prediction system, and the construction of daily "snapshots" from 3-hourly forecasts, may smooth out rapid changes or local extremes, particularly in regions with complex topography or microclimates. Small inconsistencies can also arise where the 3-hourly slices are joined, potentially introducing artificial temporal discontinuities.

All extracted variables are shown in the figure below. Other secondary variables will be calculated later based on these. We will request the same variables in the ERA5 Land and ERA5 on Single Levels datasets so as not to mix variables from these two different models when calculating secondary variables.

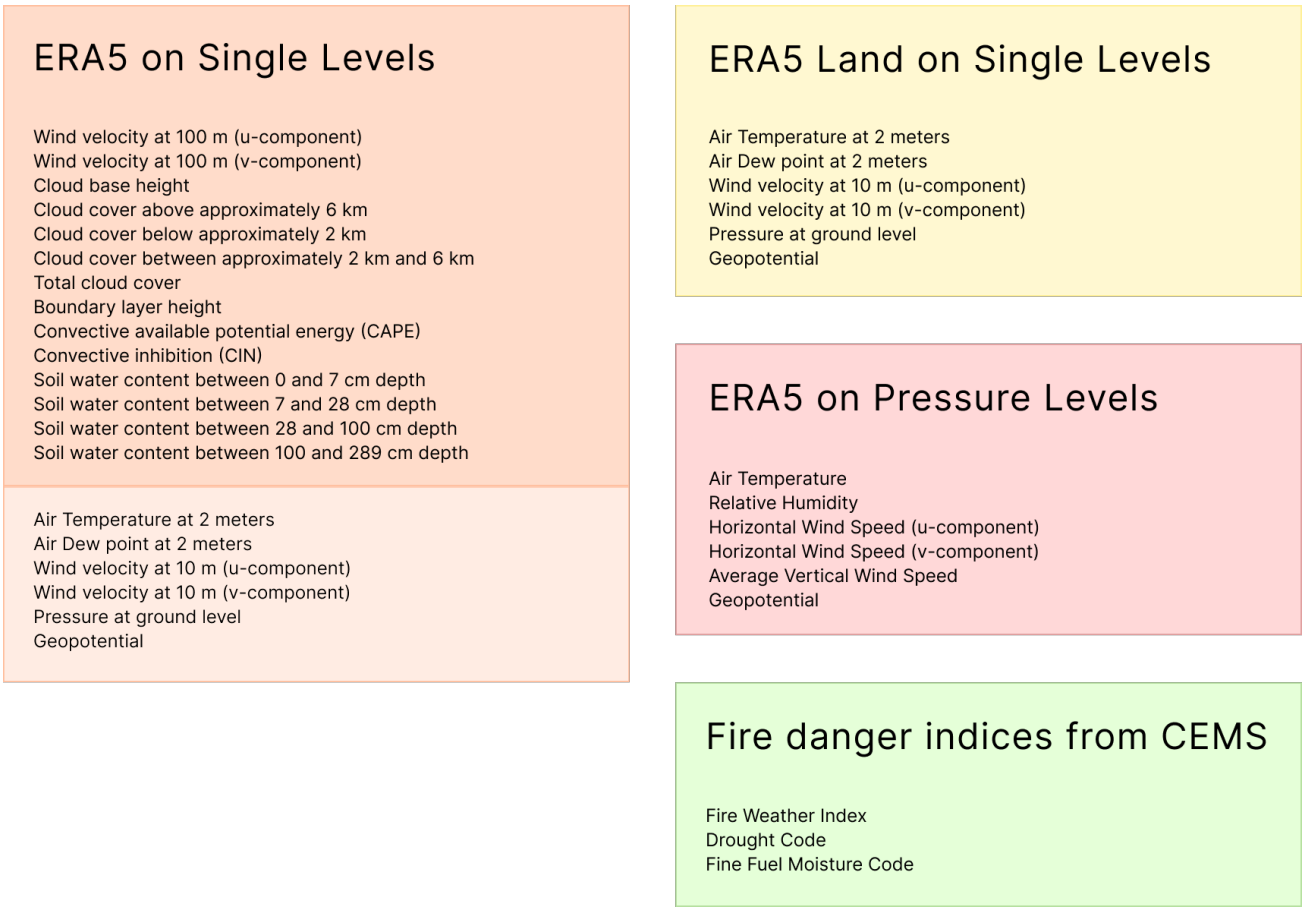


Figure 2 - Variables extracted to complement PT-FireSprd database.

Data was obtained for all 24 hours of each day on which there was an ongoing fire progression. The result was an assortment of NetCDFs that, for continental Portugal, contained the values of all variables for each combination of time, latitude and longitude requested. The dataset ERA5 on Pressure Levels had values, not only for each combination of time, latitude and longitude, but also pressure level. It was requested data for the 950hPa, 850hPa, 700hPa, 500hPa and 300hPa pressure levels.

3.1.3 The Geographic Data

Researcher Akli Benali also granted us access to the *GISdata* repository, which contains rasters on various characteristics of continental Portugal's landscape, such as:

- Average slope orientation (50m),
- Average elevation (50m),
- Dominant landform morphology, based on combining topographic position (ridge/slope/valley) derived from the pixel's relative topographical position with solar exposure (warm/neutral/cool) (90m),
- Predominant fuel model from the Fernandes et al. (2009) fuel models characterization (100m),
- Annual rasters on “years since last fire” (50m).

From the repository, we will only use the data mentioned above.

3.1.4 Other Data

Finally, we will also take into account the land use of each polygon as a determining factor in the speed of fire propagation, for which we will use the *Land Use Map* created by the Portuguese Government's *Direção-Geral do Território*. This is a vector map of mainland Portugal subdivided into land use polygons. These products are produced at 5-year intervals and, since 2018, have a consistent nomenclature, so we will only use the 2018 and 2023 maps. They can be freely downloaded from the Portuguese data catalog (dados.gov) website.

3.2 Data Preparation

3.2.1 Adding all geographical variables to the database

At this stage, we will add the environmental variables to the PT-FireSprd database to describe in detail each fire progression in its spatiotemporal context. Only the meteorological data is temporally dynamic (some data is yearly dynamic but there is no progression thats spans two years).

Most files from the *GISdata* data store were provided in an *ESRI ASCII Raster* format, a text-based grid file that stores a geographic grid where each cell contains a numeric value. They were therefore converted to TIFF raster format, as this is an image-based binary raster format that allows spatial calculations to be performed.

From the slope and elevation orientation rasters, we will calculate the average value in each fire polygon, using all raster cells that are completely within it or, if none are available, all cells that are partially contained within the polygon. This cell selection method will be used in the processing of all raster data. We would also like to point out that the average orientation is necessarily calculated as a circular average.

We will add to the dataset information on the dominant landform and fuel model type using spacial modes, since there are categorical variables.

The fuel model raster, together with the table characterizing the fuel types defined by Fernandes et al. (2009), will also be useful for calculating the average fuel load for each progression polygon. To do this, we will calculate a weighted average by multiplying the average fuel load value for each fuel type by its number of pixels in the polygon and dividing by the total number of pixels.

The “last time the area was burned” is a very important metric for estimating the rate of fire spread, as it is an indicator of the amount of real fuel accumulated in the region. We will calculate the percentage of the polygon area that was burned between 1 and 3 years ago, 3 and 8 years ago, and more than 8 years ago, from the raster “years since last fire”. It is important to note that it would make more sense to calculate the proportion in terms of vegetation area rather than total area, so that each observation would be comparable. To do this, it was necessary to distinguish between areas that have never been burned because they have no vegetation, and areas that have never been burned but could have been. This was done by rasterizing the vegetation classes of the Land Use Map (codes 3, 4 and 5) and comparing them with the “years since last fire” rasters, taking care to use the 2018 Land Use Map for the rasters up to 2020 and the 2023 map thereafter.

To define the main soil land use of each progression we also had to take into account their year, to use the correct Land Use Map. The map's key was simplified to have broader, more useful, categories: No Vegetation, Agricultural Areas, Pastures, Deciduous Forests, Eucalyptus Forests, Invasive Species

Forests (acacia forests), Coniferous Forests, Shrub and Sparse Vegetation. The table with the correspondence between the original and the adapted key is in Annex A.

3.2.2 Complementing the Meteorological datasets

Firstly we consolidated the ERA5 on Single Levels, the ERA5 Land on Single Levels and the Fire Danger Indices datasets into a single NetCDF data store with a 0.1º regular grid and hourly values for all their variables. The ERA5 on Pressure Levels data store was also interpolated to a 0.1º grid.

So, as we had more than 180 million individual values in the data store, to optimize future calculations, we excluded from it all combinations of position (latitude and longitude) and time that did not correspond to any fire progression, resulting in only 1.4% of the total values being retained. This means that almost 99% of the values would be processed, using time and computational power, but not be used in our calculations, which would be very inefficient. This process was separately done in the ERA5 on Pressure Levels dataset.

In the following table we can see which secondary meteorological variables were calculated and added to the Single Level dataset and how:

Wind Speed at 10m	From the u and v components of the wind at 10m, in km/h
Wind Direction at 10m	From the u and v components of the wind at 10m, in km/h
Relative Humidity at 2m	From the air temperature and dew point at 2m, and their water vapour pressures, in % (1)
Vapour Pressure Deficit	From the air temperature and dew point at 2m, and their water vapour pressures, in Pa (1)
Dead Fuel Moisture Content	From the relative humidity and air temperature at 2m using the Anderson et al., 2015 Model for estimating dead fuel moisture content in heathlands, in %
Soil Water Content in the soil's top 1m	From a weighted average of the water content in the 0-7 cm, 7-28 cm and 28-100 cm soil layers, in m^3/m^3
Soil Water Content in the soil's top 3m	From a weighted average of the water content in the 0-7 cm, 7-28 cm, 28-100 cm and 100-289 cm soil layers, in m^3/m^3
Lifted Condensation Level Pressure Level	From the air temperature and dew point, to define the temperature where a lifted air parcel becomes saturated, to then it convert it to LCL pressure by following a dry-adiabatic process from the surface pressure. (Bolton's (1980) analytical approximation) (1)
Lifted Condensation Level Height	Converted from the pressure level to height using the U.S. standard atmosphere, in meters (1)
Hot Dry Windy Index	From the product of the VPD and the Wind Speed at 10m
Haines Index	From the lapse rate and dew point depression between two pressure levels. They are converted into a 1-3 score, based on thresholds, and added together. The thresholds depend on the ground elevation, calculated using the geopotential
Wind Sheer of intensity between 10m and the n Pressure Level	For all n pressure levels, it is calculated as the difference in wind speed between 10m height and the wind speed at the n pressure level, in kh/h km (3)
Wind Sheer of direction between 10m and the n Pressure Level	For all n pressure levels, it is calculated as the difference in wind direction between 10m height and the wind speed at the n pressure level, in º/km (3)
Wind Sheer of intensity between 10m and 100m	For all n pressure levels, it is calculated as the difference in wind speed between 10m and 100m height, in km/h km

Wind Sheer of direction between 10m and 100m	For all n pressure levels, it is calculated as the difference in wind direction between 10m and 100m height, in º/km
Temperature gradient between the a and b pressure levels	For all consecutive a and b pressure levels (including the surface), uses their temperatures and geopotential, to calculate the interval's lapse rate, in ºC/km upwards (3)
Cloud Mixing Layer Gap	From the difference between the LCL height and the Boundary Layer Height, in meters
Level of Free Convection	From the LCL, lifts the parcel moist-adiabatically while comparing its temperature to the environmental temperature at each pressure level. The LFC the first level where the parcel becomes warmer than the environment. (1)
Convective Condensation Level	From the surface dew point, and the temperature profile, defines a mixing ratio, then finds the level in the temperature profile where the air parcel would become saturated. (1)
Equilibrium level	From the temperature and dew point profiles, lifts a parcel moist-adiabatically from the surface pressure and finds the level where its temperature matches the environment, then converts that pressure to height. If CAPE = 0, there is no buoyant energy for the parcel, so EL is no value (-1)
Ventilation Rate	From the product of the estimated Mixing Layer Height (MLH), based on the CAPE and wind speed at 950hPa, with that wind speed.
Lifted Index	From the difference between the temperature at 500 hPa and the temperature of a parcel lifted from the surface pressure to 500 hPa based on the temperature profile.

Table 1 - Meteorological variables calculated and added to the Single Level dataset.

(1) Calculated using the metpy python library
(2) In calculations with surface and pressure level variables, we used the variables from the ERA5 on SL and PL datasets, not ERA5 Land on single levels
(3) If the surface is above the pressure level of 950 hPa, the variable involving this pressure level is set to no value (-999).

In some cases, *Convective Inhibition* and *Cloud Base Height* values can be *None* when there is no clouds or no inhibition, in those cases we assigned a no value (-1). Cin values were also transformed into negative ones, since they were provided as positive values.

There must be caution when using the LCL, CCL, EL and LI values since their calculation requires the integration of the full temperature and dewpoint atmospheric profile based on the data of just 6 levels (1 surface level and 5 atmospheric pressure levels), which can bias the results.

In the ERA5 on *Pressure Levels* data store we also calculated the wind speed and direction for each pressure level. The vertical wind speed was provided in pascals per second, so it was converted to kilometers per hour by calculating the air density from its temperature. We calculated the dew point at each pressure level, ensuring that we used the saturation point of water and ice where relevant.

Lastly, we converted all pressure-level variables in the ERA5 on *Pressure Levels* dataset into separate 3D variables per level (time, latitude, longitude) and merged them with the *Single Level* dataset creating a unified, complete meteorological dataset.

3.2.3 Calculating spatiotemporal meteorological variables

Now we will add these meteorological variables to the PT-FireSprd database. Since the meteorological dataset is a combination of latitudes and longitudes, it is a group of grid points, therefore, to conduct spacial operations, we need to transform this dataset into polygons.

Then, we can compute spatial averages for each hour of each progression to determine its temperature, wind speed, etc., at that moment. From these values, we can calculate temporal averages to assign each progression its mean meteorological attributes over the time it occurs. To the Haines index, since is a categorical variable, we did not preform means, we did spatiotemporal modes. We also had to recalculate the mean wind speed and direction for each level from their mean u and v components, to have accurate values.

With all the fire progressions spacial variables averaged we can calculate some different temporal metrics like the:

- Boundary Layer fluctuation Rate, calculated by taking the difference between the temporal maximum and minimum of the spatially averaged BLH during the progression and dividing it by the time elapsed between these extremes,
- Recirculation Index, measuring how much the wind vectors change direction over the progression period (1 − L/S, following Russo et al.)
- Circular Variance, quantifying the dispersion of wind directions over time using circular statistics

3.2.4 Finalizing the Updated PT-FireSprd Database

To the PT-FireSprd database, after adding various geographic and meteorological variables, we will also be adding Rate of Spread Lags, the value of the ROS of the ongoing progression 1 hour before the progression started, and Total Fire duration, the time elapsed between the start of the fire and the start of the progression polygon.

The last variable added is *Fire Rank*, a categorical variable used to distinguish between progressions occurring simultaneously and those that are the only ones in their timeframe (from the same wildfire). It also helps in identifying cases where meteorological metrics could not be calculated, and will not be used in the modeling (negative ranks).

- Rank -4: Other missing values
- Rank -3: Partially overlapping intervals
- Rank -2: No meteorological data because the time interval does not include an hour
- Rank -1: Missing start or end date (because it could not be identified)
- Rank 1: Single fire front for specific fire at specific time
- Rank 2: Multiple fire fronts progression with the highest ROS
- Rank 3: Multiple fire fronts progression with the lowest ROS

When multiple progressions occur simultaneously, distinguishing between the one with the highest and lowest ROS is useful: the lowest ROS progression may be constrained by unidentified factors or suppression efforts that cannot be quantified, while the highest ROS progression is likely closer to the theoretical maximum ROS based on the environmental conditions that will be modeled. Therefore we will also not be using Rank 3 progressions for modeling.

In the final database, we excluded all *u* and *v* wind components and all variables from the ERA5 in *SL* dataset that are also available in the ERA5 Land, with lower precision and resolution.

The list of the final database's variables is included in the Annex B.

3.3 Data Exploration

During the analysis, we compared the distributions of the variables with the fire spread rate (ROS) and also their logarithmic transformations, in order to explore different types of relationships between the variables and the ROS. Specifically, we analyzed four relations: the linear relation (ROS vs variable), logarithmic relationship (ROS vs log(var)), exponential relationship (log(ROS) vs variable) and power relationship (log(ROS) vs log(var)). The log represents the safe logarithm, sign(x) ln (|x| + 1), for more comparable results. For each combination, we calculated the correlation, adjusted the corresponding linear regression, recorded the coefficient of determination (R^2) and the p-value of the correlation, allowing us to quantify both the strength and significance of the relationship. We will be focusing on the variables with more than 25% correlation.

We chose to exclude fires that started on 15 October 2017 from the analysis, as they were associated with very specific extreme weather conditions and very high ROS values, which could bias the results and conclusions obtained. On this date, *Hurricane Ophelia* affected the country, resulting in anomalous meteorological phenomena that reanalysis models may not be able to represent accurately. In addition, the progressions in this period have hourly resolution, leading to an overrepresentation of these events in the database, which compromises the validity of the adjustments and correlations in the models. This was also the reason for the removal of the Pedrógão Grande and Góis 2017 wildfires. The outliers of the Rate of Spread distribution (5 and 95 percentile) were also removed. In total, we excluded 35% of the dataset, 412 progressions out of the initial 1177. Thus, the exclusion of this data ensures a more robust and representative analysis of the typical relationships between variables and ROS. All circular variables were separated into their *sine* and *cosine*, transforming into continuous variables.

A complementary analysis was conducted to identify fire propagation regimes based on ROS_p (projected Rate of Spread). The goal is to group observations into ROS ranges that reflect consistent physical behavior, allowing assessment of whether the relationships between physical drivers and ROS change with its magnitude.

First, numerical data are preprocessed to ensure comparability: missing values are imputed, and variables are standardized. Categorical variables are appropriately handled so they can be included in the correlation-based analysis.

Next, the pipeline generates candidate ROS cut points using percentiles and uniform sampling and creates groups of observations based on these cuts. This approach allows exploration of different ROS segmentations to identify ranges corresponding to distinct propagation regimes.

The quality of the resulting groups is evaluated using two complementary metrics: Structural Score, that measures the heterogeneity between regimes by comparing differences in correlations between the physical drivers and ROS_p across clusters, and the Conditional ΔR^2, that quantifies the explanatory gain obtained by fitting separate models for each regime instead of a single global model.

The pipeline can automatically search for the optimal number of clusters and the best cut locations that maximize structural heterogeneity and explanatory gain. Each cluster is further analyzed through statistical summaries and the top correlations with ROS_p (*print_top_correlations*).

Finally, the results are compared visually with plots showing ΔR^2 versus number of clusters and the ROS_p distribution per cluster, facilitating physical interpretation and visual validation of the regimes.

This pipeline not only allows the identification of ROS intervals representing distinct propagation regimes, but also assesses the relative importance of the determining factors in each regime and quantifies the explanatory improvement achieved through regime-based analysis, compared to a global approach.

Another analysis was performed combining supervised machine learning, model interpretability, and unsupervised regime identification to explain the variability in the Rate of Spread (ROS) beyond

mean relationships. First, categorical variables were one-hot encoded, and all explanatory variables were imputed and standardized to ensure comparability. Next, a gradient boosting regression model (XGBoost) was trained to predict ROS, chosen for its ability to capture nonlinear relationships and interactions between variables, with model parameters optimized through cross-validation.

To interpret the model predictions, Shapley additive explanations (SHAP) were calculated, providing a consistent decomposition of each prediction into contributions from the explanatory variables. These SHAP values represent the dominant factors in fire spread for each observation and were used as the basis for identifying the regime, shifting the focus from the magnitudes of the raw variables to the underlying explanatory mechanisms.

The SHAP-based explanation matrix was standardized and projected into a low-dimensional space using *Uniform Manifold Approximation and Projection (UMAP)*. A cosine distance metric was adopted at this stage to emphasize the similarity in the direction and structure of the explanatory patterns, rather than their absolute magnitude. Subsequently, clustering was performed on the UMAP projection using the *HDBSCAN* algorithm, which was selected for its ability to automatically determine the number of clusters, identify clusters arbitrarily, and isolate outliers without forcing assignments. The *Excess of Mass (EOM)* cluster selection method was used to favor stable and well-separated regimes.

After clustering, each regime was interpreted statistically by analyzing the distribution of ROS and the relative importance of the explanatory variables within each cluster. Differences between clusters were evaluated based on the magnitude and sign of SHAP contributions, allowing the identification of dominant factors and contrasting fire propagation mechanisms between regimes. Visualization tools, including low-dimensional projections and summaries by cluster, were used to support the physical consistency and interpretability of the identified regimes.

3.4 Modeling

The modeling phase is crucial for the success of the project. Developing a model will not only allow for predictions of fire Rate of Spread given a set of input variables, but it will also provide insights on how variables relate to each other. In order to achieve this, two models were developed, a Complex Model, based on a XGBoost Regressor algorithm, and a Linear Model built using the Ordinary Least Squares (OLS) regression. The Complex Model should have the highest accuracy of the two models, being suited for when the highest precision is necessary. The Linear Model, even though has lower precision, has the benefit of being very easily interpreted and calculated. Since the dataset being used for modeling is relatively small, with only 1173 records, it was decided that, to achieve the maximum possible accuracy, both models would be trained on the full dataset, with all progressions.

3.4.1 Data handling & Feature engineering

In order to train both models, some data processing had to be performed to ensure the models could understand as clearly as possible the relationships between the input variables. Thus, only the progressions that had environmental variables calculated were selected. As previously stated, to to be more accurate in the modeling process, when multiple progressions of a fire are happening at the same time, only the one with the highest rate of spread was selected to be used in the modeling phase. This resulted in 1173 records suitable for modeling.

In both models, we will predict the ROS in the log-transformed space to address their skewed distribution, in order to normalize it and obtain more accurate results.

3.4.2 Complex Model

The complex model is based on a XGBoost algorithm for its ability to learn non-linear relationships between variables and overall robustness against missing data and overfitting[1], specially in smaller datasets like the one available. While the primary focus of the model is high accuracy, complexity is also extremely important.

Hyper-parameter tuning was conducted using *HalvingRandomSearchCV*[4], a successive halving algorithm that progressively allocates more computational resources to promising parameter combinations while eliminating poorly performing configurations early in the search process. The best hyper-parameter configuration identified during this search was saved for subsequent analysis.

To estimate the model's accuracy, using Repeated-K-Folds, the dataset was split in 5 parts, where 80% of the data was used for training and 20% for testing, this was repeated four times, originating twenty different train/test splits. The final accuracy value was calculated as a mean of the precision metrics of each test fold; the standard deviation and the confidence interval was also calculated.

The result of this pipeline is twenty different models that, when the performance metrics are averaged, give a very robust performance assessment unlike the more traditional 80/20% split only performed once. We found that using only one 20% test dataset was giving us biased metrics that did not correspond to the real performance of the model when presented with a more complete data set. However, it's important to note that this new pipeline doesn't provide direct performance metrics for the final model trained on the full dataset, since there is no data split being performed, it provides a very accurate approximation of the accuracy to expect from the final model, being a common practice applied across the Machine-Learning community[2,3] when dealing with limited records. This method can be visualized in Figure 3.

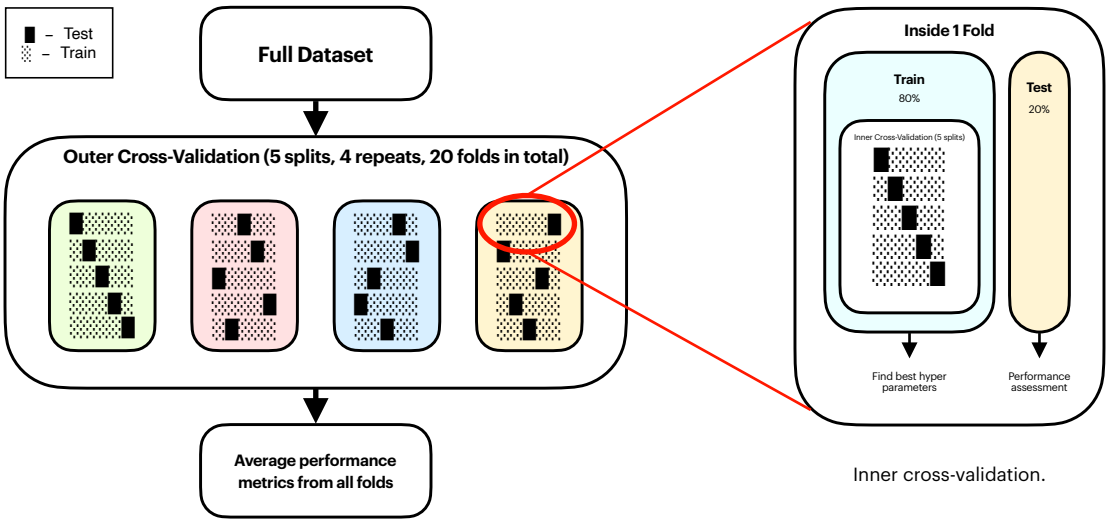


Figure 3 - Nested cross-validation methodology.

To assess the contribution of each input variable to model predictions and to reduce the model's dimensionality, *SHapley Additive exPlanations* (*SHAP*) values were computed using *TreeExplainer* on a

model trained with all 95 features. *SHAP* quantifies each feature's contribution to individual predictions by calculating contributions across all possible feature combinations. Feature importance was obtained by computing the mean absolute *SHAP* value for each feature across all samples in the dataset.

Following *SHAP* analysis, features were ranked by importance and the top *N* features were selected based on diminishing returns in predictive performance. An elbow method approach identified that six features provided good precision while substantially reducing model complexity compared to the full feature set.

Another aspect that was payed close attention to was the gap between the train R^2 and the test R^2. When the train dataset performance is significantly higher than the test dataset performance, the machine learning model may be overfitting to the train data. This gap should be evaluated considering the dataset characteristics and model complexity. We aimed for a 15% performance gap between training and testing datasets, which would suggest no overfitting was occurring. We optimized the model's complexity parameters to ensure the final model wouldn't also overfit when being trained on the full dataset.

3.4.6 Linear Model

The linear model provides an easily interpretable fire spread prediction equation, establishing fundamental relationships between input variables and rate of spread through a multiple linear regression. While less flexible than the XGBoost model, linear regressions offer clear mathematical relationships between variables that can be directly examined and validated against physical fire behavior principles. For each variable, their linear form and their safe logarithmic transformation were used as an input.

The same repeated Cross-validation methodology as the Complex Model was applied, but the linear regression pipeline contains three new sequential preprocessing steps before model fitting. First, missing values are imputed using median to maintain dataset completeness, as the model needs all values to be fitted. Second, features are standardized using z-score normalization to ensure all variables contribute on comparable scales, which is critical for linear regression where coefficient magnitudes directly reflect feature importance. Finally, the model is fit using *Ordinary Least Squares (OLS)*, learning linear relationships between features and the log-transformed target variable, the Rate of Spread.

To identify the most predictive features while minimizing model complexity, a forward feature selection algorithm was implemented. The selection process begins by evaluating all candidate features individually and selecting the univariate predictor that achieves the lowest mean absolute error in linear space. This initial feature establishes the baseline model performance. Subsequently, the algorithm iteratively evaluates all remaining features by adding each one to the current feature set, training a model, and calculating the resulting MAE. The feature that produces the largest improvement in MAE is permanently added to the selected set, this process was repeated 10 times. A critical constraint in the selection process is the exclusion of redundant linear and log-transformed versions of the same underlying variable. When a feature is selected, its corresponding transformation is automatically removed from consideration to prevent multicollinearity and ensure feature diversity. The evolution of R^2 and MAE across iterations reveals an elbow point where additional features provide minimal performance gains, guiding the final selection of the number features for the final model.

After selecting the optimal features, the final linear regression model was trained on the complete dataset using all available records. The resulting model coefficients represent the marginal effect of each standardized input variable on the log-transformed rate of spread. These coefficients can be

converted back to linear scale to show multiplicative effects, making it easier to understand how variables affect fire spread rates.

The final model equation takes the form of a linear combination of selected features and allows straightforward calculation of predicted fire rate of spread values with direct examination of feature contributions to individual predictions.

3.5 Application Deployment

To make the developed fire spread prediction models accessible to end users, a web-based platform was designed and deployed. The platform enables fire rate of spread predictions across Portugal at a 0.1° spatial resolution, integrating the trained machine learning models with live meteorological data retrieval and interactive geospatial visualization. The design prioritizes both usability and information density, being suited for researchers and operational firefighters requiring rapid field assessments during active incidents. It should be noted however, that for this interface to be used in real-time the CDS API cannot be used, since some features are only available three months prior to the present. Instead, an alternative weather *API*, ideally one with forecast capabilities should be used for this task and for the platform to be able to predict in the future.

The application follows a client-server architecture where a *Flask API* serves as the backend, handling prediction requests, data processing, and model inference, while a responsive frontend built with *HTML*, *JavaScript*, and *CSS* provides the user interface and visualization layer.

3.5.1 Frontend

The frontend was developed using *HTML*, *JavaScript*, and *CSS*, implemented as a single-page application in a *index.html* file. The single-file architecture consolidates all client-side logic and styling, making the application easier to maintain. *JavaScript* handles user input validation, *API* requests to the *Flask* backend, and real-time rendering of prediction results, while CSS provides styling to ensure usability across different screen sizes. The interface features an elegant *glassmorphism* design with modular panels for model selection, temporal parameters, and display controls positioned on the left side, while the center showcases an interactive map of Portugal with a color-coded grid visualization of fire spread predictions. Additional interface components include a timeline control for temporal analysis, real-time statistics showing minimum, average, and maximum rates of spread, map style toggles (default, satellite, terrain), and *CSV* export functionality for data analysis.

To run a prediction, the user inputs parameters in the top-left panel including the desired model, the fire event date and time, the prediction duration in hours, and the time elapsed since fire ignition. Additional model information is accessible via the "i" button next to the "MODEL" title. Once submitted, the backend accesses the *CDS API* to fetch the necessary meteorological variables for the selected temporal window. This process typically takes up to one minute depending on the prediction duration requested. After retrieving and processing the weather data, the system accesses pre-stored , yearly constant, geographic variables from *TIF* files in the backend. The model then executes predictions for each 0.1° grid cell across Portugal, generating output *TIF* files that are rendered as a color-coded heatmap on the interactive map, with fire spread rates visualized according to the legend displayed at the bottom-left of the interface.

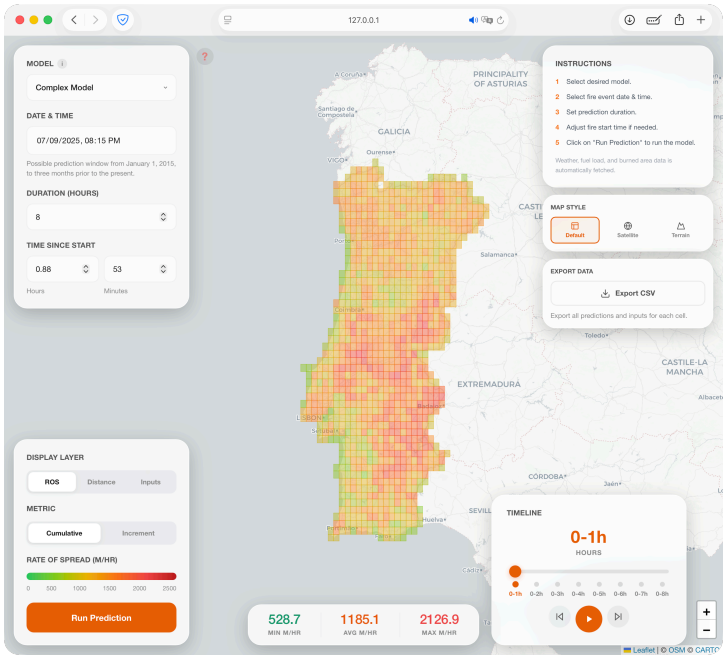


Figure 4 - Frontend interface and design.

On the bottom left, the user interface allows the user to display different data layers on the map. The default view shows fire rate of spread values, but distance traveled and individual input variables can also be selected and visualized. Below that, an option toggles between cumulative and incremental data display modes. The incremental option becomes particularly relevant for predictions exceeding one hour in duration, enabling users to observe how values evolve between consecutive time steps.

On the top right, a set of step-by-step instructions guides users through the prediction workflow. Below the instructions panel, three map style options are available. The default style presents a clean cartographic view showing regions and roads, while satellite and terrain options provide aerial imagery and topographic visualization respectively. Further below, an export functionality allows users to download all input variables and prediction results as a CSV file for external analysis.

There is also more information available when placing the mouse over the “?” button next to the top left panel. Information on how the the dataset was created, models were trained and weather variables were acquired are displayed.

Each cell on the map can be clicked to display additional data like the specific fire rate of spread value for the cell, the input variable’s values, distance traveled, and others, like seen in Figure 5. There is also a button to view charts of fire rate of spread and distance over duration, which helps understand how the fire might progress inside the cell over time. This can be seen in Figure 6.

When the user selects a duration higher than one hour, a timeline user interface appears in the bottom right. This timeline allows for precise visualizations of how the fire rate of spread changes through time. It also adapts to the metrics being displayed, wether it’s cumulative or incremental.

On the bottom, the minimum, average and maximum values are displayed of whatever variable is selected. By default this shows the fire rate of spread, but it can also show the distance and all other input variables necessary to run the model.

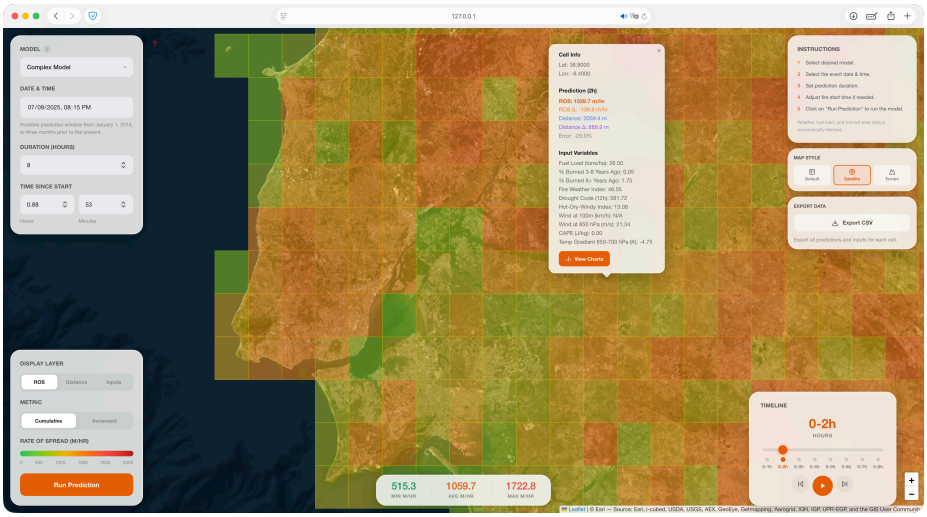


Figure 5 - Example of simulation performed using Satellite map view. Showcase of functionality to click on cell to obtain more information.

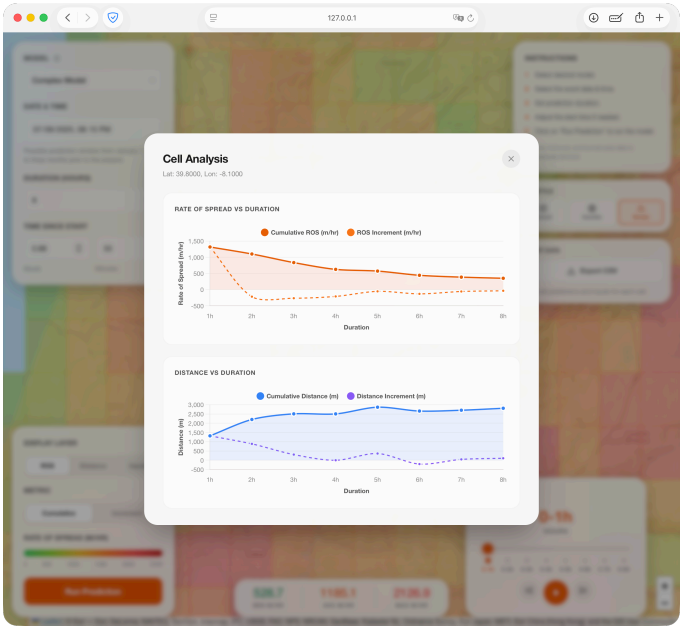


Figure 6 - Rate of spread vs Duration and Distance vs Duration charts.

3.5.2 Backend

The backend is structured around six Python scripts, each handling a specific component of the data processing pipeline. These scripts work sequentially to retrieve meteorological data, process it into features appropriate for the model, and generate wildfire spread predictions. The main application file, *app.py*, handles these components and serves as the Flask web server connecting the frontend to the prediction system.

The *CDS_API.py* script manages communication with the *Copernicus Climate Data Store API*, downloading four categories of *ERA5* climate data including single-level variables such as surface temperature and wind, pressure-level atmospheric conditions, Fire Weather Index data, and ERA5-Land

reanalysis products. The script also has caching and checks for existing local data before initiating API calls, minimizing redundant downloads.

Meteo_vars.py processes raw ERA5 data into meteorological variables for model input. It interpolates all datasets to a common 0.1° grid covering continental Portugal and calculates derived variables including wind speed at 850 hPa, vapor pressure deficit, the Hot-Dry-Windy index, and temperature gradients between atmospheric layers. These variables are the meteorological features used by the machine learning models.

Meteo_dataset.py assembles the complete meteorological dataset for the requested temporal window. It determines which hourly ERA5 data is required, calls *CDS_API.py* to retrieve any missing files, and applies *Meteo_vars.py* for processing. The output is a structured *xarray* dataset containing all necessary meteorological variables.

Create_inputs.py merges meteorological data with geographic information layers. The script integrates processed weather data with GIS datasets containing fuel load estimates, historical fire occurrence patterns, and terrain characteristics. It applies a spatial mask to filter cells outside Portugal's boundaries, calculates temporal aggregations over the requested duration, and saves results to *NetCDF* files.

Model_Prediction.py executes the final prediction stage by loading the machine learning models and generating fire spread predictions. It formats the compiled input data according to each model's requirements, runs predictions using both XGBoost and linear regression models and stores results in a *Master Table* that serves as a cache for subsequent requests with identical parameters.

The *app.py* file coordinates the entire workflow. When receiving a prediction request from the frontend, it validates input parameters including date, duration, and time since fire start, then checks whether predictions for this scenario exist in the *Master Table* cache. If cached results are available, they are retrieved immediately. Otherwise, *app.py* initiates the previous pipeline. Once predictions are generated, the system filters results to Portugal's geographic extent, formats predictions for each requested hourly interval, generates GeoTIFF output files, and transmits the results to the frontend for visualization.

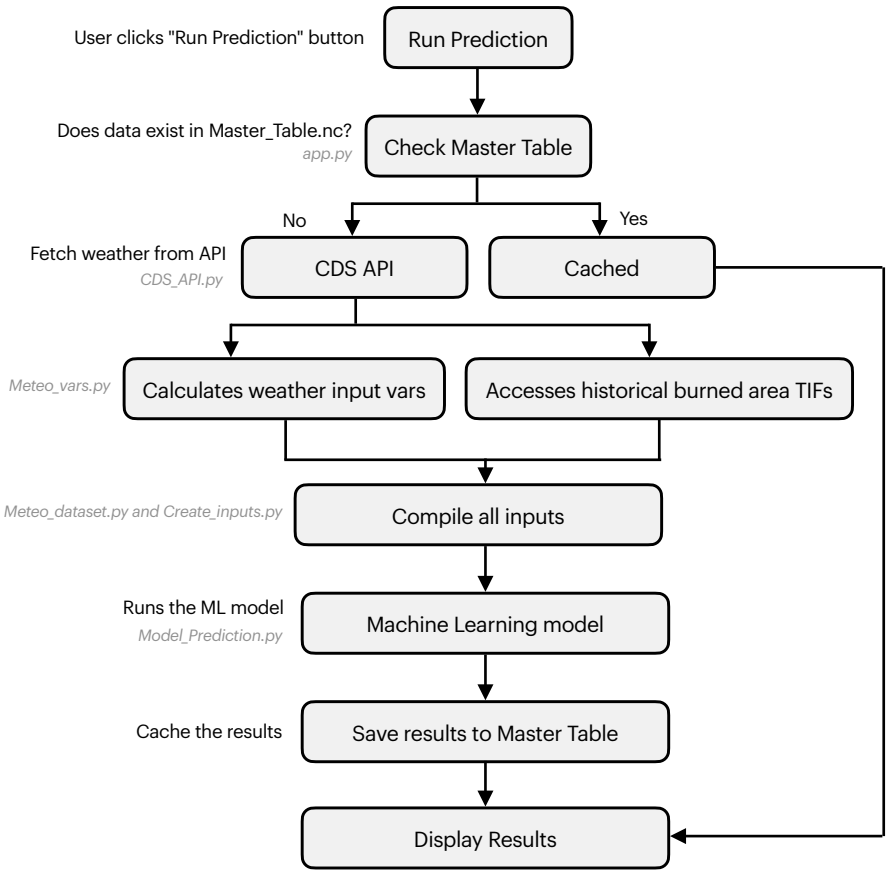


Figure 7 - Backend architecture.

It's important to mention that there are some temporal and spatial limitations. Temporal in that the duration needs to be defined, the fire date needs to be after January 1, 2015, and before three months from the present, and spatially it was assumed that in the duration defined by the user, the fire wouldn't leave a 0.1º by 0.1º degree cell, which is the case in 65-85% of the records. This had to performed due to the way the PT-FireSprd was created.

4. Results and Discussion

4.1 The updated PT-FireSprd database

The updated database comprises 1177 fire progressions characterized by 99 environmental and topographic variables that reflect the main physical drivers of fire spread. These include meteorological conditions (wind speed and direction at multiple atmospheric levels, temperature, relative humidity, VPD), atmospheric stability indices (CAPE, CIN, Haines Index), fire weather indices (FWI, FFMC, DC), fuel moisture content, and terrain attributes.

Descriptive statistics reveal substantial variability across the model input variables (Table 9 and 10). Wind speed at 850 hPa ranged from 0.3 to 66.2 m/s (Coefficient of Variation (CV) = 61%), representing the strongest bivariate predictor of ROS (r = 0.38). The Hot-Dry-Windy index (HDW) spanned from 513 to 111,052 (CV = 71%), while the Drought Code (DC) ranged from 78 to 1,171 (r = 0.30). Atmospheric instability, captured by CAPE, exhibited extreme right skewness (2.96), with most progressions occurring under stable conditions (median ≈ 0 J/kg) but values reaching 1,272 J/kg during

convective events. The temperature lapse rate between 800-700 hPa (gT_8_7) averaged -7.5 °C/km, consistent with typical atmospheric stratification. Soil water content (0-100 cm) showed limited variability (CV = 19%), while fire history variables indicated that most progressions occurred in areas where over 59% of the landscape had burned more than 8 years prior. Progression durations ranged from 10 minutes to 23 hours (median = 1.5 h), with shorter progressions associated with higher spread rates (r = -0.31).

The target variable, Rate of Spread (ros_p), has a pronounced right-skewed distribution (skewness = 2.85, kurtosis = 10.62), with a mean of 900.9 m/h significantly exceeding the median of 518.2 m/h. Values range from 17.0 to 8,949.2 m/h, with a coefficient of variation of 122.2%. The D'Agostino-Pearson test confirms non-normality (statistic = 736.6, $p < 10^{-160}$), suggesting a log-transformation likely appropriate for modeling. The complete list of variables is provided in Annex A.

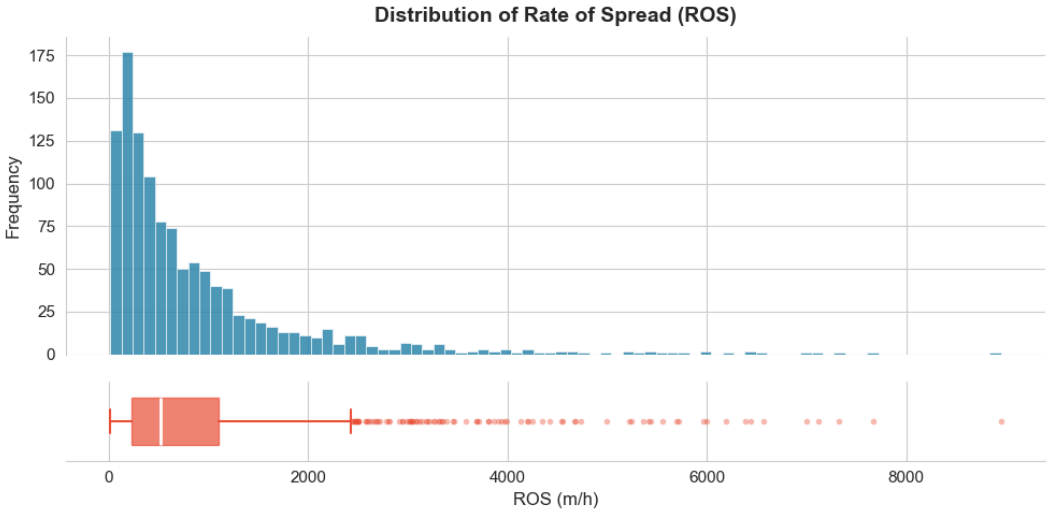


Figure 8 - Histogram and box plot of ROS.

4.2 Global Rate of Spread drivers analysis

The correlation analysis, for the four relations (linear, logarithm, power-law and exponential) showed that there are 21 variables with more than 25% correlation with the Rate of Spread, 20 of those representing power-law relations. The variables with the highest values are the Rate of Spread lag, progression duration, Hot-Dry-Windy index, and wind velocity at 10m and 100m.

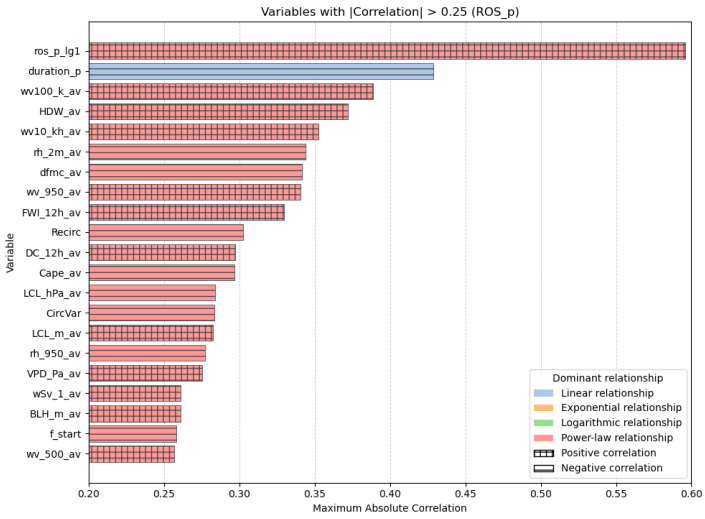


Figure 9 - Bivariate correlation analysis between environmental variables and ROS (|r| > 0.25), colored by dominant relationship type.

4.2.1 Temporal dynamics of fire propagation

Progressions with high ROS values tend to have short durations, evidenced by the strong negative relationship between progression duration and ROS (r = −0.429). This indicates that, shorter progressions, were probably defined due to the rapid advance of the fire, and longer progressions were traced due to the absence of data sources, or also due to a slow and steady rate of spread at night, for example. This strong relationship is more an artifact of the methodology used to define progression than a relationship with ROS.

The ROS observed in the immediately preceding progression (ros_p_lg1) has the highest absolute correlation among all variables analyzed (|r| = 0.596), following a power relationship. This result shows a strong temporal persistence of propagation, indicating that the recent behavior of the fire is the main predictor of its immediate evolution and that the dynamics of the system show high autocorrelation over time.

4.2.2 Fuel Moisture characterization

Dead Fuel Moisture Content (DFMC) exhibits a strong negative relationship with ROS (|r| = 0.342), confirming the central role of the water status of fine fuels in controlling fire spread. This control is closely linked to atmospheric dryness: relative humidity at 2 m (rh_2m_av) also shows a strong negative power relationship with ROS (|r| = 0.344), constituting one of the most direct atmospheric constraints on spread. Progressions with high ROS are rare under high relative humidity, further supporting the presence of physical thresholds governing fuel–atmosphere coupling. The relevance of atmospheric moisture extends beyond the surface layer, as relative humidity at lower tropospheric levels (rh_950_av) also displays a significant correlation with ROS (|r| = 0.277). This indicates that fire propagation responds to a vertically coherent dry atmosphere, favoring sustained fuel desiccation rather than to near-surface conditions alone. In this context, the *Vapor Pressure Deficit (VPD)* emerges as a key integrative variable, presenting a significant positive relationship with ROS (|r| = 0.276). By synthesizing the combined effects of temperature and humidity, VPD provides a more effective explanation of extreme fire spread than either variable considered in isolation.

4.2.3 Wind and atmospheric circulation

Wind emerges as one of the main controls of ROS. Average wind speeds at 10 m (wv10_kh_av) and 100 m (wv100_k_av) show very strong power relations (|r| = 0.353 and |r| = 0.389, respectively), confirming the dominant role of wind in flame inclination and efficient heat transport.

Wind shear between 10 m and 100 m (wSv_1_av) exhibits a positive correlation with ROS (|r| = 0.261), highlighting the role of vertical wind gradients in fire behavior. Strong wind shear can tilt flames, enhance turbulent mixing, and promote spotting by lofting embers, thereby facilitating faster and more erratic propagation, particularly under conditions of already high surface wind speeds.

Variables related to wind and smoke dynamics, such as *Recirc* (coastal smoke recirculation) and *CircVar* (circular variance of wind direction), show strong negative correlations with ROS (|r| ≈ 0.29). These results indicate that greater wind variability and pronounced recirculation patterns can locally reduce ROS by enhancing turbulence, which mixes different air masses, and by promoting smoke recirculation, which can limit oxygen supply to the fire and repeatedly deflect or split the advancing fire front. In contrast, more stable and directionally consistent winds favor continuous and rapid propagation.

	Cluster -1		Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
1	HDW_av	0.0889	CircVar	0.0792	HDW_av	0.0604	HDW_av	0.1112	CircVar	0.0895	HDW_av	0.1332	HDW_av	0.0913
2	CircVar	0.0572	HDW_av	0.0683	CircVar	0.0504	CircVar	0.0577	HDW_av	0.0790	wco_500_av	0.0477	CircVar	0.0759
3	wco_500_av	0.0409	rh_300_av	0.0641	wsin10_av	0.0474	wco_500_av	0.0380	wco_500_av	0.0470	CircVar	0.0408	wco_500_av	0.0375
4	f_load_av	0.0306	wco_500_av	0.0486	rh_700_av	0.0467	wsin10_av	0.0333	wv100_k_av	0.0335	Liftldx_av	0.0329	wv100_k_av	0.0310
5	wv100_k_av	0.0281	wwv_500_av	0.0456	wco_500_av	0.0435	wv100_k_av	0.0321	8_ny_fir_p	0.0319	f_load_av	0.0295	Recirc	0.0301
6	wScos_9_av	0.0278	Recirc	0.0348	wScos_9_av	0.0372	f_load_av	0.0268	wScos_9_av	0.0298	wScos_9_av	0.0283	wScos_9_av	0.0289
7	wsin10_av	0.0264	f_load_av	0.0296	f_load_av	0.0339	rh_700_av	0.0259	wSv_1_av	0.0286	wv100_k_av	0.0231	DC_12h_av	0.0282
8	rh_700_av	0.0251	3_8y_fir_p	0.0225	wsi_950_av	0.0296	Recirc	0.0233	Recirc	0.0262	LCL_hPa_av	0.0223	f_load_av	0.0241
9	DC_12h_av	0.0246	wsin10_av	0.0223	wv100_k_av	0.0259	wScos_9_av	0.0206	rh_700_av	0.0237	rh_700_av	0.0209	wsin10_av	0.0240
10	Recirc	0.0235	DC_12h_av	0.0223	wwv_500_av	0.0258	wv10_kh_av	0.0201	wsin10_av	0.0235	wsin10_av	0.0186	wwv_500_av	0.0228

(Outliers)	Irregular Pyroconvection driven Propagation	Windy-Dry Propagation	Extreme Weather driven Complex Propagation	Wind-Turbulence driven Propagation	Extreme Drought driven Propagation	Wind-Turbulence Limited Propagation

ROS Response		ROS Response		ROS Response		ROS Response		ROS Response		ROS Response		ROS Response	
average	608.9	average	751.4	average	541.6	average	945.1	average	534.0	average	986.7	average	326.6
median	515.0	median	561.0	median	350.5	median	854.1	median	434.1	median	884.7	median	259.0
standard dev.	560.2	standard dev.	600.7	standard dev.	487.7	standard dev.	588.5	standard dev.	320.2	standard dev.	516.6	standard dev.	233.3

Figure 11 - Fire propagation regimes identified by HDBSCAN clustering, showing dominant drivers and ROS statistics per cluster.

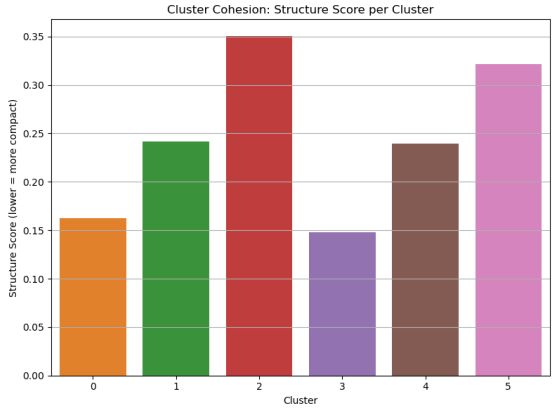


Figure 12 - Cluster cohesion measured by structure score (lower = more compact).

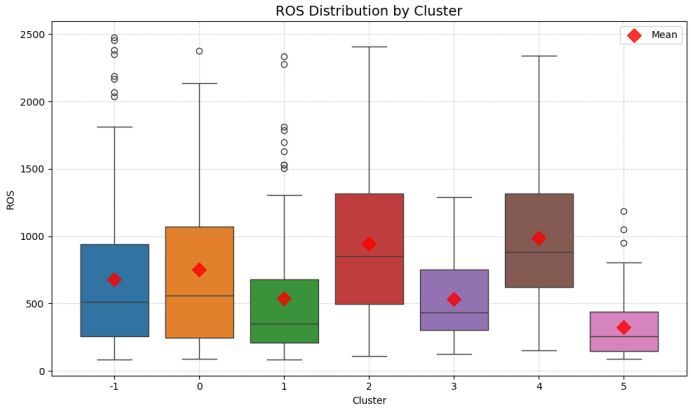


Figure 13 - ROS distribution across identified propagation regimes, with mean (diamond) and median (line) indicated.

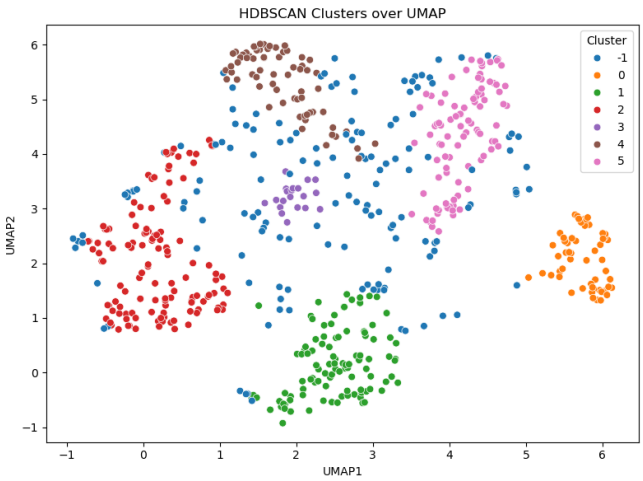


Figure 14 - UMAP visualization of HDBSCAN clusters, showing spatial separation of propagation regimes in reduced feature space.

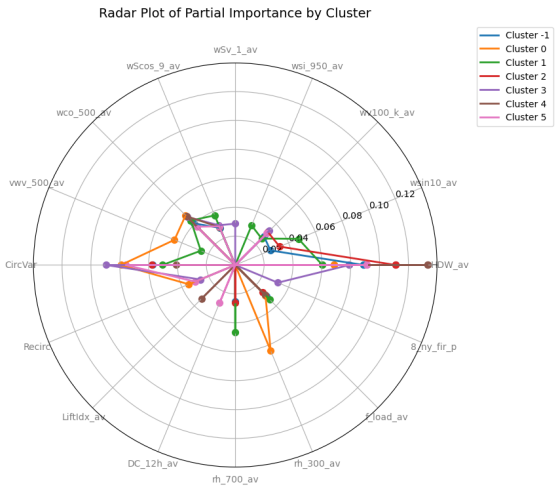


Figure 15 - Variable importance profiles across propagation regimes, highlighting distinct drivers for each cluster.

Wind at 500 hPa (wv_500_av) exhibits a moderate positive relationship with ROS (|r| = 0.257), indicating that larger-scale wind patterns in the mid-troposphere can contribute to fire propagation. Although the effect is weaker than that of near-surface winds, elevated winds at 500 hPa may enhance the advection of heat, smoke, and dry air into the boundary layer, indirectly supporting the maintenance of high ROS values during extreme events.

4.2.4 Vertical structure of the atmosphere

CAPE shows a strong negative correlation (|r| = 0.297), indicating that episodes of rapid spread occur predominantly in stable and dry atmospheres, which are not very favorable for the development of deep convection. The *Lifted Condensation Level* (LCL) consistently appears among the most relevant variables, both when expressed in pressure (LCL_hPa_av, r ≈ −0.284) and in metres (LCL_m_av, r ≈ 0.283), reflecting drier and deeper atmospheres, which are highly favourable to fuel drying.

Boundary Layer Height (BLH) has a significant positive correlation with ROS (|r| = 0.261). However, the most extreme ROS values tend to be concentrated in situations with a relatively low boundary layer, suggesting that the confinement of heat, dryness and wind in a shallower layer can intensify propagation.

4.2.5 Other Indices

Among the composite indices, the *Hot-Dry-Windy* index (HDW) stands out with one of the highest correlations in the entire set (|r| = 0.372), confirming its ability to effectively synthesize the main physical controls of rapid fire spread, combining fuel dryness and wind speed in a single variable.

Drought Code (DC_12h_av) and *Fire Weather Index* (FWI_12h_av) show positive correlations with ROS (|r| = 0.297 and |r| = 0.330, respectively), reflecting the influence of accumulated fuel dryness and integrated fire danger on propagation. DC captures long-term moisture deficits in deep fuels, while FWI provides a more comprehensive measure of fire potential by combining weather and fuel conditions. Together, these indices allow to effectively explain variations in ROS beyond the effect of instantaneous meteorological variables.

4.3 Analysis of different regimes of fire behavior

To assess whether there are different determinants of ROS based on their magnitude, the clustering procedure was performed for 1 to 5 clusters, evaluating the segmentation according to three complementary criteria: the structural score, which measures the differentiation of correlations between clusters; the conditional mean ΔR^2, which quantifies the explanatory gain when segmenting compared to a global approach; and the normalized sum of statistically significant correlation scores, as it reflects the overall strength and relevance of the relationships captured within each cluster. This multi-metric assessment allowed for an informed selection of the optimal number of clusters, balancing both the clarity of separation and the explanatory power of segmentation.

For 1 cluster, where all observations remain together, and there are no differentiated regimes; the structural score and conditional ΔR^2 are zero. Subsequent clusters show an increase in the structural score, from 0.21 to 0.77 when considering 2 to 5 clusters, indicating that the correlations between ROS_p and the explanatory variables become progressively more distinct as segmentation increases.

However, the average conditional ΔR^2 is negative in all segmentations with more than 1 cluster, suggesting that, although the structural score increases with a higher number of clusters, suggesting more differentiated correlation patterns within the smaller groups, the negative average conditional ΔR^2 highlights that these subdivisions do not enhance the overall explanatory power. As an example, for two clusters, the ΔR^2 is negative but close to zero and the dataset is split at ~2000m/h and the higher cluster has ~35% percent correlation with the soil water content at 100cm and 28cm, however, there is not a gain in correlation inside both clusters, it decreased a lot.

The apparent differentiation arises mainly from noise captured by smaller clusters, generating different but extremely weak correlation patterns with ROS, rather than significant changes in the regime. Therefore, considering both robustness and explanatory gain, the optimal segmentation is a single cluster, reflecting that ROS_p exhibits consistent global correlate ion patterns in it's magnitude, with the explanatory variables and does not naturally fragment into distinct, clearly interpretable regimes.

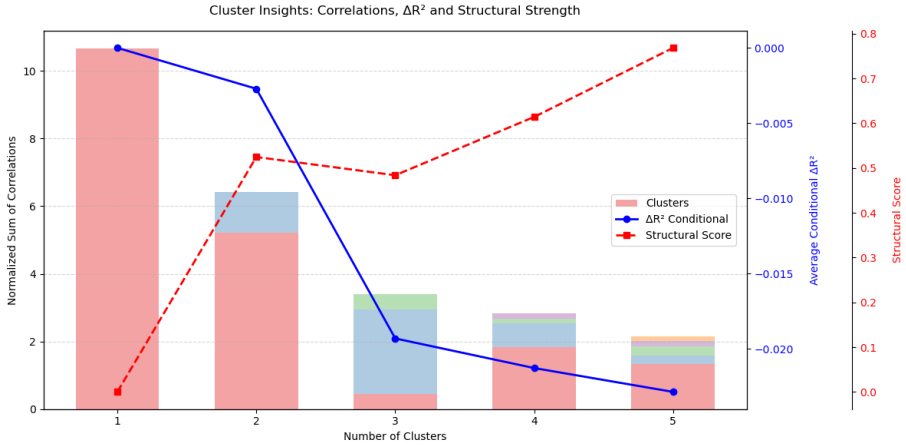


Figure 10 - Trade-off between cluster differentiation (structural score) and explanatory power (ΔR^2) for ROS segmentation.

The *HDBSCAN* clustering identified six clusters in the dataset, six different sets of observations with similar characteristics in relation to the explanatory variables of the ROS, with another 147 points classified as outliers (cluster −1).

Cluster 0 corresponds to an irregular pyroconvective propagation regime, dominated mainly by atmospheric dynamics and not only by surface conditions. The high variability of circulation (CircVar = 0.0792) and intense recirculation (Recirc = 0.0348) indicate an unstable and heterogeneous wind field, favoring erratic fire behavior. The strong influence of wind and vertical velocity at 500 hPa (wco_500 = 0.0486; wvv_500 = 0.0456) reveals the presence of persistent updrafts, while high relative humidity at 300 hPa (rh_300 = 0.0641) favors the development and maintenance of deep convective columns associated with pyroconvection. Together, these factors promote strong fire-atmosphere coupling, resulting in high but highly variable propagation rates, consistent with the high average values and large dispersion of ROS observed in this cluster.

Cluster 1 characterizes a surface propagation regime strongly influenced by wind and fuel dryness. The fire front is mainly driven by near-surface wind (wsin10 = 0.0474; wsi_950 = 0.0296) and near-ground wind gradients (wScos_9 = 0.0372), determining the direction and rate of fire advance. The presence of updrafts at 500 hPa (wvv_500 = 0.0435) suggests that there is some vertical transport of heat and smoke, but it does not generate intense pyroconvection. Finally, the average fuel load (f_load = 0.0339) and relative humidity at 700 hPa (rh_700 = 0.0467) modulate fuel drying, favoring faster propagation on days with relatively dry air at medium levels. In this regime, propagation is mainly controlled by wind and fuel dryness, with some secondary effect from vertical atmospheric dynamics.

Cluster 2 characterizes a complex propagation regime driven by extreme surface conditions. Propagation is mainly guided by near-surface wind (wsin10_av = 0.0333) and wind at intermediate heights, such as 100m (wv100_k_av = 0.0321), which determine the direction and speed of the fire front. This wind effect is reinforced by the Hot-Dry-Windy index (HDW_av = 0.1112), which combines high temperatures, low humidity, and strong winds, creating conditions that are extremely favorable for rapid fire spread. The fuel load (f_load_av = 0.0268) modulates the intensity and duration of the fire, allowing the fire to maintain high energy throughout its spread. Together, these factors define a regime of rapid and intense spread, with very high ROS and high variability, typical of extreme fire events.

Cluster 3 represents a propagation regime strongly influenced by wind turbulence and local recirculation. The fire front is driven by winds at intermediate heights (wv100_k_av = 0.0335; wco_500_av = 0.0470) and by the wind gradient near the ground (wScos_9_av = 0.0298; wSv_1_av = 0.0286), which create an irregular advance pattern. Circulation variability (CircVar = 0.0895) and air recirculation (Recirc = 0.0262) reinforce the erratic propagation, keeping the fire active and redistributing embers. The relatively high value of old fuel (8_ny_fir_p = 0.0319) suggests that this older material could generate more ash and long-lasting embers, contributing to increased ROS in this regime. Together, these factors define an irregular and turbulent propagation regime, with moderate ROS and difficult-to-predict behavior.

Cluster 4 characterizes a propagation regime strongly influenced by extreme surface conditions and accumulated drought, described as “extreme drought-driven propagation.” The main driver is the very high Hot-Dry-Windy index (HDW_av = 0.1332), which combines strong winds, high temperatures, and low humidity, creating conditions extremely favorable to rapid fire spread. Propagation is also modulated by the Lifted Index (LiftIdx_av = 0.0329), indicating that some atmospheric instability favors updrafts that can intensify propagation, although to a lesser extent than HDW. In addition, the fuel load (f_load_av = 0.0295) contributes to the intensity and duration of the fire, providing material available to sustain the fire front. Together, these factors define a regime with very high ROS, typical of days with extreme heat, wind, and extreme drought conditions.

Cluster 5 represents a regime described as “Wind-Turbulence Limited Propagation,” characterized by warm surface conditions and significant winds, but with relatively low ROS (average 326.6 m/h). The main drivers include the high Hot-Dry-Windy index (HDW_av = 0.0913) and high circulation variability (CircVar = 0.0759), which indicate strong wind influence and local instability. The presence of wind at intermediate height (wv100_k_av = 0.0310) and recirculation (Recirc = 0.0301) suggests that, despite the potential for propagation, atmospheric dynamics disperse part of the fire energy, limiting the rate of advance. Accumulated drought (DC_12h_av = 0.0282) and fuel load (f_load_av = 0.0241) provide sufficient material for combustion, but the combined effect of turbulence and instability creates irregular and localized propagation, explaining moderate ROS in this regime.

Cluster -1, which groups together outliers that did not fit into any of the main clusters, represents a propagation regime with varied and heterogeneous characteristics. The high Hot-Dry-Windy index (HDW_av = 0.0889) indicates that strong winds, high temperatures, and low humidity still influence fire propagation. The fuel load (f_load_av = 0.0306) provides sufficient material to sustain combustion, while the wind at intermediate height (wv100_k_av = 0.0281) and wind gradients near the ground (wScos_9_av = 0.0278) modulate the direction and rate of advance of the front. Together, these factors result in high but highly variable ROS (mean 608.9 m/h; standard deviation 560.2), reflecting the irregular and unpredictable nature of these fires, which do not clearly fit any specific regime pattern.

The results show that the Structure Score varies, with lower values indicating more compact and well-separated clusters (e.g., cluster 3, SS ≈ 0.15) and higher values showing more dispersed groups (e.g., cluster 2, SS ≈ 0.350). Thus, we can assume that clusters are better at capturing the nuances of ROS than a single cluster. Note that cluster -1, of the outliers, may be relevant but appears to represent intermediate conditions between clusters. These findings should be further analyzed and validated before using these conclusions to cement their robustness.

It would be interesting to further tune the XGBoost model in order to confirm these conclusions, as this was not possible due to hardware limitations.

4.8 Complex Model

The hyperparameter search space is summarized in Table 3. Tree depth was limited to 5–7 levels, balancing expressiveness with overfitting risk. The learning rate was sampled between 0.05 and 0.15 to allow conservative gradient updates. Subsampling parameters (subsample, colsample_bytree, colsample_bylevel, colsample_bynode) ranged from 0.6 to 0.9, introducing randomness at multiple levels to reduce variance and decorrelate trees. The min_child_weight (12–40) prevents splits on sparse data regions, while max_delta_step (0–5) constrains weight updates for stability. Regularization was controlled through gamma (0.5–2.0) for split penalty, reg_alpha (0.5–2.5) for L1 regularization, and reg_lambda (2.5–6.0) for L2 regularization. The relatively high L2 values favor smoother predictions.

Parameter name	Parameter range
max_depth	5, 6, 7
learning_rate	uniform(0.05, 0.10)
subsample	uniform(0.6, 0.3)
colsample_bytree	uniform(0.6, 0.3)
colsample_bylevel	uniform(0.6, 0.3)
colsample_bynode	uniform(0.6, 0.3)
min_child_weight	randint(12, 40)
gamma	uniform(0.5, 1.5)
reg_alpha	uniform(0.5, 2.0)
reg_lambda	uniform(2.5, 3.5)
max_delta_step	randint(0, 5)

Table 3 - Table with parameters range of XGBoost model.

Parameter name	Parameter value
n_estimators	320
max_depth	7
learning_rate	0.117
min_child_weight	28
gamma	0.863
subsample	0.790
colsample_bytree	0.896
colsample_bylevel	0.643
colsample_bynode	0.747
reg_alpha	1.956
reg_lambda	3.787
max_delta_step	2

Table 4 - Table with best parameters found for XGBoost model.

The elbow method defined that 7 input variables should be chosen, they can be found on Table 2.

The Complex model, based on the XGBoost algorithm, achieved an R^2 of 0.5565 (Figure 16) using the repeated k-fold cross-validation methodology mentioned previously. In log-transformed space, MAE was 0.459 and RMSE was 0.583. Converting to the original linear scale, these correspond to MAE of 384.3 m/h and RMSE of 728.3 m/h. It's important to note that this is just an approximation of the final model's performance, since it was trained on the full dataset. The model showed a train-test R^2 gap of 0.1546, suggesting a slight degree of overfitting, though still within acceptable bounds for modeling. However, the final deployed model was trained on all 1173 records rather than the ~938 used per fold, which should reduce this effect.

The selected features reinforce known drivers of fire behavior: fire duration relates directly to spread distance, burned land cover history reflects fuel load and continuity, while meteorological variables, wind speed, soil moisture and HDW, represent the atmospheric conditions conducive to fire

propagation. The model's reliance on these interpretable predictors supports its potential utility for operational applications and future research studies.

Variable name	Variable full name	Variable description
duration_p	Duration associated with ros_p	The time elapsed between two consecutive polygons.
3_8y_fir_p	Percentage of area burned between 3 and 8 years before	The percentage of area that was burned between 3 and 8 years before the progression's wildfire year.
8_ny_fir_p	Percentage of area burned more than 8 years before	The percentage of area that was burned more than 8 years before the progression's wildfire year.
sW_100_av	Average soil water content between 28 and 100cm depth	Average soil water content between 28 and 100cm depth.
f_start	Time since the begining of the wildfire	Minutes elapsed since the first ignition of this fire
HDW_av	Average Hot Dry Windy	Average index representing the potential for large fire growth under hot, dry, and windy conditions, Calculated temporal average from temperature, humidity, and wind.

Table 2 - Table of Complex Model input variables.

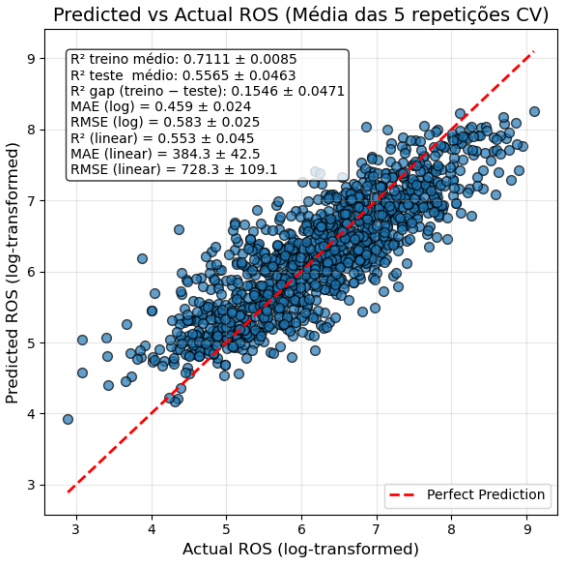


Figure 16 - All 20 Complex Model split's results condensed into one plot in logarithmic space.

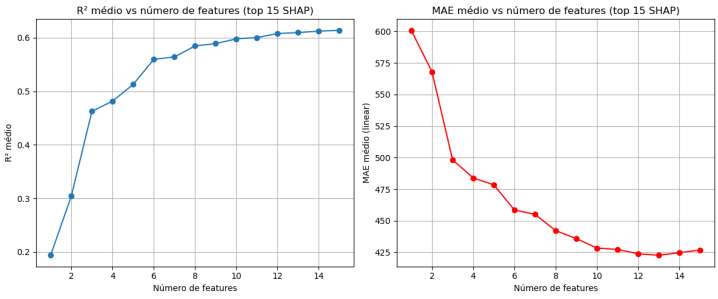


Figure 17 - Plots of R^2 and MAE versus number of input features of Complex model.

Metric	Value
R^2	0.5565
MAE *(m/h)*	384.3 ± 42.5
RMSE *(m/h)*	728.3 ± 109.1

Table 5 - Performance metrics of Complex model.

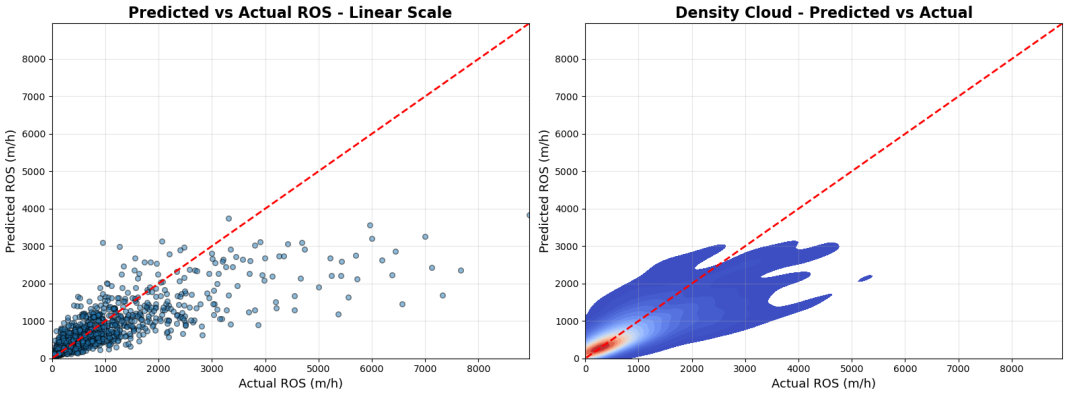


Figure 18 - All 20 Complex Model split's results condensed into one plot in linear space.

An extensive error analysis was conducted to characterize the model's predictive behavior across different conditions and magnitudes. The residual analysis in linear scale reveals patterns in prediction errors. In Figure 19, the residuals vs fitted values plot shows clear heteroscedasticity, with variance increasing for higher predicted values, also known as funnel effect, and a tendency toward negative residuals (underprediction) as predictions increase. The Q-Q plot confirms significant deviation from normality, indicating the presence of extreme errors. However, it's important to keep in mind that in logarithmic space these trends are no longer present and we see homoscedasticity and a normal distribution in the Q-Q plot. The reason why we're displaying the results in linear space is for better interpretability of predictions and corresponding errors. Since the model was trained in logarithmic space, these problems arise when converting it back to linear scale.

The bias analysis by ROS range is particularly revealing. The model exhibits positive bias (overprediction) for low-intensity fires in the 17-518 m/h range, with mean overestimation of 39-93 m/h. This transitions to negative bias (underprediction) for fires above 518 m/h, reaching -1021.6 m/h for the highest intensity range (1527-8949 m/h). This pattern indicates the model performs best for medium-intensity fires while struggling with extreme values at both ends of the ROS spectrum. In the lower interval we believe the reason is fire suppression from firefighters since they can more easily suppress slower-spreading fires, causing observed ROS values to be lower than they would be, under natural conditions. The exceptional fires from the October 2017, that had exceptionally high ROS values, represent statistical outliers with unique physical mechanisms that seem to deviate substantially from a expected fire behavior. Combined with the limited number of samples in this upper range, the model lacks sufficient examples to adequately learn these anomalous patterns, resulting in systematic underprediction of extreme fire spread rates. As was said previously, the environmental vairables might not be able to represent these very complex phenomena.

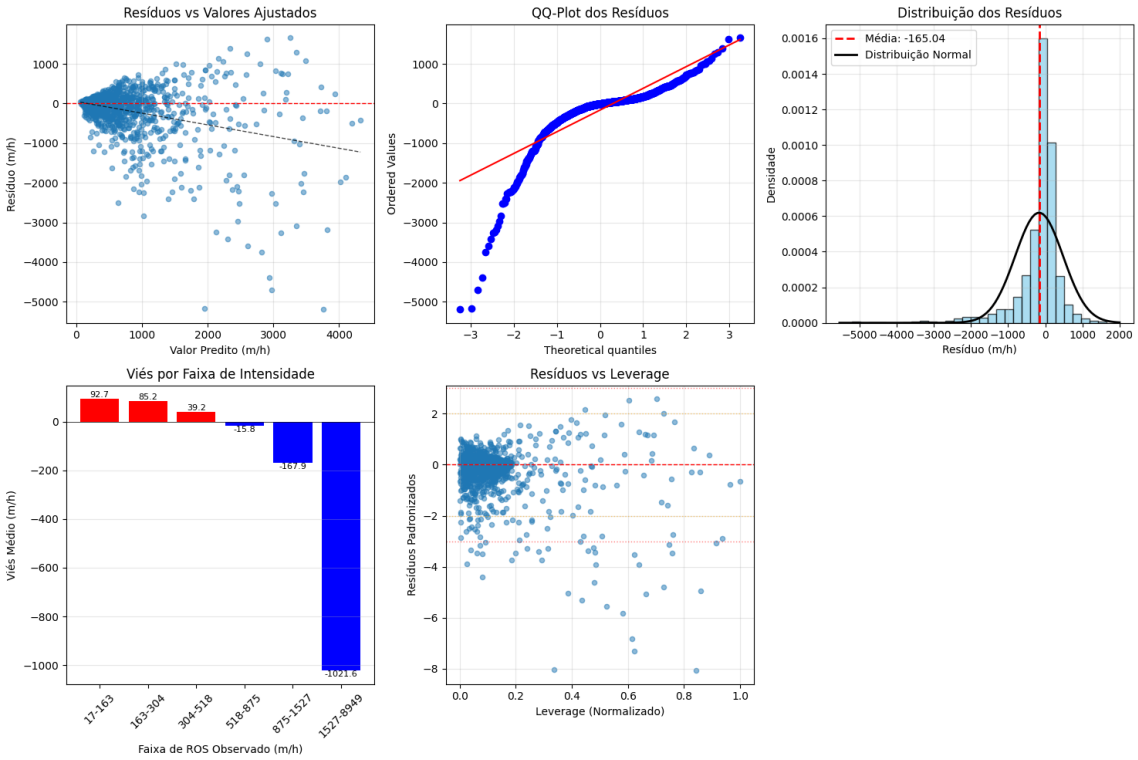


Figure 19 - Residual analysis of Complex model.

In Figure 20, we define bad predictions as the ones that register an error higher or equal to 35%. This is considered by the community to be the ideal maximum error allowed for fire predictions. The percentual error (erro percentual) vs frequency (frequência) plot shows that the majority of errors occur between -100% and 100%, with very few observations above 300%. Over half of the features were calculated to have an higher than acceptable error, this is affected by the modeling being done with the log transformation and not the linear value and the lack of progressions to train the model with among other sources.

The observed vs predicted scatter plot shows bad predictions occurring both at higher and lower ROS. It can also be observed that at higher observed ROS values the model systematically underpredicts, as discussed previously. The error comparison boxplots confirm substantially higher variance and more extreme outliers among poor predictions.

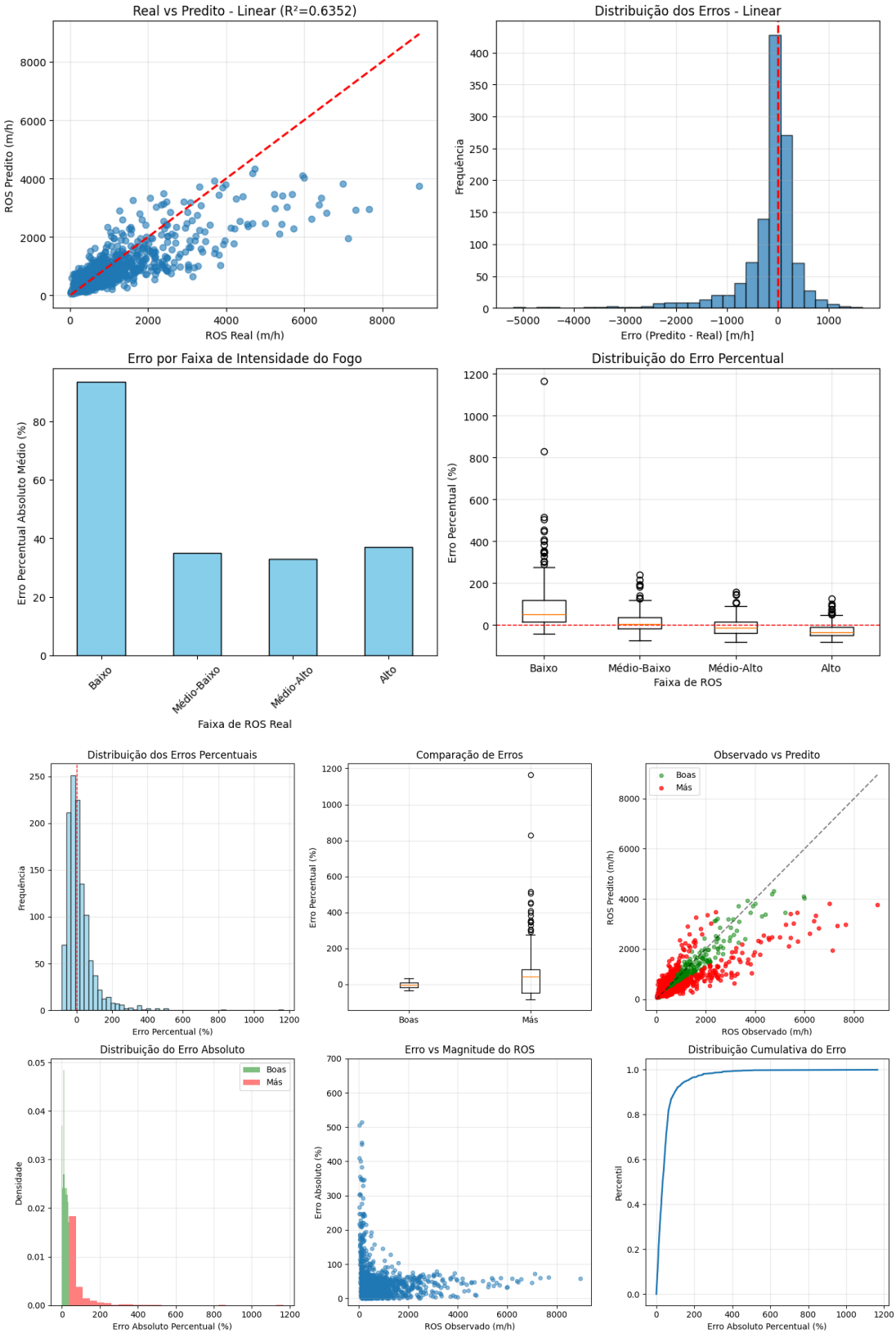


Figure 20 - Error analysis and identification of good and bad predictions of Complex model.

SHAP (SHapley Additive exPlanations) analysis provides insights on the relationships between the fire rate of spread and the input variables. Duration (duration_p) dominates feature importance with a mean |SHAP| of approximately 0.45, exhibiting a strong negative relationship, meaning, longer fire durations are associated with lower predicted ROS. The Fire Weather Index (FWI_12h_av) and wind velocity (wv100_k_av) show positive relationships consistent with fire behavior theory. Notably, the

difference in SHAP importance between good and poor predictions suggests that weather conditions introduce prediction uncertainty not fully captured by the available features, since there is essentially no difference between the two.

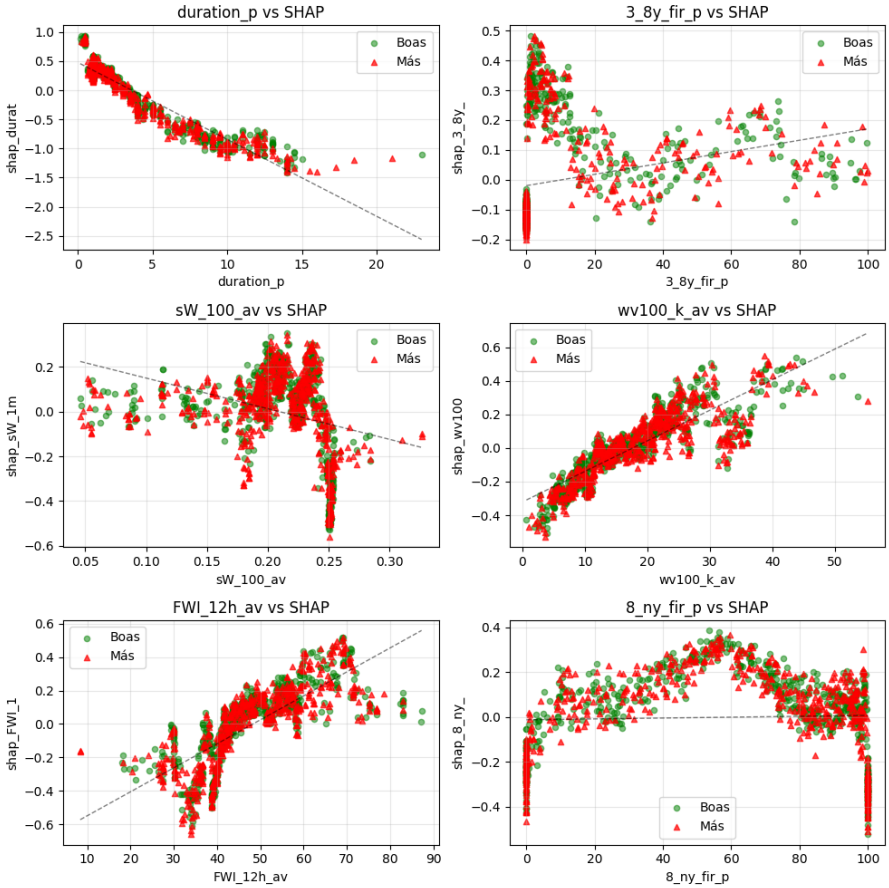


Figure 21 - SHAP value comparison between good and bad predictions across all input variables of Complex model.

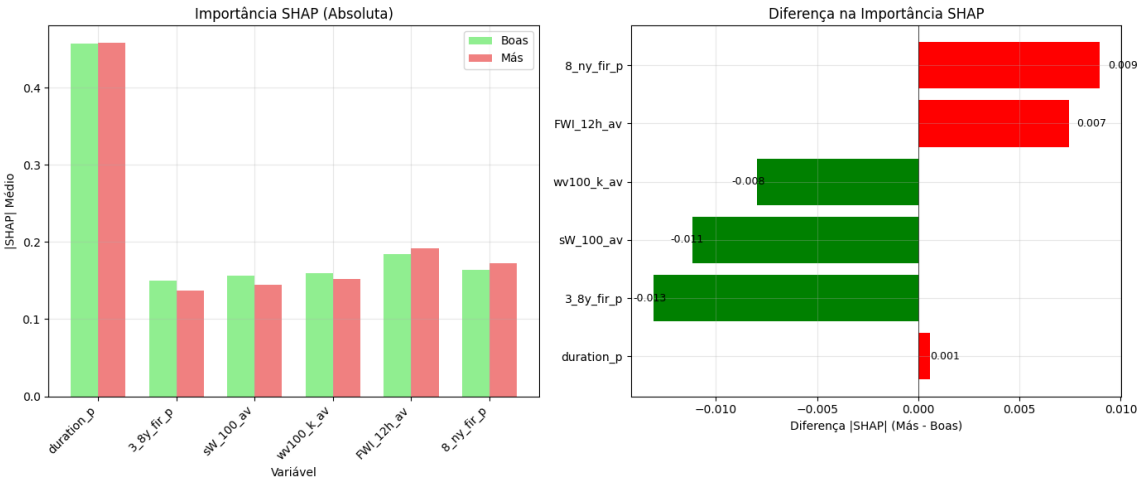


Figure 22 - Absolute mean SHAP value comparison between of good and bad predictions across all input variables of Complex model.

Another aspect analyzed was the presence of clusters within the prediction errors to identify systematic patterns in model failures. To determine the optimal number of clusters for the high-error observations, multiple validation metrics were employed (Figure 23). While the elbow method suggested k=2 and the Calinski-Harabasz index favored k=5, the silhouette score achieved its maximum value of 0.241 at k=6, which was adopted as the final configuration. The Davies-Bouldin index, though optimal at k=8, showed only marginal improvement beyond k=6. The hierarchical clustering dendrogram corroborated this choice, revealing natural groupings at approximately that level.

Clustering high-error predictions reveals that most clusters (0, 1, 3, 5) share a common failure, strong overprediction, with bias between +97% and +108% and errors exceeding 120%. Cluster 2 stands out with the lowest error (64%) and negative bias (-40%), despite representing fires with observed ROS 462% higher than low-error cases. This suggests the model handles intense fires better than slow-developing ones, where it systematically overshoots, possibly due to combat efforts. Cluster 4 falls in between (99% error, +48% bias). Sample sizes range from n=14 (Cluster 2) to n=108 (Cluster 1), so the Cluster 2 finding should be treated cautiously.

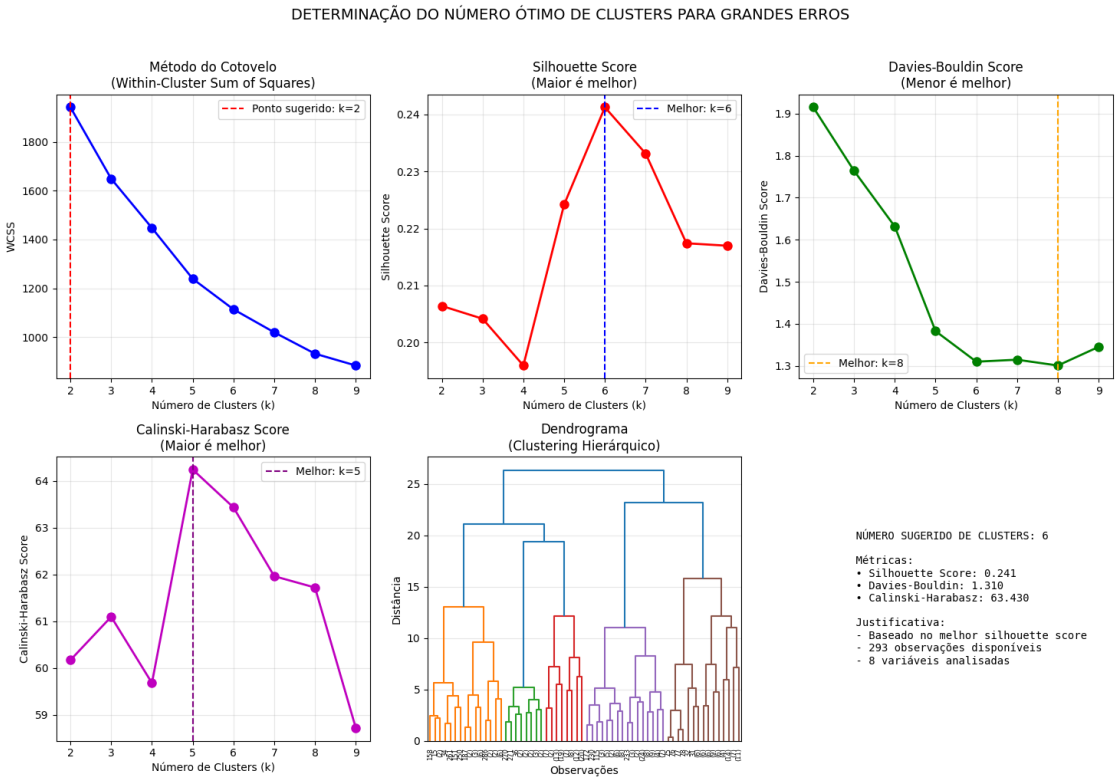


Figure 23 - Determining the best number of clusters for big errors of Complex model.

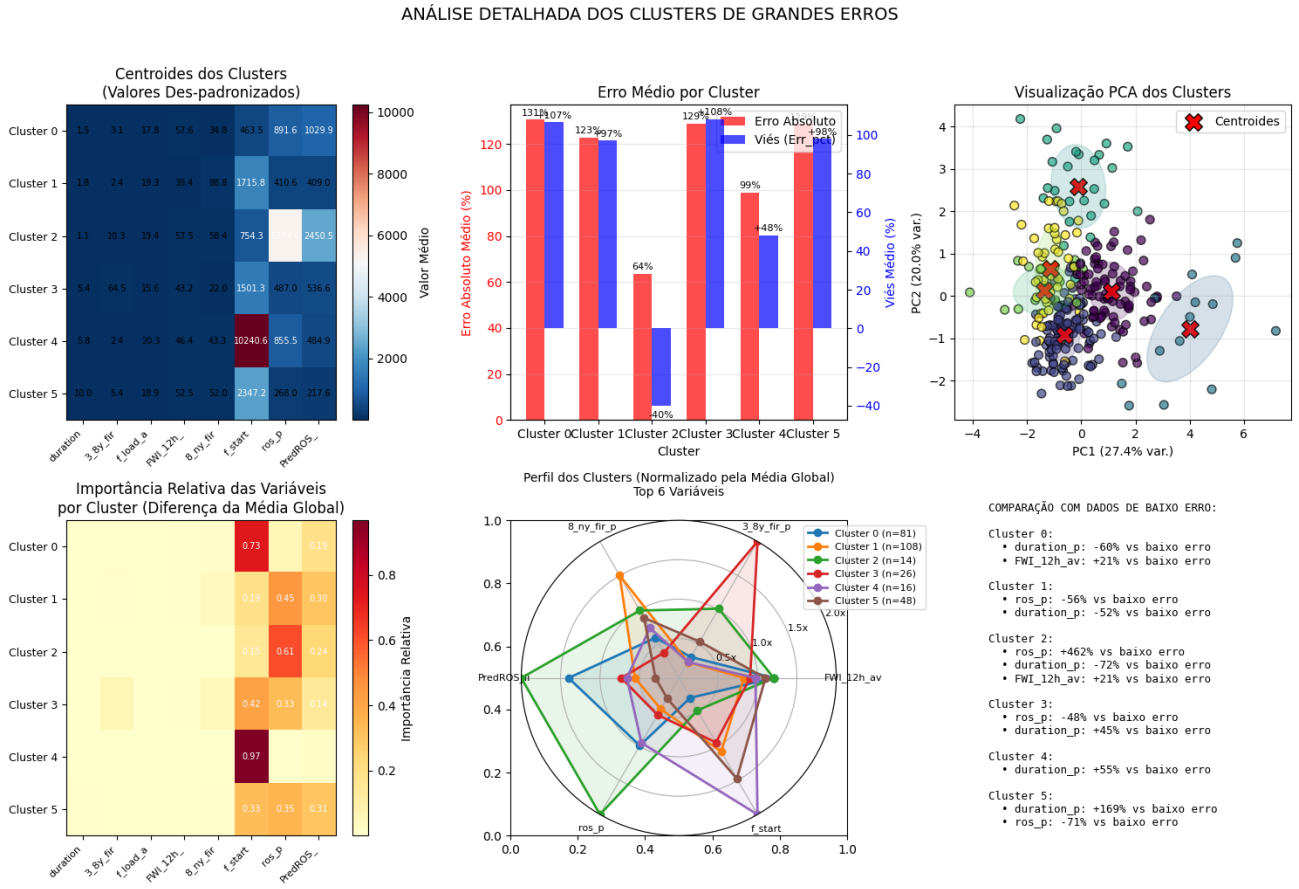


Figure 24 - Detailed analysis of clusters for big errors of Complex model.

To further understand model behavior, a sensitivity analysis was conducted examining the 90th percentile (P90) of observed versus predicted ROS across different quantile ranges of each predictor variable (Figure 23). The analysis reveals that while the model captures the general directional relationships, systematic discrepancies emerge at extreme values. For duration_p, both observed and predicted P90 values decrease with longer fire durations, though the model tends to underpredict at all durations. The 8_ny_fir_p (8-year fire history) variable shows a peaked relationship with maximum ROS around the 40-50 range, which the model approximates but fails to fully capture at the extremes, also underpredicting throughout the whole range. Similarly, for HDW_av (Hot-Dry-Wind index), the model increasingly underestimates as conditions become more severe, with divergence exceeding 1000 m/h at the highest HDW values. This pattern of underestimation under extreme conditions suggests the model's training distribution may not adequately represent tail events. However, sW_100_av (average soil water content between 28 and 100cm depth) presents a quite accurate dynamic of ROS, even if it's still underpredicting.

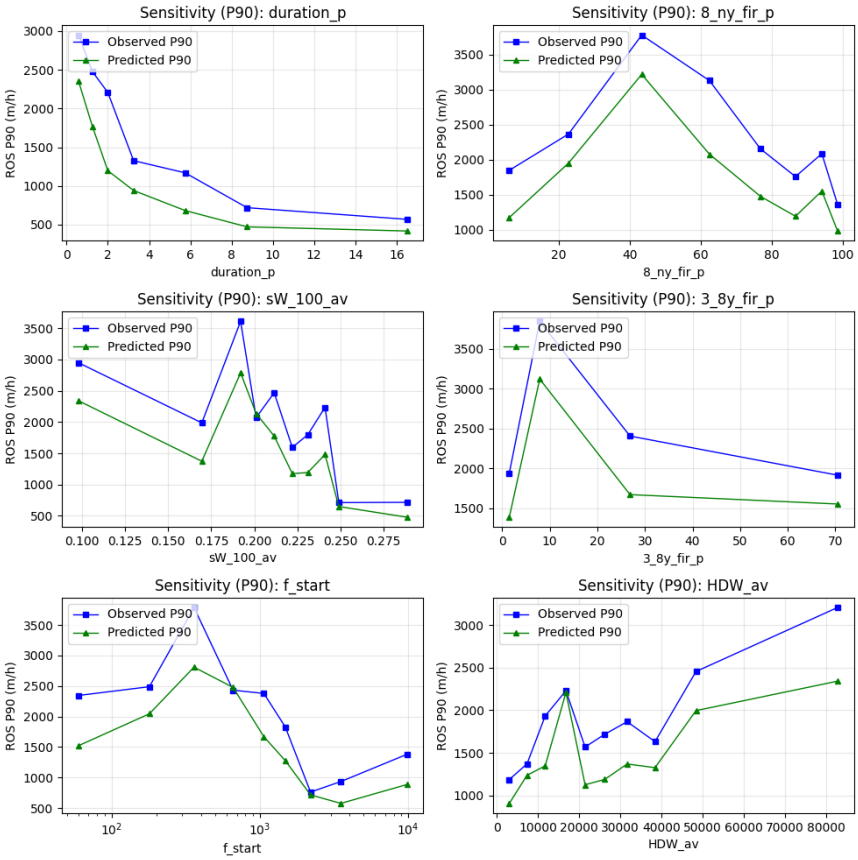


Figure 25 - Sensitivity analysis of input variables across binnes of 90th percentile of Complex model.

The interaction heatmaps show how error varies across feature combinations. Most of the cells don't reach 65% mean absolute error, this means that only some more specific combinations of values have very high error, while most have a more acceptable value. Many cells also have few samples (n < 50), making local estimates somewhat unreliable. High errors in rapid, fuel-driven fires may indicate real prediction difficulty, or just sparse data in those regions.

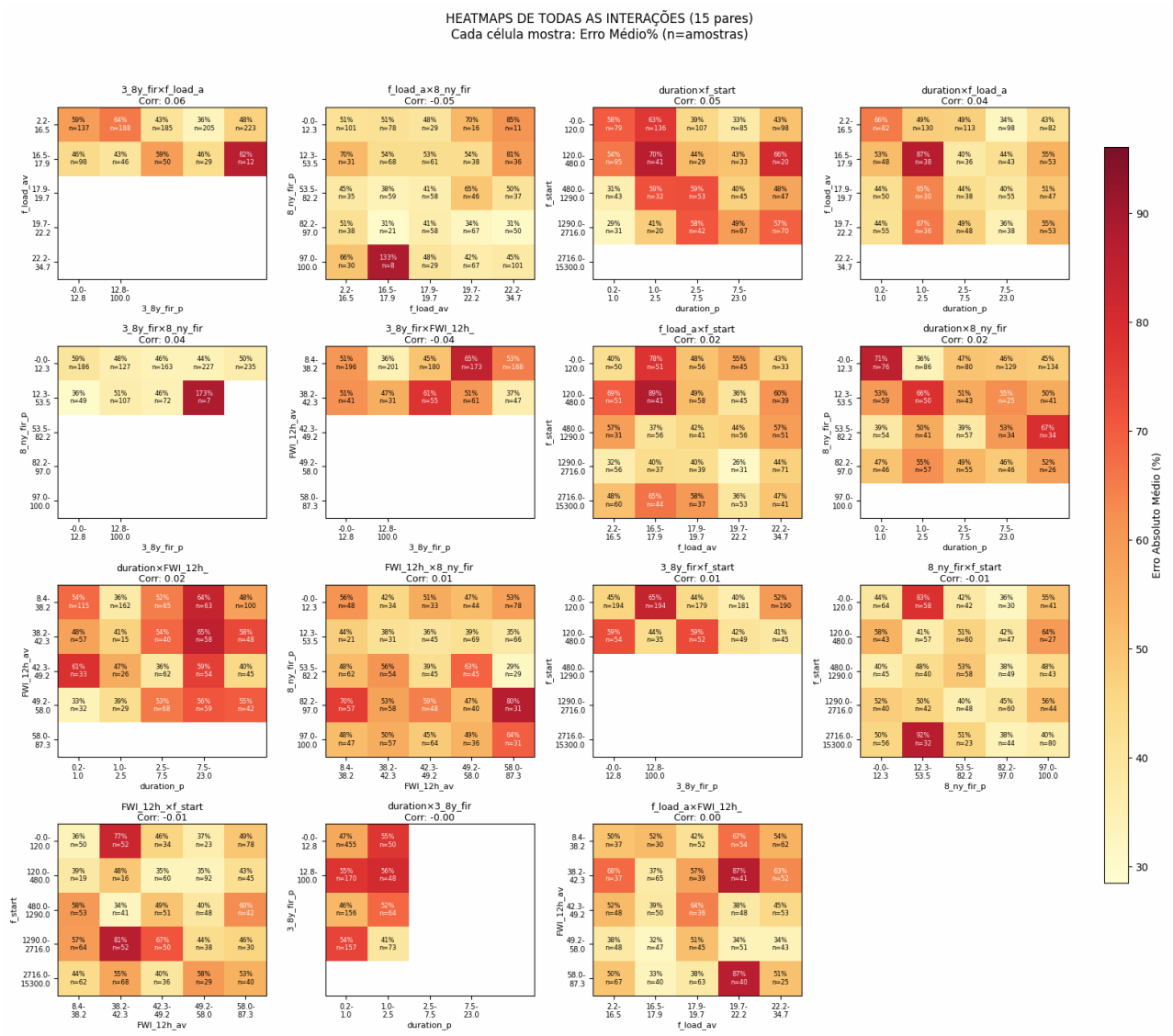


Figure 26 - Interaction heatmaps for Complex model.

We find the results satisfactory considering the limited size of the dataset. As mentioned previously, the dataset is quite small, to develop a model as complex as this one we would need more records, ideally a minimum of around 5000 records in our opinion. Fire spread is influenced by numerous factors that are very complexly related, operating at scales difficult to capture with remote sensing data, including fine-scale fuel arrangement, local topographic effects, and instantaneous wind variability. Since the database is built on fire progression averaged values, the fine detail of fire dynamics is lost to a certain extent, we would like to have included variables such as temporal and spacial maximums, minimums, amplitudes but it would multiplicate the already very high number of input variables, exponentially increasing the computational power needed. We also suspect that some variables may have been not faithfully represented by reanalysis models which may skew the results. An R^2 of 0.53 indicates the model captures roughly half of the variance in rate of spread, even so, the model achieves acceptable performance, specially on medium sized fire rate of spread ranges.

For future work, the model should be revised and its error analysis should guide necessary modifications. We believe there could still be some room for improvement after careful reevaluation of used records, feature engineering, and model complexity tuning. This being said, it won't improve significantly until more fire events are registered in the database.

4.9 Linear Model

The 7 linear model's input variables, obtained from the elbow method based on forward feature selection, can be found on Table 5.

Variable name	Variable full name	Variable description
duration_p	Duration associated with ros_p	The time elapsed between two consecutive polygons.
3_8y_fir_p	Percentage of area burned between 3 and 8 years before	The percentage of area that was burned between 3 and 8 years before the progression's wildfire year.
HDW_av	Average Hot Dry Windy	Average index representing the potential for large fire growth under hot, dry, and windy conditions, Calculated temporal average from temperature, humidity, and wind.
wv_850_av	Average Horizontal Wind Speed at 850 hPa	Average Horizontal Wind Speed at the 850hPa pressure level
Cape_av_log	Average convective available potential energy.	Atmospheric instability and potential for thunderstorm uplift.
gT_8_7_av	Average Temperature Gradient between the 700hPa and 500hPa pressure levels	Average rate of temperature change between the 850hPa and the 700hPa pressure levels
DC_12h_av_log	Average Drought Code	Average moisture content of deep, compact organic layers at noon on the day in question

Table 6 - Table of Linear Model input variables.

The Simple model achieved an R^2 of 0.422 in log-transformed space using seven selected variables. In the original scale, MAE is 500.0 m/h and RMSE is 894.8 m/h. The model produces an interpretable equation where each coefficient represents the effect of a standardized predictor on log-transformed rate of spread.

$$\ln(ROS) = 6.2374 + 0.2240\, HDW\_av + 0.1116\, wv\_850\_av - 0.4613\, duration\_p - 0.2202\, Cape\_av\_log$$
$$-0.1744\, gT\_8\_7\_av + 0.1756\, 3\_8y\_fir\_p\_log + 0.2365\, DC\_12h\_av\_log$$

$$ROS = 511.55 \times 1.2513^{HDW\ av} \times 1.1181^{wv\ 850\ av} \times 0.6305^{duration\ p} \times (Cape\ av)^{-0.2202} \times$$
$$0.8399^{gT\ 8\ 7\ av} \times (3\ 8y\ fir\ p)^{0.1756} \times (DC\ 12h\ av)^{0.2365}$$

The coefficients align with expected fire behavior dynamics. Fire duration shows a strong negative relationship (β = -0.4613), as longer fires tend to have lower average spread rates due to their progression defining process being associated with periods of reduced activity and fewer data sources, during the night also. The Hot-Dry-Windy index (β = 0.2240) and wind speed at 850 hPa (β = 0.1116) show positive effects, reflecting the role of atmospheric conditions in driving fire spread. The Drought Code (β = 0.2365) captures fuel moisture deficit. The temperature gradient between 700 hPa and 500 hPa (β = -0.1744) shows a negative effect, suggesting that steeper mid-tropospheric lapse rates, associated with atmospheric instability, may disrupt organized fire spread. CAPE shows a negative coefficient (β =

-0.2202), possibly indicating that convective instability disrupts organized fire spread or correlates with precipitation events.

While the Simple model is less accurate than the Complex model, its transparent structure offers advantages for interpretation and operational use where understanding individual factor contributions is valuable.

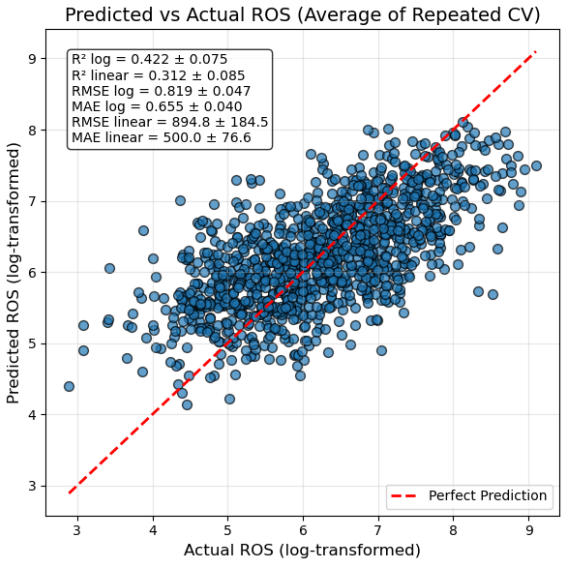


Figure 27 - All 20 Linear Model split's results condensed into one plot in logarithmic space.

Metric	Value
R^2	0.422
MAE *(m/h)*	500.0 ± 76.6
RMSE *(m/h)*	894.8 ± 184.5

Table 7 - Performance metrics of linear model.

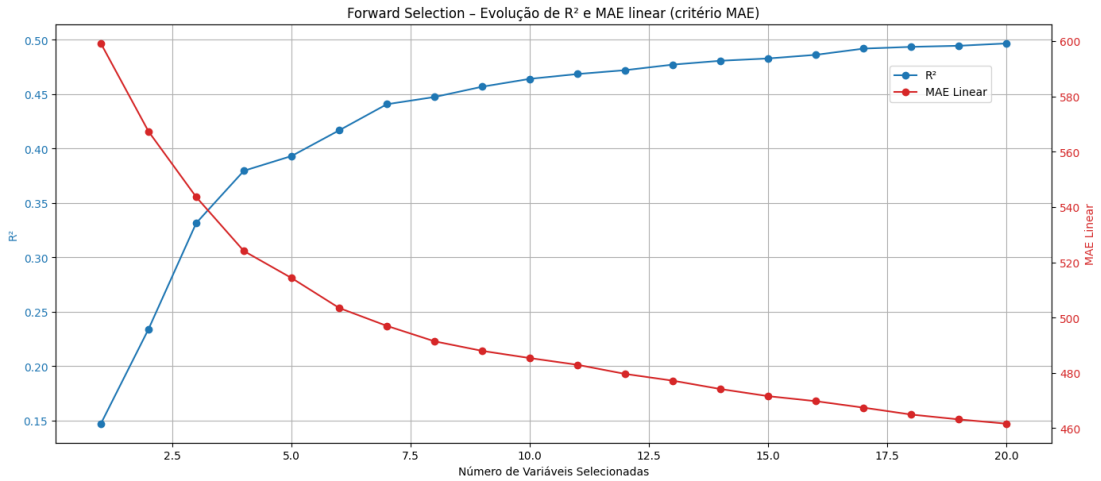


Figure 28 - Plots of R^2 and MAE versus number of input features.

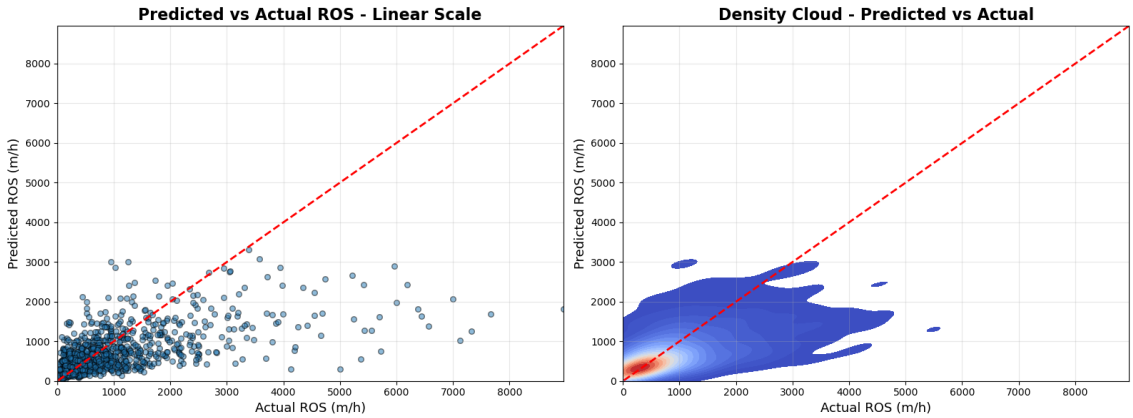


Figure 29 - All 20 Linear Model split's results condensed into one plot in linear space.

Regarding error analysis, we see very similar results compared to the Complex model, the residuals vs fitted values plot shows clear heteroscedasticity, with variance increasing for higher predicted values, and a tendency toward negative residuals (underprediction) as predictions increase. The QQ-plot also shows a significant deviation from normality.

The bias analysis by intensity range is also identical to the Complex model. The model exhibits positive bias (overprediction) for low-intensity fires in the 17-518 m/h range, but here with a higher mean of overestimation of 81.4-203.8 m/h. This transitions to negative bias (underprediction) for fires above 518 m/h, reaching -1548.5 m/h for the highest intensity range (1527-8949 m/h).

The observed vs predicted plot shows a visually bigger cloud of bad predictions compared to the Complex model, which is expected considering the worse performance metrics obtained.

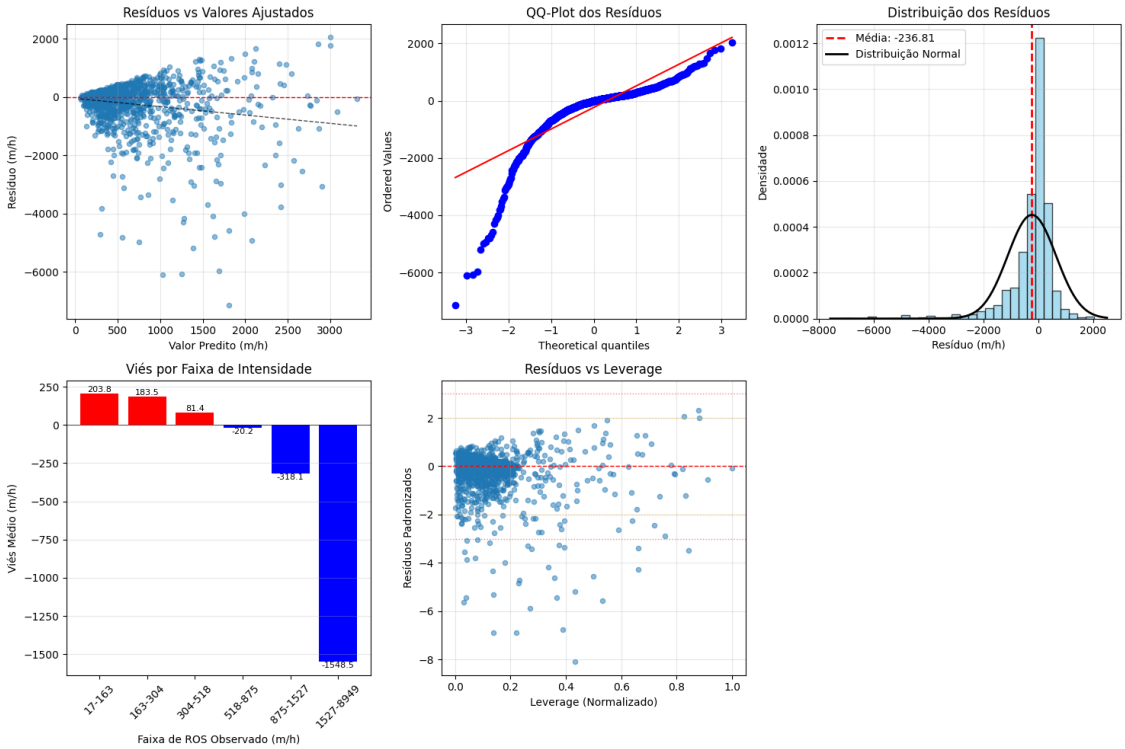
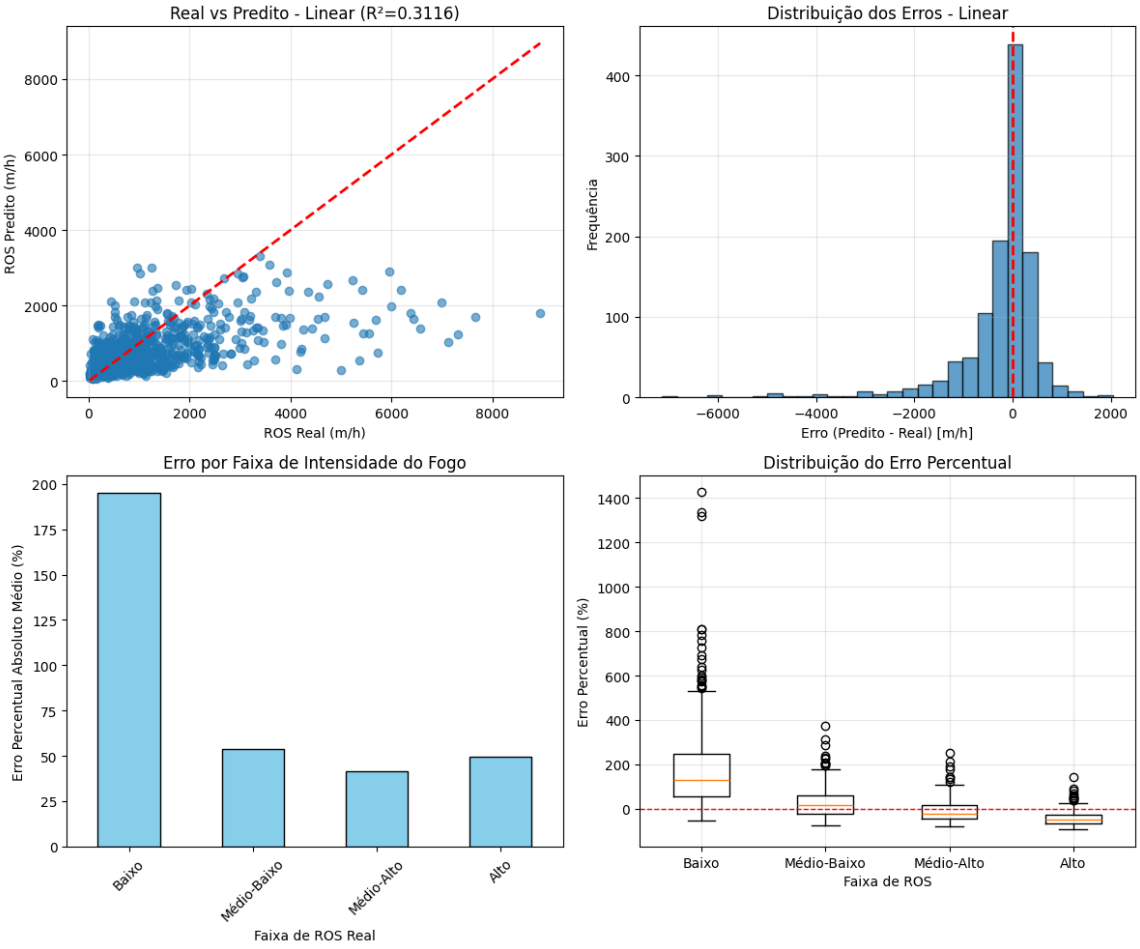


Figure 30 - Residual analysis of Linear model.



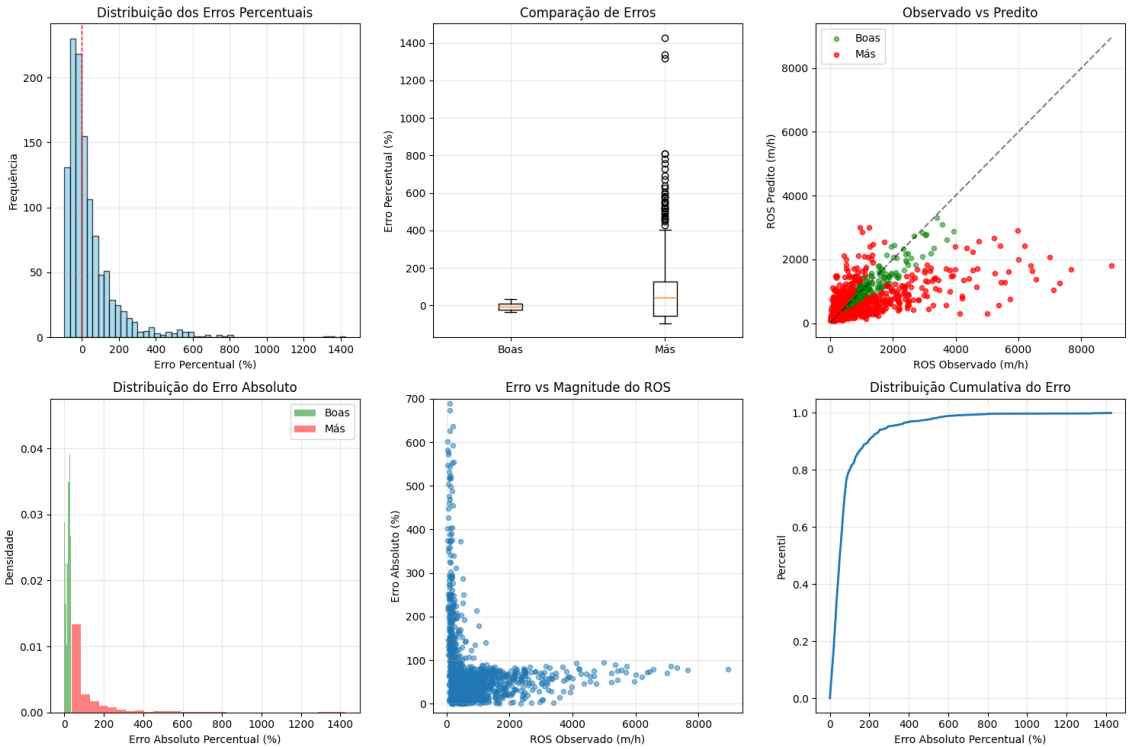


Figure 31 - Error analysis and identification of good and bad predictions of Linear model.

Just like in the Complex model, clusters for high errors were performed, this time four clusters was the appropriate number. Clustering high-error predictions for this model reveals severe overprediction as the dominant failure mode. Clusters 0, 1, and 2 all show bias between +198% and +261%, with predicted ROS roughly 2 to 3× the observed values. These clusters correspond to fires with low observed spread rates (55–83% below low-error cases), suggesting the model consistently overshoots for slow fires. Cluster 3 shows the opposite pattern: observed ROS is 342% higher than low-error cases, but the model underpredicts by 86%. This indicates the model struggles with the amplitude of predictions, overestimating slow fires and underestimating fast ones. However, Cluster 3 contains only 11 samples, limiting confidence in this finding.

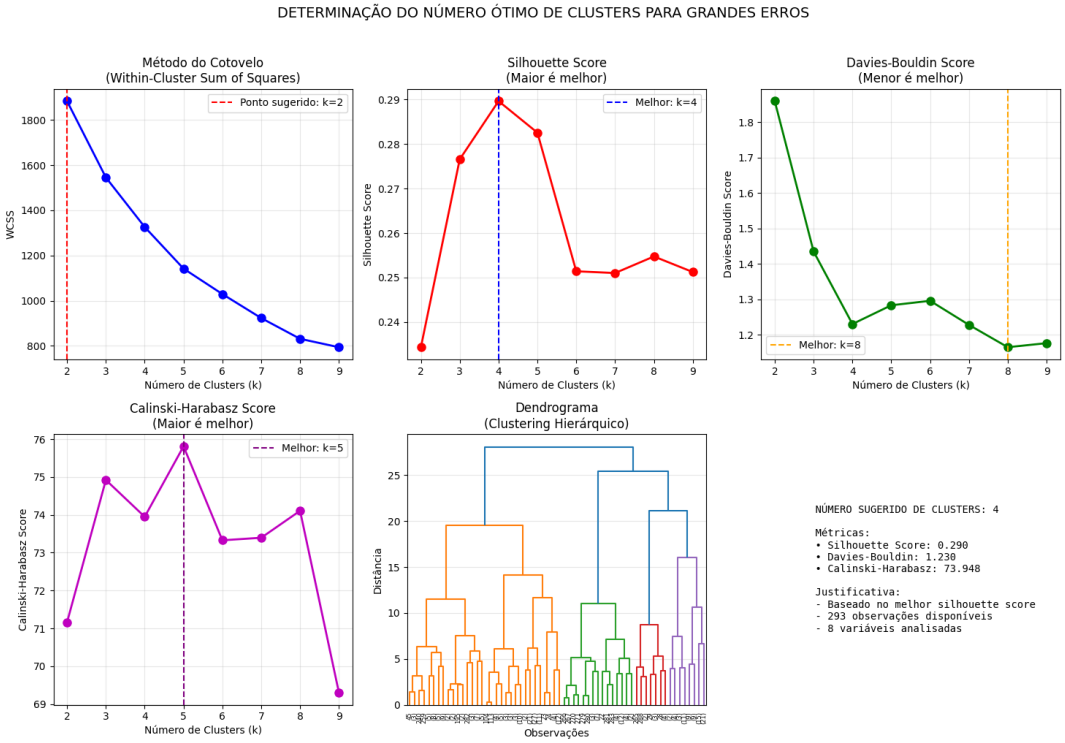


Figure 32 - Determining the best number of clusters for big errors of Linear model.

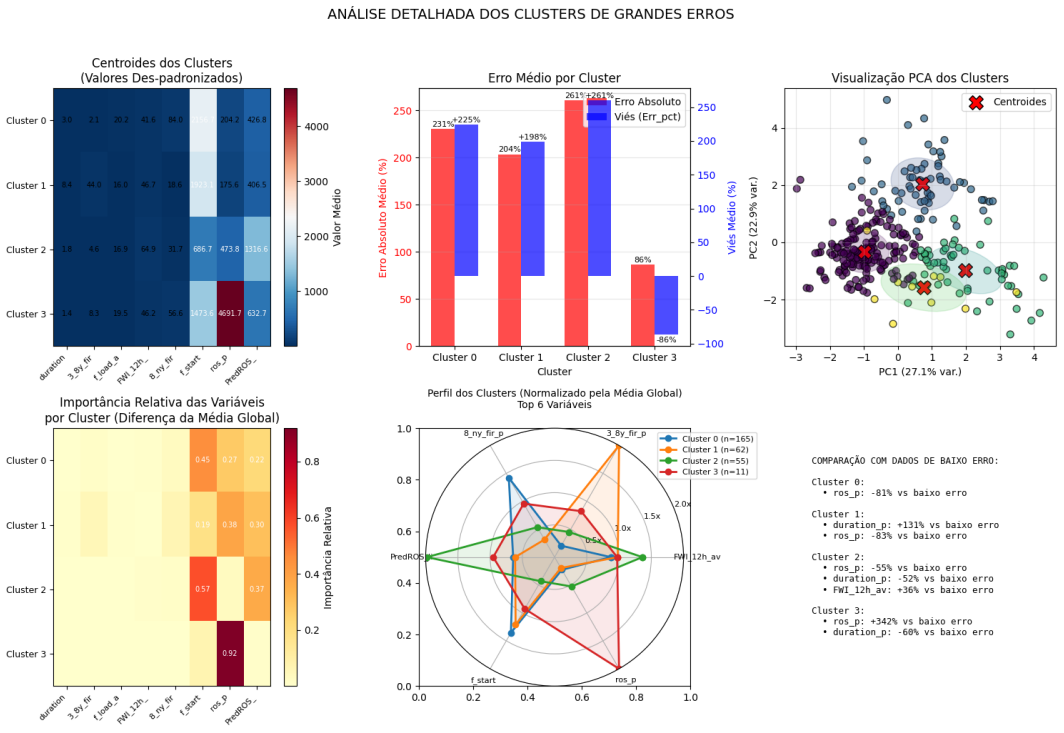


Figure 33 - Detailed analysis of clusters for big errors of Linear model.

A sensitivity analysis was also performed. The model captures general directional trends but systematically underpredicts across nearly all variables. For duration_p, both curves decrease with longer durations, though predictions falls below observations throughout. The 3_8y_fir_p variable shows declining ROS with more fire burn history, and the model follows this trend but consistently undershoots. HDW_av shows the same trend of underpredicting, but it manages to capture the dynamics accurately. For wv_850_av (850 hPa wind velocity), observed ROS increases with wind speed,

but the model underestimates this sensitivity and doesn't fully capture its dynamics. Cape_av the model fails to capture accurately, instead declining monotonically. The gT_8_7_av variable shows a similar story, with underpredictions and small dynamics captured. DC_12h_av (Drought Code) shows the best agreement, both curves increase with drought severity, and the model tracks dynamics reasonably well.

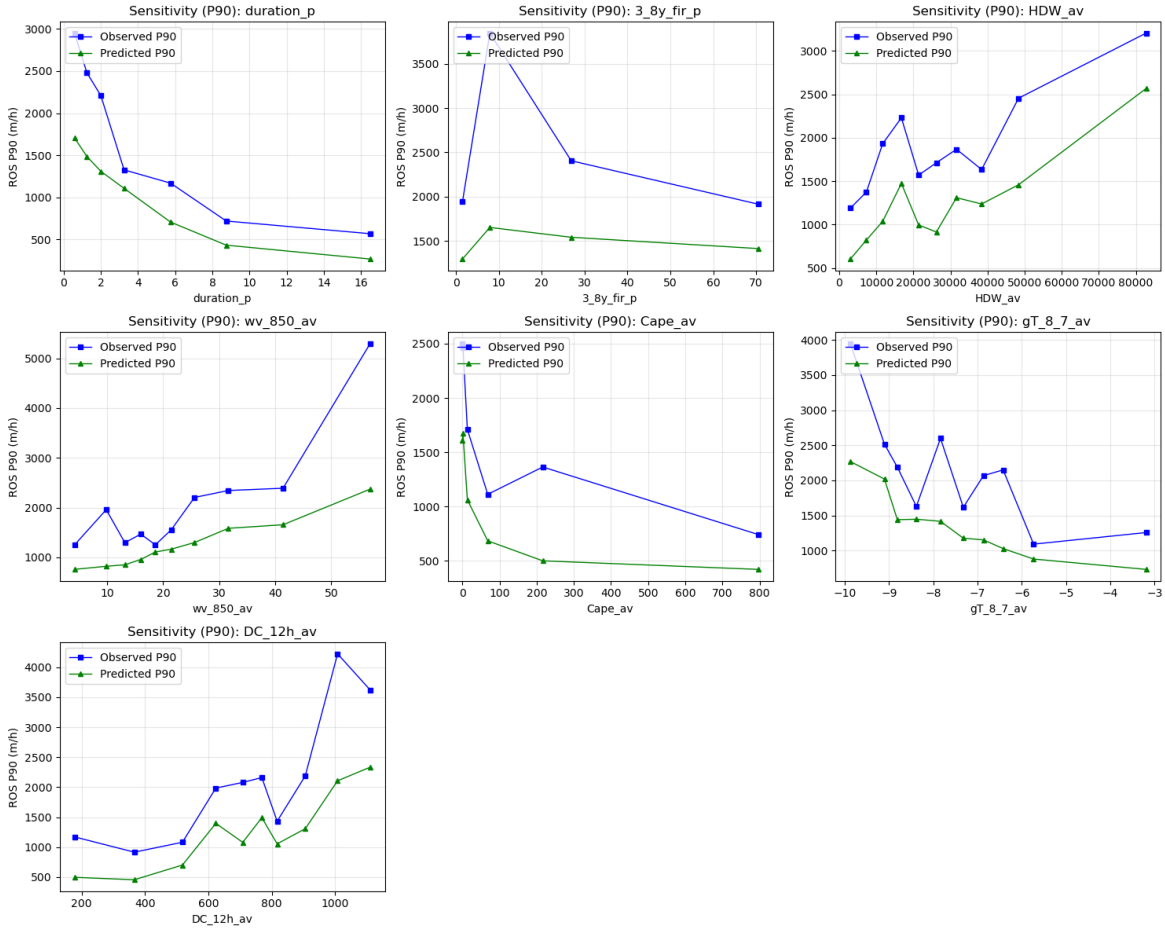


Figure 34 - Sensitivity analysis of input variables across binnes of 90th percentile of Linear model.

The interaction heatmaps, unlike the Complex model, show cells exceeding 200% mean absolute error, and overall we see higher values than in the complex model. Many cells contain few samples (n < 50), making local estimates unreliable. The concentration of very high errors in specific regions, often involving drought conditions (high DC) combined with atmospheric instability indicators, may reflect a special difficulty to predict or simply sparse data in those extreme conditions.

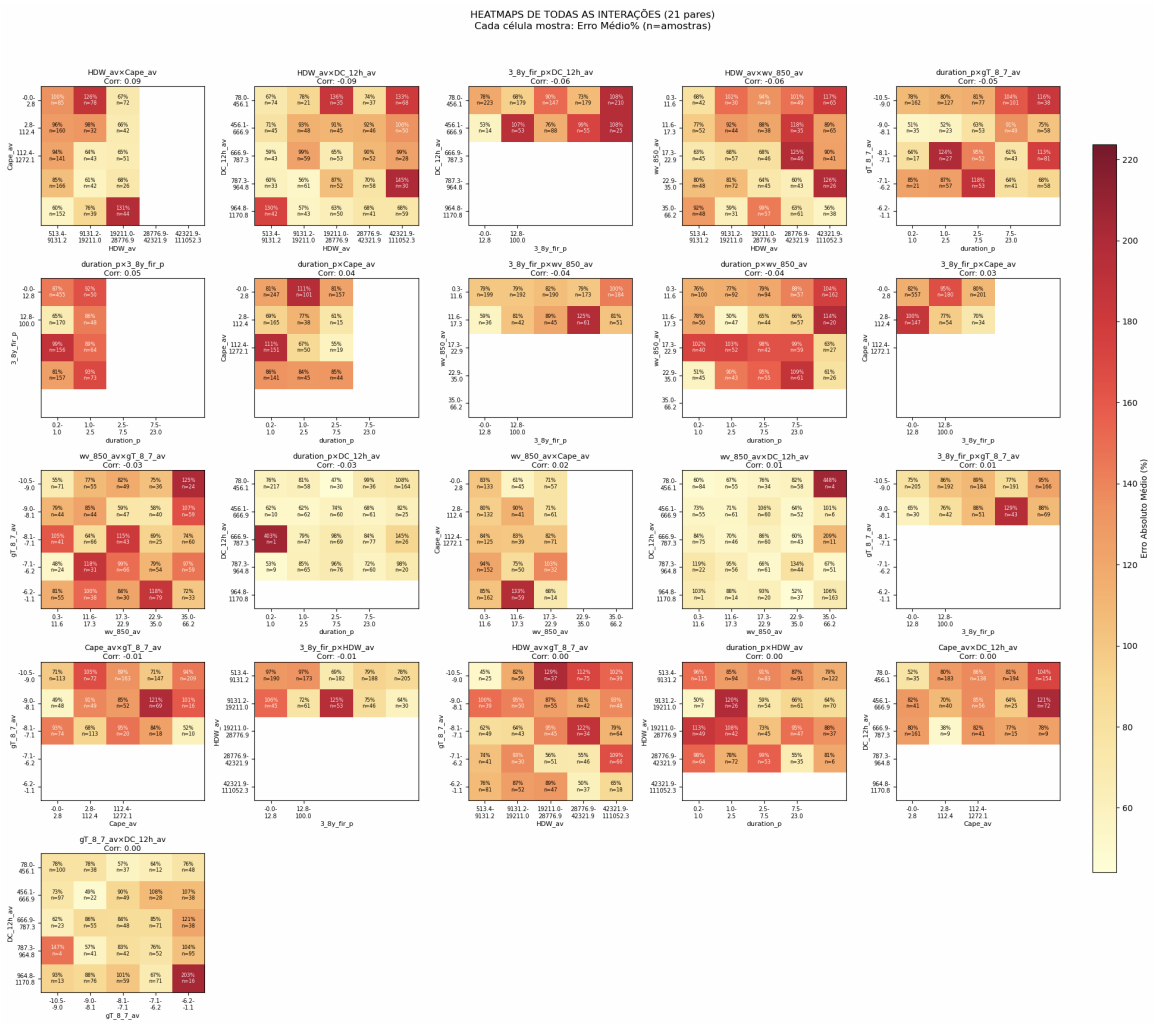


Figure 35 - Interaction heatmaps for Linear model.

We're less satisfied with the model's results than the comple model's. Despite this, it has the benefit of being easily calculated and interpreted, unlike the Complex model. The same concerns can be raised as those in the Complex model.

One important thing is that the model's capacity to capture variable relationships appears limited when trained across the full ROS range. As shown in the error analysis, performance varies substantially between slow and fast fires, with systematic overprediction for low ROS events and underprediction for high ROS. This suggests the underlying relationships between predictors and spread rate may differ across fire intensity regimes, and a single model struggles to represent both simultaneously. So, for future work, two improvements should be considered. First, replacing OLS with Ridge regression could provide modest performance gains. Second, training separate models for distinct ROS ranges, for example, one for fires below median ROS and another for fires above, may allow each model to better capture the specific dynamics of slow-developing versus fast-spreading fires. This separate approach could reduce the bias patterns observed in the current model, where predictions compress toward the mean and fail to represent extreme values at either end of the distribution.

4.10 Deployment

The developed platform provides a solid foundation for operational fire spread prediction. Its modular architecture allows for future modifications to accommodate additional functionality as needs evolve, whether through new input variables, alternative models, or expanded geographic coverage.

With further refinement, the interface could be adapted to better serve firefighters in the field, prioritizing simplicity and rapid interpretation of results under operational conditions. Features such as mobile optimization and streamlined outputs would improve usability during active fire events. This process should be an active debate between the researchers and firefighters to ensure the platform becomes as efficient as possible.

Currently, the main bottleneck is data retrieval from the CDS API, which introduces significant latency, sometimes up to several minutes. Replacing or supplementing this with faster data sources would substantially reduce execution time, making the platform more practical for time-sensitive applications.

For future prediction capability, integration with a weather forecast API is necessary. The current implementation relies on observed meteorological data, limiting its use. Incorporating forecast data would enable predictions of fire behavior hours or days ahead, significantly increasing operational value for fire management planning.

5. Conclusion

This work successfully fulfilled the objectives proposed within the scope of FIRE-HACK, enriching the PT-FireSprd database with a broad and physically consistent set of meteorological and geographical variables, allowing for an unprecedented characterization of the environmental context associated with the spread of large fires in Portugal.

The analysis showed that the rate of spread of fire (ROS) is controlled by an integrated set of factors, notably the temporal persistence of fire behavior, but there's still strong influence of atmospheric and fuel dryness (dfmc and relative humidity at 2m relate the most), wind intensity (wind speed at 10 and 100m and 950hPa), local turbulence (Recirc and Circvar) and vertical structure (LCL and Cape). Composite indices such as Hot-Dry-Windy and FWI are one of the best metrics to define the ROS since they aggregate various dimensions of the ROS drivers.

The results indicate that, there may be different mechanisms that control propagation that influence the ROS in different ways, but these conclusions seem somewhat weak and would need to be validated in a more rigorous way.

From an operational standpoint, the development of two complementary models, a complex, high-precision model based on XGBoost and a simpler, more interpretable linear model, allows us to respond to different usage needs, from real-time decision support to scientific analysis and communication. The models tended to eliminate self-correlation, using variables that correspond to different driver types (time since fire start/ROS lag, HDW, Soil Water content/ Drought Code, area burned, Cape/atmospheric temperature gradient). The XGBoost model proved particularly effective in capturing nonlinear relationships and complex interactions between variables, offering better predictive performance and greater capacity to represent rapid propagation scenarios (MAE = 384m/h), while the linear model stood out for its transparency and ease of physical interpretation, allowing coefficients to be directly related to known mechanisms of fire behavior (MAE = 500m/h).

The implementation of an interactive web platform demonstrates the practical applicability of the results, bringing scientific knowledge closer to end users. Together, this work contributes significantly to improving fire behavior prediction, strengthening the ability to anticipate, plan, and manage forest fire risk in a context of increasing climate complexity.

6. References

1. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

2. Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics. 2020 May 1;36(10):3093-3098. doi: 10.1093/bioinformatics/btaa046. PMID: 31985777; PMCID: PMC7776094.

3. Nested versus non-nested cross-validation. Scikit-learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html

4. HalvingRandomSearchCV. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingRandomSearchCV.html

5. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., . . . Thépaut, J. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803

6. Lundberg & Lee (2017) Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.07874

7. Alexander, M. E., & Cruz, M. G. (6 2013). Limitations on the accuracy of model predictions of wildland fire behaviour: A state-of-the-knowledge overview. The Forestry Chronicle, 89(03), 372–383. doi:10.5558/tfc2013-067

8. Cruz, M. G., & Alexander, M. E. (9 2013). Uncertainty associated with model predictions of surface and crown fire rates of spread. Environmental Modelling & Software, 47, 16–28. doi:10.1016/j.envsoft.2013.04.004

9. McLauchlan, K. K., Higuera, P. E., Miesel, J., Rogers, B. M., Schweitzer, J., Shuman, J. K., … Watts, A. C. (2020). Fire as a fundamental ecological process: Research advances and frontiers. Journal of Ecology, 108(5), 2047–2069. doi:10.1111/1365-2745.13403

10. Vinodkumar, V., Dharssi, I., Yebra, M., & Fox-Hughes, P. (9 2021). Continental-scale prediction of live fuel moisture content using soil moisture information. Agricultural and Forest Meteorology, 307, 108503. doi:10.1016/j.agrformet.2021.108503

11. Anderson, W. R., Cruz, M. G., Fernandes, P. M., McCaw, L., Vega, J. A., Bradstock, R. A., … van Wilgen, B. W. (2015). A generic, empirical-based model for predicting rate of fire spread in shrublands. International Journal of Wildland Fire, 24(4), 443–460. doi:10.1071/wf14130

12. Lamane, H., Mouhir, L., Moussadek, R., Baghdad, B., Kisi, O., & El Bilali, A. (2 2025). Interpreting machine learning models based on SHAP values in predicting suspended sediment concentration. International Journal of Sediment Research, 40(1), 91–107. doi:10.1016/j.ijsrc.2024.10.002

13. Antonini, A. S., Tanzola, J., Asiain, L. ia, Ferracutti, G. R., Castro, S. M., Bjerg, E. A., & Ganuza, M. ia L. (9 2024). Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task. Applied Computing and Geosciences, 23, 100178. doi:10.1016/j.acags.2024.100178

14. Pereira, M. G., Trigo, R. M., da Camara, C. C., Pereira, J. M. C., & Leite, S. M. (3 2005). Synoptic patterns associated with large summer forest fires in Portugal. Agricultural and Forest Meteorology, 129(1–2), 11–25. doi:10.1016/j.agrformet.2004.12.007

15. Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., & Woollings, T. (2022). Atmospheric blocking and weather extremes over the Euro-Atlantic sector -- a review. Weather and Climate Dynamics, 3(1), 305–336. doi:10.5194/wcd-3-305-2022

16. Grünig, M., Seidl, R., & Senf, C. (2022). Increasing aridity causes larger and more severe forest fires across Europe. Global Change Biology, 29(6), 1648–1659. doi:10.1111/gcb.16547

17. Kobziar, L. N. (2014). Fire Ecology, 10(1), 88–91. doi:10.4996/fireecology.1001088

18. Vallejo, V. R., Arianoutsou, M., & Moreira, F. (2012). Fire Ecology and Post-Fire Restoration - Approaches in Southern European Forest Types. In F. Moreira & et al. (Eds), Post-Fire Management and Restoration of Southern European Forests (p. null). Springer.

19. Mather, A. S., & Pereira, J. M. C. (2006). Transição Florestal e Fogo em Portugal. Null, null, 257-null.

20. Fernandes, P. A. M. (2006). Silvicultura Preventiva e Gestão de Combustíveis: Opções e Optimização. null: Universidade de Trás-os-Montes e Alto Douro, Departamento Florestal \textbar Centro de Estudos em Gestão de Ecossistemas.

21. Pereira, J. M. C., Carreiras, J. M. B., Silva, J. M. N., & Vasconcelos, M. J. (2006). Alguns Conceitos Básicos sobre os Fogos Rurais em Portugal. In null (Ed.), null (pp. 133-). null: null.

22. Castro Rego, F., Morgan, P., Fernandes, P., & Hoffman, C. (2021). Fuel and Fire Behavior Description. In Springer Textbooks in Earth Sciences, Geography and Environment (pp. 101–114). doi:10.1007/978-3-030-69815-7_6

23. Scott, A. C., Bowman, D. M. J. S., Bond, W. J., Pyne, S. J., & Alexander, M. E. (2014). Fire management applications of wildland fire behaviour knowledge. In Fire on Earth: An Introduction (First Edition, pp. 373-). null: John Wiley & Sons, Ltd.

24. Fernandes, P., Gonçalves, H., Loureiro, C., Fernandes, M., Costa, T., Cruz, M. G., & Botelho, H. I. (2023). Modelos de Combust\' ivel Florestal para Portugal. Vila Real, Portugal: Universidade de Trás-os-Montes e Alto Douro.

25. Pimont, F., Ruffault, J., Martin, N., & Dupuy, J.-L. (2018, October 19). Why is the Effect of Live Fuel Moisture Content on Fire Rate of Spread Underestimated in Field Experiments in Shrublands? Retrieved from https://doi.org/10.20944/preprints201810.0459.v1

26. DaCamara, C. C., Bento, V. I. A., Nunes, S. I. A., Lemos, G., Soares, P. M. M., & Trigo, R. M. (2024). Impacts of fire prevention strategies in a changing climate: an assessment for Portugal. Environmental Research: Climate, 3(4), 045002. doi:10.1088/2752-5295/ad574f

27. Hardy, J. P. (2001). Ventilation Index and Its Application to Air Pollution Management. null: United States Forest Service.

28. McDonald, J. M., Srock, A. F., & Charney, J. J. (2018). Development and Application of a Hot-Dry-Windy Index (HDW) Climatology. Atmosphere, 9(7), 285. doi:10.3390/atmos9070285

29. Srock, A. F., Charney, J. J., Potter, B. E., & Goodrick, S. L. (2018). The Hot-Dry-Windy Index: A New Fire Weather Index. Atmosphere, 9(7), 279. doi:10.3390/atmos9070279

30. Russo, A., Gouveia, C., Levy, I., Dayan, U., Jerez, S., Mendes, M., & Trigo, R. (6 2016). Coastal recirculation potential affecting air pollutants in Portugal: The role of circulation weather types. Atmospheric Environment, 135, 9–19. doi:10.1016/j.atmosenv.2016.03.039

31. Butler, B., Quarles, S., Standohar-Alfano, C., Morrison, M., Jimenez, D., Sopko, P., … Page, W. (2019). Exploring fire response to high wind speeds: fire rate of spread, energy release and flame residence time from fires burned in pine needle beds under winds up to 27 m s\textminus1. International Journal of Wildland Fire, 29(1), 81–92. doi:10.1071/wf18216

32. Benali, A., Guiomar, N., Gonçalves, H., Mota, B., Silva, F., Fernandes, P. M., … Sá, A. C. L. (2023). The Portuguese Large Wildfire Spread database (PT-FireSprd). Earth System Science Data, 15(8), 3791–3818. doi:10.5194/essd-15-3791-2023

7. Annexes

Annex A

Adapted ID	Adapted Category	Corresponding Codes in the original Land Use Maps
1	Agricultural Areas	2.1. (Temporary crops), 2.2. (Permanent crops), 2.3. (Heterogenous Agricultural Areas), 2.4. (Protected agriculture and nurseries).
2	Pastures	3.1. (Improved pastures and spontaneous pastures), 4.1. (Agroforestry areas), 4.2. (Silvopastoral areas).
3	Deciduous Forests	5.1.1.1 (Cork oak forests), 5.1.1.2 (Holm oak forests), 5.1.1.3 (Other oak forests), 5.1.1.4 (Chestnut forests), 5.1.1.5 (Carob forests), 5.1.1.5 (Other dociduous forests).
4	Eucalyptus Forests	5.1.1.6 (Eucalyptus forests).
5	Invasive Species Forests	5.1.1.7 (Acacia forests).
6	Coniferous Forests	5.1.2.1 (Maritime pine forests), 5.1.2.2 (Stone pine forests), 5.1.2.3 (Other coniferous forests).
7	Shrub	6.1. (Scrubland)
8	Sparse Vegetation	7.1.3 (Sparse vegetation)
9	No Vegetation	1.1. (Residential areas), 1.2. (Areas for economic activities), 1.3. (Facilities), 1.4. (Infrastructure), 1.5. (Transportation), 1.6. (Areas of exploration of geological resources), 1.7. (Vacant lots without construction and areas under construction), 1.8. (Green spaces), 7.1.1 (Beaches, dunes, and sands) 7.1.2 (Rocky areas), 8.1. (Wetlands) 9.1. (Inland water bodies), 9.3. (Transitional and coastal water bodies)

Table 7 - Table of the correspondence between the original Land Use Map's key and the Adapted key.

Annex B

Table 8 - Table of the variables of the updated PT-FireSprd database and their description.

Database Attribute Name	Attribute Name	Attribute Description
fid	Fire ID	A unique numerical identifier for each wildfire in the database
fname	Fire Name	The name of the wildfire, e.g. Gouveia_10082015
year	Year of the Fire	The year in which the wildfire occurred
id	Polygon ID	A unique identifier for each fire progression polygon within a specific wildfire
type	Type of spread polygon	Indicates the type of the progression polygon
sdate	Start date and hour	The date and time when the fire began burning the area represented by the polygon
edate	End date and hour	The date and time when the fire finished burning the area represented by the polygon
inidoy	Start day-of-year	The Julian day of the year when the polygon started burning with hours in decimal values
enddoy	End day-of-year	The Julian day of the year when the polygon finished burning with hours in decimal values
source	Source of the polygon's geometry	Provides information on the source of the data. i.e. satelite (sat), forest service (fserv), etc.
zp_link	Numerical link to the respective z ploygon	Numerical link between an ignition or active flaming zone (z) polygon and a wildfire progression (p) polygon
burn_perio	Progression's burning order	Order of the polygons' burning in a homogenous fire run
area	Burned-area extent	Burned extent, this corresponds to the polygon's area
growth_rat	Fire Growth rate	The rate at which the fire consumed area in the polygon
ros_i	Average rate of spread since the active flame zone	Average rate of spread calculated since the previous ignition/active flaming area
ros_p	Partial rate of spread	Partial rate of spread calculated between two consecutive polygons
spdir_i	Spread direction associated with ros_i	The direction of the forward spread from the previous ignition/active flaming area
spdir_p	Spread direction associated with ros_p	The direction of the forward spread between two consecutive polygons
int_i	Fireline intensity associated with ros_i	Fireline intensity since the previous ignition/active flaming area
int_p	Fireline intensity associated with ros_p	Fireline intensity between two consecutive polygons
duration_i	Duration associated with ros_i	The time elapsed from the previous ignition/active flaming area
duration_p	Duration associated with ros_p	The time elapsed between two consecutive polygons
qc	Confidence flag	An empirical score (1-5) indicating confidence in the reconstructed progression of the wildfire
elev_av	Average elevation	The average elevation within the polygon.
aspect_av	Average aspect	The average compass direction the slope faces in the polygon
landform	Dominant Landform	The most common ALOS terrain classification in the polygon
land_use	Dominant Land Use Class	The most prominant classification of the land use in the polygon
land_use_d	Dominant Land Use Class Name	The corresponding description of the Dominant Land Use, land_use.

1_3y_fir_p	Percentage of area burned less than 3 years before	The percentage of area that was burned less than 3 years before the progression's wildfire year.
3_8y_fir_p	Percentage of area burned between 3 and 8 years before	The percentage of area that was burned between 3 and 8 years before the progression's wildfire year.
8_ny_fir_p	Percentage of area burned more than 8 years before	The percentage of area that was burned more than 8 years before the progression's wildfire year.
fuel_model	Dominant Fuel model	The most common fuel model in the polygon, based on Paulo Fernandes' fuel model classification for Continental Portugal.
f_load_av	Average fuel load	The average amount of burnable plant material based on each fuel model's proportions
sW_1m_av	Average soil water content until 1m depth	Average soil water content until 1m depth
sW_3m_av	Average soil water content until 3m depth	Average soil water content until 2.97m depth
sW_7_av	Average soil water content between 0 and 7cm depth	Average soil water content between 0 and 7cm depth
sW_28_av	Average soil water content between 7 and 28cm depth	Average soil water content between 7 and 28cm depth
sW_100_av	Average soil water content between 28 and 100cm depth	Average soil water content between 28 and 100cm depth
sW_289_av	Average soil water content between 100 and 289cm depth	Average soil water content between 100 and 289cm depth
t_2m_C_av	Average temperature at 2 meters	Average air temperature registered at 2 meters above ground.
d_2m_C_av	Average dew point at 2 meters	Average measurement of the air's moisture content 2 meters above ground.
rh_2m_av	Average relative humidity at 2 m	Average relative humidity registered at 2 meters above ground.
VPD_Pa_av	Average vapour pressure deficit	Average difference between the amount of moisture in the air and the maximum moisture the air can hold at a given temperature.
sP_hPa_av	Average pressure at ground level	The average ground pressure value registered at ground level.
gp_m2s2_av	Average Geopoential	Average potential of the Earth's gravity field
dfmc_av	Average dead fuel moisture content	Average moisture content of dead fuels.
HDW_av	Average Hot Dry Windy	Average index representing the potential for large fire growth under hot, dry, and windy conditions, Calculated temporal average from temperature, humidity, and wind.
Haines_av	Average Haines Index	Average Haines index for the potential for large fire growth due to dry, unstable air
FWI_12h_av	Average Fire Weather Index	Average FWI, the composite index that summarizes the overall fire danger at noon on the day in question
DC_12h_av	Average Drought Code	Average moisture content of deep, compact organic layers at noon on the day in question
FFMC_12h_a	Average Fine Fuel Moisture Code	Average moisture content of surface organic materials, such as grasses, needles, and small twigs at noon on the day in question
wv10_kh_av	Average wind velocity at 10 m	Average wind velocity registered at 10 meters above ground.
wdir10_av	Average wind direction at 10 m	Average wind direction registered at 10 meters above ground.
wv100_k_av	Average wind velocity at 100 m	Average wind velocity registered at 100 meters above ground.
wdir100_av	Average wind direction at 100 m	Average wind direction registered at 100 meters above ground.
Recirc	Costal Air recirculation	Describes the tendency for air masses or wind patterns to circulate back towards the area.
CircVar	Wind direction Circular variance	A statistical measure of wind direction variability at 10m to quantify how much the wind directions deviate from the mean direction in a time interval
t_950_av	Average Air Temperature at 950 hPa	Average Air Temperature at the 950hPa pressure level

t_850_av	Average Air Temperature at 850 hPa	Average Air Temperature at the 850hPa pressure level
t_700_av	Average Air Temperature at 700 hPa	Average Air Temperature at the 700hPa pressure level
t_500_av	Average Air Temperature at 500 hPa	Average Air Temperature at the 500hPa pressure level
t_300_av	Average Air Temperature at 300 hPa	Average Air Temperature at the 300hPa pressure level
rh_950_av	Average Relative Humidity at 950 hPa	Average Relative Humidity at the 950hPa pressure level
rh_850_av	Average Relative Humidity at 850 hPa	Average Relative Humidity at the 850hPa pressure level
rh_700_av	Average Relative Humidity at 700 hPa	Average Relative Humidity at the 700hPa pressure level
rh_500_av	Average Relative Humidity at 500 hPa	Average Relative Humidity at the 500hPa pressure level
rh_300_av	Average Relative Humidity at 300 hPa	Average Relative Humidity at the 300hPa pressure level
wv_950_av	Average Horizontal Wind Speed at 950 hPa	Average Horizontal Wind Speed at the 950hPa pressure level
wv_850_av	Average Horizontal Wind Speed at 850 hPa	Average Horizontal Wind Speed at the 850hPa pressure level
wv_700_av	Average Horizontal Wind Speed at 700 hPa	Average Horizontal Wind Speed at the 700hPa pressure level
wv_500_av	Average Horizontal Wind Speed at 500 hPa	Average Horizontal Wind Speed at the 500hPa pressure level
wv_300_av	Average Horizontal Wind Speed at 300 hPa	Average Horizontal Wind Speed at the 300hPa pressure level
wdi_950_av	Average Horizontal Wind Direction at 950 hPa	Average Horizontal Wind Direction at the 950hPa pressure level
wdi_850_av	Average Horizontal Wind Direction at 850 hPa	Average Horizontal Wind Direction at the 850hPa pressure level
wdi_700_av	Average Horizontal Wind Direction at 700 hPa	Average Horizontal Wind Direction at the 700hPa pressure level
wdi_500_av	Average Horizontal Wind Direction at 500 hPa	Average Horizontal Wind Direction at the 500hPa pressure level
wdi_300_av	Average Horizontal Wind Direction at 300 hPa	Average Horizontal Wind Direction at the 300hPa pressure level
vwv_950_av	Average Vertical Wind Speed at 950 hPa	Average Vertical Wind Speed at the 950hPa pressure level
vwv_850_av	Average Vertical Wind Speed at 850 hPa	Average Vertical Wind Speed at the 850hPa pressure level
vwv_700_av	Average Vertical Wind Speed at 700 hPa	Average Vertical Wind Speed at the 700hPa pressure level
vwv_500_av	Average Vertical Wind Speed at 500 hPa	Average Vertical Wind Speed at the 500hPa pressure level
vwv_300_av	Average Vertical Wind Speed at 300 hPa	Average Vertical Wind Speed at the 300hPa pressure level
gp_950_av	Average Geopotential at 950 hPa	Average Geopotential at the 950hPa pressure level
gp_850_av	Average Geopotential at 850 hPa	Average Geopotential at the 850hPa pressure level
gp_700_av	Average Geopotential at 700 hPa	Average Geopotential at the 700hPa pressure level
gp_500_av	Average Geopotential at 500 hPa	Average Geopotential at the 500hPa pressure level
gp_300_av	Average Geopotential at 300 hPa	Average Geopotential at the 300hPa pressure level
gT_s_9_av	Average Temperature Gradient between the surface layer and 950hPa pressure level	Average rate of temperature change between the surface (2m) and the 950hPa pressure level
gT_9_8_av	Average Temperature Gradient between the 850hPa and 700hPa pressure levels	Average rate of temperature change between the 950hPa and the 850hPa pressure levels
gT_8_7_av	Average Temperature Gradient between the 700hPa and 500hPa pressure levels	Average rate of temperature change between the 850hPa and the 700hPa pressure levels
gT_7_5_av	Average Temperature Gradient between the 500hPa and 300hPa pressure levels	Average rate of temperature change between the 700hPa and the 500hPa pressure levels
gT_5_3_av	Average Temperature Gradient between the 950hPa and 850hPa pressure levels	Average rate of temperature change between the 500hPa and the 300hPa pressure levels

wSv_9_av	Average Wind Shear velocity between 10 m and 950hPa pressure level	Average wind velocity change between 10m and 950hPa (aproximtely 500m above ground)
wSdir_9_av	Average Wind Shear direction between 10 m and 950hPa pressure level	Average wind direction change between 10m and 950hPa (aproximtely 500m above ground)
wSv_7_av	Average Wind Shear velocity between 10 m and 700hPa pressure level	Average wind velocity change between 10m and 700hPa (aproximtely 3km above ground)
wSdir_7_av	Average Wind Shear direction between 10 m and 700hPa pressure level	Average wind direction change between 10m and 700hPa (aproximtely 3km above ground)
wSv_5_av	Average Wind Shear velocity between 10 m and 500hPa pressure level	Average wind velocity change between 10m and 500hPa (aproximtely 6km above ground)
wSdir_5_av	Average Wind Shear direction between 10 m and 500hPa pressure level	Average wind direction change between 10m and 500hPa (aproximtely 6km above ground)
wSv_1_av	Average Wind Shear velocity between 10 m and 100m	Average wind velocity change between 10m and 100 m
wSdir_1_av	Average Wind Shear direction between 10 m and 100 m	Average wind direction change between 10m and 100 m
CBH_m_av	Average cloud base height	Average height of the cloud base above ground.
HigCC_p_av	Average cloud cover above approximately 6km	Average cloud cover percentage on model levels with a pressure less than 0.45 times the surface pressure.
LowCC_p_av	Average cloud cover below approximately 2km	Average cloud cover percentage on model levels with a pressure greater than 0.8 times the surface pressure
MidCC_p_av	Average cloud cover between approximately 2km and 6km	Average cloud cover percentage on model levels with a pressure between 0.45 and 0.8 times the surface pressure
TotCC_p_av	Average total cloud cover	Average total cloud cover percentage
Cape_av	Average convective available potential energy	Atmospheric instability and potential for thunderstorm uplift.
Cin_av	Average convective inhibition	Average energy that suppresses upward air motion.
BLH_m_av	Average boundary layer height	Height of the atmosphere layer mixed by the Earth's surface.
BLH_m_rt	Rate of change of the Boundary Layer Height	Rate of change of the Boundary Layer Height
LCL_hPa_av	Average lifting condensation level	Average pressure level at which an air parcel becomes saturated.
LCL_m_av	Average lifting condensation height	Average height at which an air parcel becomes saturated.
LFC_hPa_av	Average Level of Free Convection	Average height where a rising air parcel becomes warmer than the surrounding air and begins to rise freely.
CCL_hPa_av	Average Convective Condensation Level	Average height where air, when heated from the surface, reaches saturation and cloud formation begins
EL_m_av	Average Equilibrium Level	Average height at which a rising parcel of air is at the same temperature as its environment.
LiftIdx_av	Average Lifted Index	Average difference between ambient temperature and a lifted parcel’s temperature at 500hPa.
VentIdx_av	Average ventilation rate	Average ventilation rate. Product of wind speed and mixing height, indicating potential for fire growth
CMLG_av	Average Convective Mixing Layer Gap	Average distance between the Boundary Layer Height and the Lifting Condensation Level
ros_p_lg1	Average rate of spread 1h before	Average rate of spread since the active flame zone 1 hour before the start of the progression
f_start	Time since the begining of the wildfire	Minutes elapsed since the first ignition of this fire
fire_rank	Progression's Rate of Spread Temporal Rank	Index indicating whether the progression is the the one that has the highest ROS in that period of a given wildfire.

Annex C

Descriptive statistics of principal variables.

Variable	Description	Mean	Median	Standard Deviation	Min	Max	r(ROS)
wv_850_av	Wind speed 850 hPa (m/s)	23.5	19.9	14.3	0.3	66.2	0.38
wv100_k_av	Wind speed at 100 m (km/h)	18.0	16.6	9.3	0.7	55.2	0.34
FWI_12h_av	Fire Weather Index (12h average)	47.7	45.5	12.4	8.4	87.3	0.32
DC_12h_av	Drought Code	700.2	750.7	267.1	78.0	1170.8	0.31
HDW_av	Hot-Dry-Windy index	27268	23623	19257	513	111052	0.29
t_2m_C_av	Temperature at 2 m (°C)	26.6	26.1	5.5	13.9	39.1	0.08
rh_2m_av	Relative humidity at 2 m (%)	40.2	37.1	16.9	8.9	90.8	-0.25
VPD_Pa_av	Vapor Pressure Deficit (Pa)	2343	2113	1234	156	6363	0.15
dfmc_av	Dead fuel moisture content (%)	9.9	9.3	3.2	4.2	20.0	-0.23
duration_p	Progression duration (h)	3.7	1.5	3.9	0.2	23.0	-0.31
3_8y_fir_p	Area burned 3–8 years prior (%)	10.1	0.0	21.6	0.0	100.0	0.00
8_ny_fir_p	Area burned >8 years prior (%)	59.1	71.4	37.3	0.0	100.0	−0.06
sW_100_av	Soil water 0–100 cm (m^3/m^3)	0.21	0.22	0.04	0.05	0.33	−0.24
f_start	Fire start time (min from midnight)	1593	840	2130	0	15300	−0.19
Cape_av	CAPE (J/kg)	92.3	0.01	207.6	0.0	1272.1	−0.19
gT_8_7_av	Temp. gradient 800–700 hPa (°C/km)	−7.46	−7.57	1.57	−10.5	−1.1	−0.16

Table 9 - Table of Descriptive Statistics of Key Environmental Variables.

Statistic	N	Mean (m/h)	Median (m/h)	SD (m/h)	Min (m/h)	Max (m/h)	Skewness	Kurtosis	CV (%)
Value	1,173	900.9	518.2	1,101.2	17.0	8,949.2	2.85	10.62	122.2

Table 10 -Table with distribution statistics of Rate of Spread (*ros_p*)

List of Figures

List of Tables