INSTITUTO SUPERIOR DE AGRONOMIA
Universidade de Lisboa

Detecting and Assessing Pollution Events from Wildfires Using Remote Sensing and Meteorological Data

A Data Science Approach

Sofia Margarida Matias Rodrigues

Dissertação para a obtenção do Grau de Mestre em

Ciências dos Dados em Agricultura, Alimentação, Floresta e Ambiente

Orientador(es): Ana Russo
Célia Gouveia

(Versão provisória)

2025

ULISBOA UNIVERSIDADE DE LISBOA

Acknowledgments

Abstrato

Os incêndios florestais contribuem significativamente para a poluição do ar ao libertarem matéria particulada (PM) e gases tóxicos na atmosfera, com as alterações climáticas a preverem um aumento na atividade dos incêndios florestais e na propagação do fumo, agravando os riscos para a saúde. O trabalho a ser desenvolvido será apoiado por um quadro orientado por dados para monitorizar e avaliar a poluição do ar causada por incêndios florestais, uma questão urgente de saúde e ambiental que afeta a população global. O trabalho irá analisar os resultados dos modelos atmosféricos atuais, juntamente com indicadores de deteção remota, como o Fire Radiative Power (FRP) e o Fire Radiative Energy (FRE), combinados com dados meteorológicos e aprendizagem automática, para melhorar a deteção de eventos de poluição. O objetivo é analisar os impactos transfronteiriços das emissões de incêndios florestais, avaliar tecnologias de deteção remota (e.g., MODIS, SEVIRI, Sentinel) no monitoramento de incêndios florestais e examinar métodos de ciência de dados para monitorização ambiental. Na secção de dados e metodologia, será descrita a integração de dados meteorológicos e de deteção remota, com modelos de aprendizagem automática, como Random Forest, XGBoost e Redes Neuronais, usados para classificar eventos de poluição e mapear padrões espaço-temporais do fumo. A validação do modelo será realizada através da comparação dos resultados com eventos históricos extremos de incêndios florestais para verificar a sua precisão. O modelo será, então, avaliado pelo seu desempenho preditivo e fornecerá algumas perceções sobre os padrões de dispersão do fumo dos incêndios florestais, identificando fatores-chave que contribuem para os eventos de poluição. Para concluir o trabalho, serão apresentados os destaques do estudo, demonstrando como os dados de deteção remota e meteorológicos podem melhorar o monitoramento da qualidade do ar e apoiar o planeamento de políticas públicas. Serão propostos trabalhos futuros para melhorar as capacidades de monitoramento em tempo real, integrar fontes de dados adicionais e aplicar as descobertas em quadros ambientais e de saúde mais amplos. Esta pesquisa tem o potencial de informar intervenções estratégicas, reforçando ainda mais as ferramentas de tomada de decisão para gerir a poluição provocada por incêndios florestais.

Palavras-Chave: Fogos, Poluição do Ar, Dados Meteorológicos, Deteção Remota, Aprendizagem Automática

RESUMO PT 1200-1500 PALAVRAS

How to build a good abstract:

Importance of the study

Gap in the existing literature

Objective of the study

Method used to conduct the study

Key findings of the study

Implications of the study

Abstract

Wildfires contribute significantly to air pollution by releasing PM and toxic gases into the atmosphere, with climate change projected to increase wildfire activity and the spread of smoke, heightening health risks. The work that will be developed will be supported by a data-driven framework to monitor and assess air pollution from wildfires, a pressing health and environmental issue that affects the global population. The work will analyse current atmospheric models outputs coupled with remote sensing indicators, like Fire Radiative Power (FRP) and Fire Radiative Energy (FRE), combined with meteorological data and machine learning to improve pollution event detection. It aims to look upon transboundary impacts of wildfire emissions, evaluate remote sensing technologies (e.g. MODIS, SEVIRI, Sentinel) in wildfire monitoring, and examine data science methods for environmental monitoring. For the data and methodology section, it will describe the integration of meteorological and remote sensing data, with machine learning models, such as Random Forest, XGBoost, and Neural Networks, used to classify pollution events and track spatial-temporal patterns of smoke. Model validation will be performed by comparing results with historical extreme wildfire events to verify accuracy. Then the model will be evaluated by its predictive performance and have some insights into wildfire smoke dispersion patterns, identifying key factors contributing to pollution events. To conclude the work, the highlights of the study will be shown, demonstrating how remote sensing and meteorological data can improve air quality monitoring and support policy planning. Future work will be proposed to enhance real-time monitoring capabilities, integrate additional data sources, and apply findings within broader environmental and health frameworks. This research has the potential to inform strategic interventions, further strengthening decision-making tools for managing wildfire-driven pollution.

Key Words: Wildfires, Air Pollution, Meteorological Data, Remote Sensing, Machine Learning

Table of Contents

List of Figures

List of Tables

List of Abbreviations

IPMA – Instituto Português do Mar e da Atmosfera

WRF-Chem – Weather Research and Forecasting Model with Chemistry

WRF – Weather Research and Forecasting

SEVIRI – Spinning Enhanced Visible and Infrared Imager

MSG – Meteosat Second Generation

WHO – World Health Organization

WMO – World Meteorological Organization

FRP – Fire Radiative Power

FRE – Fire Radiative Energy

CDS – Climate Data Store

ADS – Atmosphere Data Store

ECMWF – European Centre for Medium-Range Weather Forecasts

ML – Machine Learning

AI – Artificial Intelligence

API – Application Programming Interface

CAMS – Copernicus Atmosphere Monitoring Service

PM – Particulate Matter

IFS – Integrated Forecasting System

EAC4 – ECMWF Atmospheric Composition Reanalysis 4

FIRMS – Fire Information for Resource Management System

MODIS – Moderate Resolution Imaging Spectroradiometer

NetCDF – Network Common Data Form

GADM – Database of Global Administrative Areas

1.Introduction

Wildfires are a derivative of a fire that is uncontrolled that spreads across vegetation, often fuelled by dry conditions, strong winds, and abundant combustible material. They can take large proportions because they are usually situated in rural and forested areas with difficult access. Historically, they could be a beneficial tool to certain natural landscapes by clearing the underbrush and allowing seed release for some species. However, climate change and anthropogenic influence contribute to extreme episodes and in consequence larger burned areas, which causes a major imbalance in the ecosystems and carbon storage (Knorr, Dentener, Lamarque, Jiang, & Arneth, 2017). Wildfires are increasing its frequency, severity and duration, bringing concerns to the health of those affected by its exposure to a mixture of hazardous air pollutants, like particulate matter ($PM_{2.5}$ in specific), nitrogen dioxide (NO_2) and ozone (O_3) (De Sario, Katsouyanni, & Michelozzi, 2013). Even though they are aggravated by meteorological extremes, like droughts, heat waves and high winds, wildfires can also be a source that impacts the climate by releasing large amounts of carbon dioxide (CO_2) and other greenhouse gases into the atmosphere (World Health Organization, 2025). Wildfires are not isolated events; they are increasingly linked to extreme weather conditions. Droughts, heatwaves, and high winds create the perfect conditions for wildfires to ignite and spread, while climate change continues to amplify these extreme events.

Meteorological extremes or extreme weather events are characterized as rare events “at a particular place and time of year, with unusual characteristics in terms of magnitude, location, timing, or extent” (World Meteorological Organization, 2025). Climate change is one of the main reasons for the increase in the impacts of extremes events that take place, and wildfires are directly linked to that, in a sense that often occur alongside or a result of other climate extremes (e.g. heatwaves, droughts or high winds) (AghaKouchak, et al., 2020). The Mediterranean region is particularly susceptible to weather and climate variability, being more prone to larger occurrences of fires, while being directly related to extreme events (Bento, et al., 2023). Beyond fuelling wildfires, extreme weather also worsens air pollution. The combination of prolonged droughts, increased temperatures, and intensified fires releases vast amounts of hazardous pollutants into the atmosphere, affecting air quality and human health on a global scale.

The World Health Organization (WHO) defines air pollution as a contamination by any chemical, physical or biological agent that changes the natural characteristics of the atmosphere, either indoors or outdoors. There is a large amount of air pollution sources, the most common categories being transportation vehicles (that run on combustion motors), industrial facilities and fires (World Health Organization, 2025). Key air pollutants that pose significant public health risks include particulate matter (PM), carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). Data and studies provided by WHO show that 99% of the global population breathes contaminated air, based on the established guideline limits for each pollutant (World Health Organization, 2025).

That being said, there is a concern that wildfires will become more aggravated over the years, leading to further degradation of air quality. So, it is important to develop studies relating these two topics, air pollution and wildfires, to come up with strategies that promotes interventions and initiatives.

Traditional approaches for studying wildfire air pollution emissions, such as emission inventories, ground-based monitoring, satellite observations, and atmospheric dispersion models, have been essential in estimating pollutant emissions and tracking smoke transport (Li, Zhang, Kondragunta, & Csiszar, 2018). However, these methods come with significant caveats, including uncertainties in emission factors, limited spatial and temporal coverage of ground-based sensors, satellite retrieval errors due to cloud cover, and computational constraints in high-resolution atmospheric modelling. Additionally, atmospheric dispersion models, while commonly used to track wildfire smoke, require heavy computational resources and often exhibit discrepancies between predicted and actual pollution concentrations (Schneider, Lee, Santos, & Abbatt, 2021). Alternatively, remote sensing provides near-real-time data with products, which allow for more effective monitoring of wildfire emissions. However, as the availability of large-scale remote sensing and meteorological datasets grows, researchers face increasing challenges in efficiently processing and analysing these vast amounts of data. This presents a key research gap where traditional methods struggle with computational performance and real-time prediction capabilities (Ceamanos, et al., 2023). To address these limitations, the integration of machine learning and data science is essential, enabling more robust pattern recognition, improved predictive accuracy, and efficient data processing. By leveraging data-driven approaches, this study aims to enhance the understanding of wildfire-induced air pollution using data science resources like machine learning.

Commented [SM1]: Maybe refrase it

As a subset of Artificial Intelligence (AI), Machine learning (ML) allows systems to learn from data and improve their performance over time without explicit programming, playing a crucial role in extracting insights, detecting patterns, and automating processes across different topics. While all ML is considered AI, not all AI encompasses machine learning. As data generation grows exponentially, ML helps analyse vast and large datasets to uncover trends and make predictions. It includes supervised, unsupervised, semi-supervised, and reinforcement learning, each suited for different tasks. ML uses various algorithms, such as neural networks for complex data, regression for predictions, clustering for grouping similar data, and decision trees for rule-based predictions. By training on data, ML models identify complex relationships, enabling automatic pattern discovery, predictive analysis and large-scale data processing. This makes it crucial to select the right learning approach based on data structure, resources, and application needs (Chen, 2024).

The study will use data from two different sources that will be described in more detail further. Copernicus is the source for two of the different datasets used – the Climate Data Store (CDS) and the Atmosphere Data Store (ADS). Copernicus, the European Union's Earth observation program, provides free environmental data through satellites, ground sensors, and airborne measurements (Copernicus, 2025). From CDS, the study will use ERA5 hourly data, a reanalysis dataset combining weather models with observations to reconstruct past climate and weather patterns (Copernicus, 2025). ADS will provide CAMS global reanalysis (EAC4) dataset, which is the fourth-generation global reanalysis of atmospheric composition (Copernicus, 2025).

Fire Radiative Power (FRP) is one of the variables used to characterize wildfires, it is obtained "from the radiance at the 4-μm band of satellite sensors and represents the instantaneous radiative energy that is released from actively burning fires". FRP can be used in many ways to provide information about biomass burning, land cover dynamics and hydrological cycles. But for this study, it gives insights about the rate of emissions in relation to the rate of biomass combustion. This allows to estimate trace gas and aerosol emissions or smoke production. The variable can be obtained from multiple polar-orbiting and geostationary satellites and can be obtained on one of the main sources already mentioned, the MODIS (Li, Zhang, Kondragunta, & Csiszar, 2018). Based on its proportionality to the amount of burned biomass, higher values of FRP indicate more severe fires and in consequence larger levels of smoke production, leading to higher emissions of particulate matter and other pollutants (Durao, Alonso, Russo, & Gouveia, 2024). Fire Radiative Power (FRE) is estimated via temporal integration from the measures of FRP (during the lifetime of a fire). By representing the total amount energy release during a fire, it can also provide the total amount of consumed biomass (Instituto de Meteorologia, 2009).

Commented [SM2]: Usar apenas se se usar no futuro

Model selection

Still to write

General and specific objectives

The study has the purpose of detecting and assessing wildfire pollution events by using a combination of remote sensing, atmospheric monitoring data and machine learning approaches, focusing on FRP and FRE outputs as indicators of wildfire-induced pollution. In order to achieve the proposed general goal, four objectives will be pursued: 1) identify key indicators from remote sensing and meteorological data that correlates with pollution events due to wildfires – those can be fire activity parameters, atmospheric pollutant concentrations, and meteorological conditions; 2) develop a machine learning-based model to detect pollution events using data from sources already mentioned; 3) analyse the spatial and temporal impacts of wildfire smoke on air quality in affected areas; and finally, 4) analyse the impact of compound extreme events on wildfire-related pollution events.

Research questions to answer

2.State of the Art

A relevant example of integrating fire activity with air quality analysis is presented by (Moura, et al., 2024), who investigates the impacts of biomass burning in Sinop, a Brazilian city. Their study combines satellite-derived FRP with atmospheric pollutant measurements, meteorological data, and modelling datasets to assess how wildfire events influenced air quality. Statistical analyses revealed that ozone concentrations rose immediately following fires and remained elevated for several days, while particulate matter increased a few days after fire events. FRP showed a strong correlation with ozone levels, confirming its value as an indicator of fire intensity, while temperature exhibited a consistent positive correlation with all pollutants, highlighting its role in pollutant dynamics. These results demonstrate the effectiveness of integrating remote sensing, atmospheric data, and statistical analyses to quantify wildfire-driven pollution events. Although this research focused on the Amazon, where biomass burning is largely associated with deforestation and agriculture, the methodology provides a transferable framework for other regions. Applying similar approaches in Southern and Central Europe, where wildfires are increasingly driven by climate extremes, can improve the understanding of wildfire-pollution interactions and support strategies for air quality management and public health protection.

COMPARAR COM O QUE FOI FEITO NA MINHA TESE?

3.Data

3.1.Study Area

The study area selected for this research is located in Southern and Central Europe, defined by the generic geographical boundaries of latitude 34,5ºN to 66ºN and longitude -12ºE to 36ºE, as illustrated in Figure 1. Within this broader area, four specific regions – Portugal, Spain, Italy and Greece – were chosen due to its high susceptibility to wildfires and pollution episodes. These countries frequently experience large-scale fires, are sensitive to atmospheric pollutant accumulation, and are particularly relevant within the context of Mediterranean climate dynamics. This selection provides an ideal framework for analysing the interactions between wildfire activity and pollutant concentrations across diverse yet climate-vulnerable European regions.

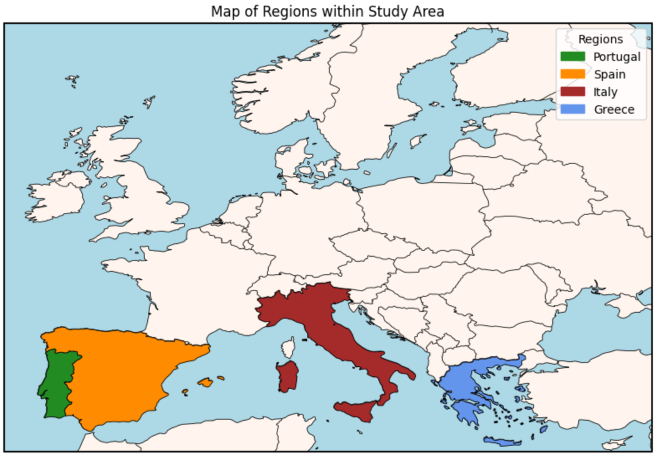


Figure 1 - Regions defined for the current study within the chosen study area

While examining the study area, it is impossible to overlook the devastatinf history of catastrophic wildfires that have marked all chosen countries. Each one has endured tragic fire events that not only results in severe ecological losses but also claimed human lives and destroyed homes, agricultural lands, pastures and other critical infrastructures. Portugal, for instance, experienced one of the deadliest fire seasons in 2017, where 64 people were killed and more than 250 injured, and 30 000 hectares of land burned over 5 days. Other devastating fires were registered in the years of 2013, 2006, 2005 and 2003 (Expatica, 2025). Spain has likewise faced severe wildfire disasters, such as in the summer of 2012, due to a welding negligence and combined with strong winds and high temperatures, more than 30 000 hectares burnt. At the same time this fire was spreading another negligence provoked a second fire in another region that burned 22 500 hectares. Spain had a range of area burned from 11 500 – 47 000 hectares in the years of 2022, 2021, 2017, 2012, 2007, 2005, 2004 and 2003 (Fascinating Spain, 2024). In Italy, . Greece, tragically, has witnessed as well repeated large-scale fires . Together, these events illustrate the current importance of keep developing studies in the matter. NOT FINISHED

3.2.Sources and Datasets

Copernicus is the European Union’s Earth observation program, and main source of data, that provides free and open data about our planet’s environment, by using a network of satellites, ground-based sensors and airborne measurements. It delivers real-time information on climate change, land use, oceans, atmospheric conditions and emergency responses to natural disasters like wildfires or floods. Their data can be used to provide relevant information to help service providers, public authorities, international organizations or even for academic and research purposes (Copernicus, 2025).

This research utilizes data from three main databases:

1. Climate Data Store (CDS) provides the dataset called “ERA5 hourly data on single levels from 1940 to present”. ERA5 is defined as the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather, replacing the ERA-Interim reanalysis. Reanalysis is a method used to create long-term, consistent datasets of past atmospheric conditions by combining weather model simulations with real-world observations, in this way it is reconstructing historical climate and weather patterns instead of predicting the future. A key process behind reanalysis is data assimilation, where new observational data is continuously merged with previous model predictions to improve accuracy. This dataset is updated daily, having a delay of about 5 days that is used to prevent providing inaccurate data (Copernicus, 2025).
2. Atmosphere Data Store (ADS) provides the dataset called CAMS (Copernicus Atmosphere Monitoring Service) global reanalysis (EAC4 - ECMWF Atmospheric Composition Reanalysis 4) that is the fourth-generation global reanalysis of atmospheric composition. It combines model data with worldwide observations through data assimilation, creating a consistent, long-term dataset. By integrating new observations every 12 hours, it improves estimates of atmospheric conditions. Unlike real-time forecasts, reanalysis allows for the incorporation of refined historical data, enhancing accuracy over time (Copernicus, 2025).
3. NASA’s Active Fire Data provides the dataset that contains FRP values. The Fire Information for Resource Management System (FIRMS) is the tool providing near real-time active fire data from the Moderate Resolution Imaging Spectroradiometer (MODIS) carried by Aqua and Terra satellites (NASA, 2025).

3.2.1.Download

Commented [SM3]: Before or after variables?

The process to download data from the two Copernicus databases, ERA5 and CAMS-EAC4, was facilitated by the source providing an Application Programming Interface (API). In each

Commented [SM4]: Explain?

website's database is possible to select different aspects that it's needed in the downloadable dataset, such as variables to use, date range, time of observation, geographical area, and data and download format. In the case of CAMS database, it was also possible to determine pressure and model levels.

Because these two databases have different time ranges, from the ERA5 database it was downloaded the years 1979 to 2024 for every month, day and time available, for all the variables needed and generic area of study. The data retrieved from CAMS-EAC4 was done in two separate API's, where it is determined by single level and multi-level. In the latter it was opted to download with a pressure level of 1000 hPa, to represent values from the surface. The downloaded years for this dataset was 2003-2024 and also for every time available.

Commented [SM5]: Mencionar variaveis antes ou depois?

All data files were provided in a zip file that contained its respective Network Common Data Form (NetCDF) files. Using this kind of format allows to store large and multidimensional datasets.

Data about wildfires was provided by FIRMS's way of downloading its data that required creating a new request with the wanted data. It consisted in selecting a custom region – the same coordinates used to define the study area. The fire source was MODIS (both Aqua and Terra spacecrafts) from the most from the first data available (November 2000) to end of 2024 in a csv format. The request was approved and sent to via email.

In order to be able to correlate every data downloaded, it was decided to use only the period of time that was identical throughout the data (2003-2024).

3.3.Variables

3.3.1.Meteorological Data

The ERA5 dataset has a global coverage of gridded data with a horizontal resolution of 0.25ºx0.25º, it equivales to approximately 28 km. It provides data from 1940 to present and has a temporal resolution of 1 hour. It contains many variables, but for this study, these are the ones that are going to be used:

Commented [SM6]: Get this verified

- 10m u-component of wind ($m.s^{-1}$) – horizontal speed of air moving towards the east, at 10 meters above the surface of the Earth (Copernicus, 2025);
- 10m v-component of wind ($m.s^{-1}$) – horizontal speed of air moving towards the north, at 10 meters above the surface of the Earth (Copernicus, 2025);
- 2m temperature (K) – air temperature at 2 meters above the surface of land, sea or inland waters (Copernicus, 2025);

- Boundary layer height (m) – it represents the depth of air next to the Earth’s surface which is most affected by the resistance to the transfer of momentum, heat or moisture across the surface. The values can vary, when low as a couple of meters – such as cooling air, or much higher values, when talking about a desert in the middle of a hot sunny day. It is also stated that if the boundary layer height is low, there can be present higher concentrations of pollutants, emitted from the Earth’s surface (Copernicus, 2025);

Commented [SM7]: Not being used at the moment

- Total precipitation (m) – this variable contains the accumulated liquid and frozen water that falls to the Earth’s surface. It sums the large-scale precipitation and convective precipitation. The cloud scheme in the ECMWF, Integrated Forecasting System (IFS), represents the formation and dissipation of clouds and large-scale precipitation. The convective precipitation is generated by the convection scheme in the IFS. Total precipitation parameter does not include other forms of water, like fog, dew or precipitation that evaporates in the atmosphere (Copernicus, 2025).

3.3.2.Pollutant Data

The CAMS-EAC4 dataset has a global coverage with a gridded horizontal resolution of 0.75ºx0.75º, equivalent to approximately 80 km. It has a temporal coverage of 2003 to 2024, where this dataset is updated twice a year with 4 to 6 month delay. Its temporal resolution is of 3 hours and has a vertical resolution of 60 models and different pressure levels (hPa). The variable retrieved from this dataset were divided into 2 groups:

- Single-level:
    - Particulate matter d < 1 µm (PM_1) ($kg.m^{-3}$)
    - Particulate matter d < 2.5 µm ($PM_{2.5}$) ($kg.m^{-3}$)
    - Particulate matter d < 10 µm (PM_{10}) ($kg.m^{-3}$)
- Multi-level, where it was specified to retrieve data from pressure level 1000 hPa, that correspond to the Earth’s surface:
    - Carbon monoxide ($kg.kg^{-1}$)
    - Nitrogen dioxide ($kg.kg^{-1}$)
    - Nitrogen monoxide ($kg.kg^{-1}$)
    - Ozone ($kg.kg^{-1}$)

Commented [SM8]: Not being used but could be referenced for future improvements or works

3.3.3.Wildfire Data

The FRP dataset represents the radiative energy released by active fires, derived from satellite sensors. The point dataset contains the respective coordinates of the observed energy, contains the date and time of the registration, the value of the FRP in MW, among others but only this will be relevant to the research.

EXPLAIN FIRE FOCI

Commented [SM9]: But where????

3.4. Data Processing/Preparation

Commented [SM10]: Desenvolver

3.4.1. Masks

After selecting which countries to study, it was essential to transform those into some kind of masks that were a good representative of respective region. Based on the chosen grid, the one corresponding to the CAMS dataset (0,75ºx0,75º),), it was developed a way to transform shapefiles into pixel polygons. Because Portugal and Spain share boundaries, it was defined to attribute each pixel that fell in the boundary to the one country that contained more area (this was made to ensure that the majority of fires that occur in Spain would be well represented). The results, Figure 2, were later used to perform different statistics for its respective countries.

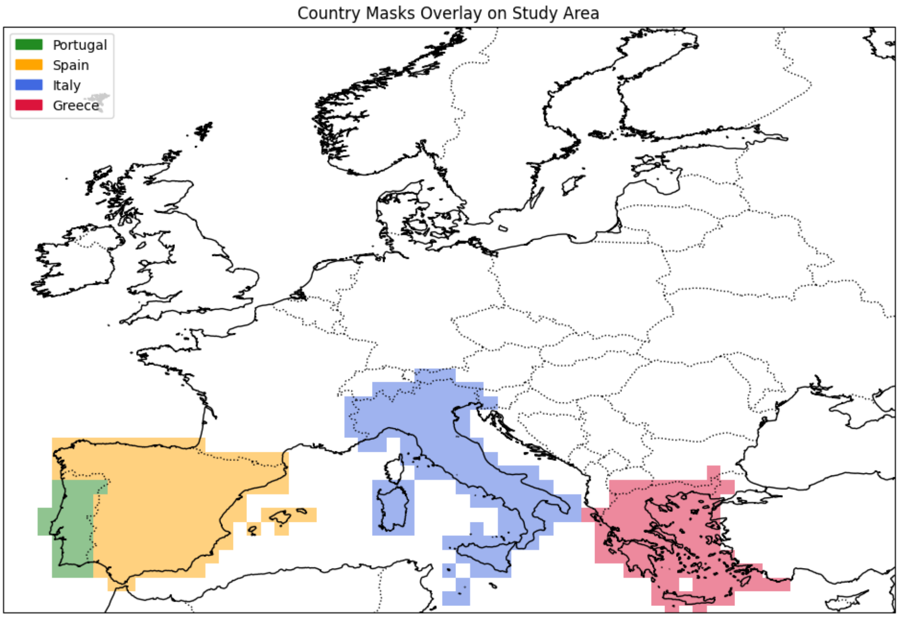


Figure 2 - Pixel masks created to represent the countries chosen within study area. Resolution 0.75ºx0.75º

Renaming some variables names to match given data.

Conversion of coordinates from 0º to 360º to -180º to 180.

Crop to study area coordinates.

Unit conversion was performed to the pollutants data to match the WHO air quality guidelines recommended levels, available at (WHO, 2025). These guidelines will further be used to study the subject. Scripts were created to make the following conversions:

- Pollutant concentration multiplied by pressure and divided by the multiplication of R and temperature, where pressure = 1e5 Pa (equal to 1000 hPa) and R = 287.0500676 (J/kg.K). $\frac{[Pollutant] * pressure}{(R * t)}$
    - Carbon monoxide (kg.kg-1) to mg/m^3
    - Nitrogen dioxide (kg.kg-1) to µg/m^3
    - Nitrogen monoxide (kg.kg-1) to µg/m^3
    - Ozone (kg.kg-1) to µg/m^3
- Multiply values by 1e9
    - Particulate matter d < 1 µm (PM1) (kg.m-3) to µg/m^3
    - Particulate matter d < 2.5 µm (PM2.5) (kg.m-3) to µg/m^3
    - Particulate matter d < 10 µm (PM10) (kg.m-3) to µg/m^3
- Temperature from Kelvin to Celsius, subtract 273.15

Masks for each region of the study were created for future usage in analysing the influence of wildfires in pollutants concentrations. Shapefiles were downloaded for its respective region in EPSG:4326 (WGS84) from the Dataset of Global Administrative Areas (GADM) official website. Based on the same grid parameters as the pollutant's datasets and boundaries of the defined study area, the creation of these masks will also follow the same settings. For Portugal and Spain since they share boundaries, it was created a restriction that chooses which pixel to assign based on the maximum area in the grid cell, this was set based on the observation of the predominance of intense fires closer to their boundaries, like it is possible to see in Figure 2. For Italy and Greece, the grid cells don't overlap so it processes the intersection with the shapefiles normally. The created masks are represented in Figure 3.

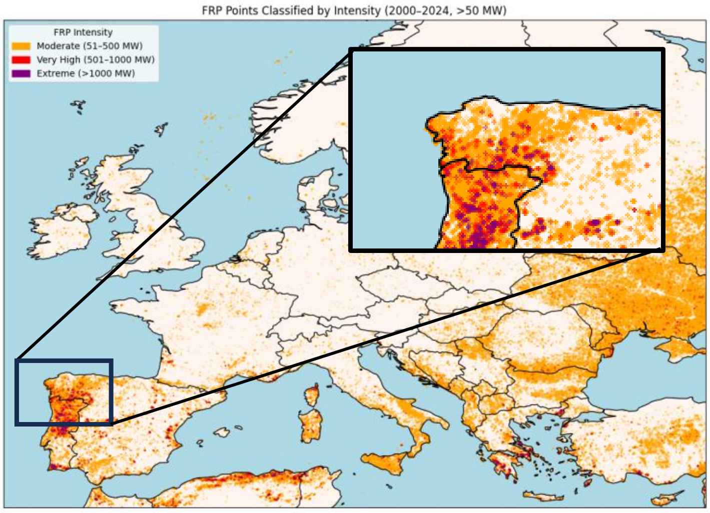

FRP Points Classified by Intensity (2000–2024, >50 MW)
FRP Intensity
🟧 Moderate (51–500 MW)
🟥 Very High (501–1000 MW)
🟪 Extreme (>1000 MW)

Figure 3 - FRP points discriminated by their intensity

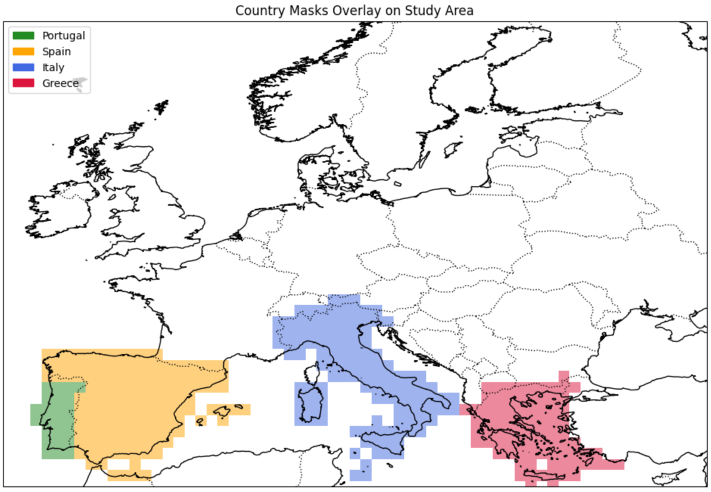


Figure 4 - Masks of each region over study area

The raw FRP dataset contains fires of low intensity (≤ 50 MW), such as minor agricultural fires – prescribed burns, that were removed because they don’t impose significant alterations in pollutants concentration levels.

3.5.Data Understanding

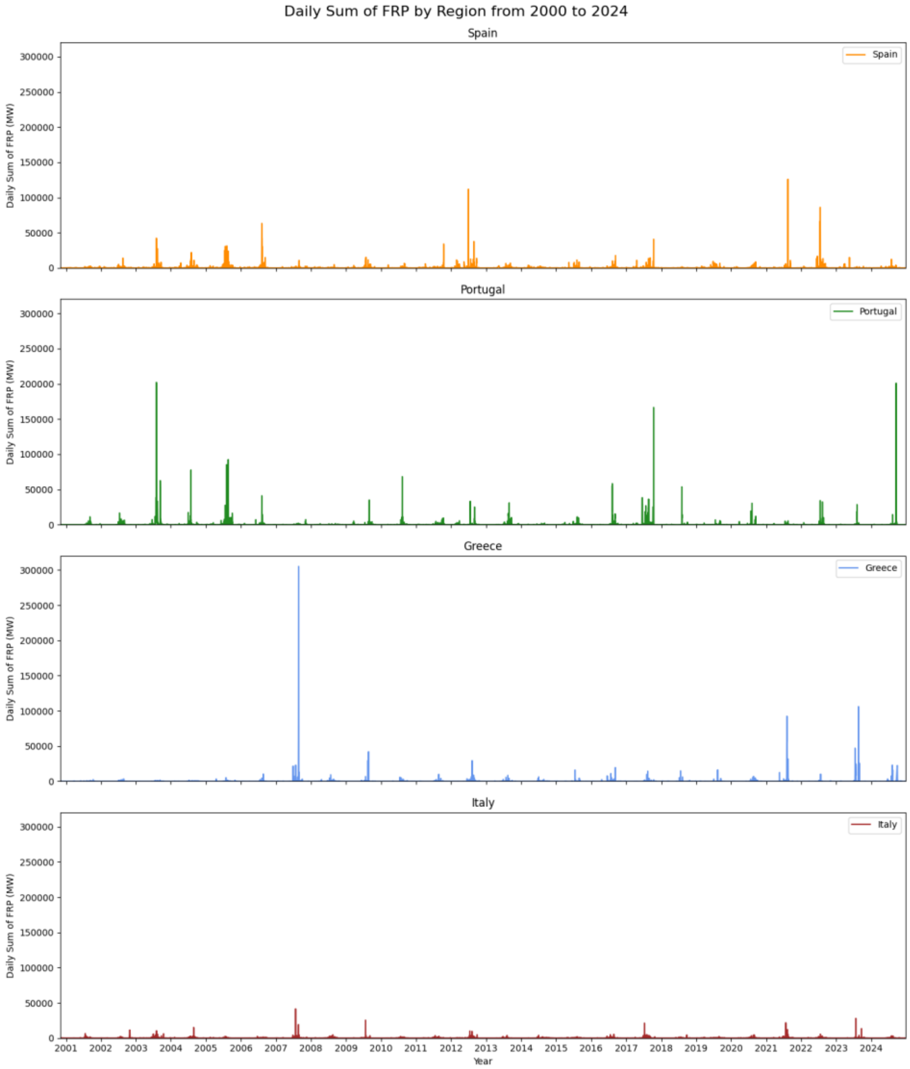
To investigate the temporal dynamics of wildfires and their potential influence on air quality, time series analysis were conducted for both FRP and atmospheric pollutant concentrations. These plots allow to better understand trends, seasonal patterns and interannual variability across the respective data.

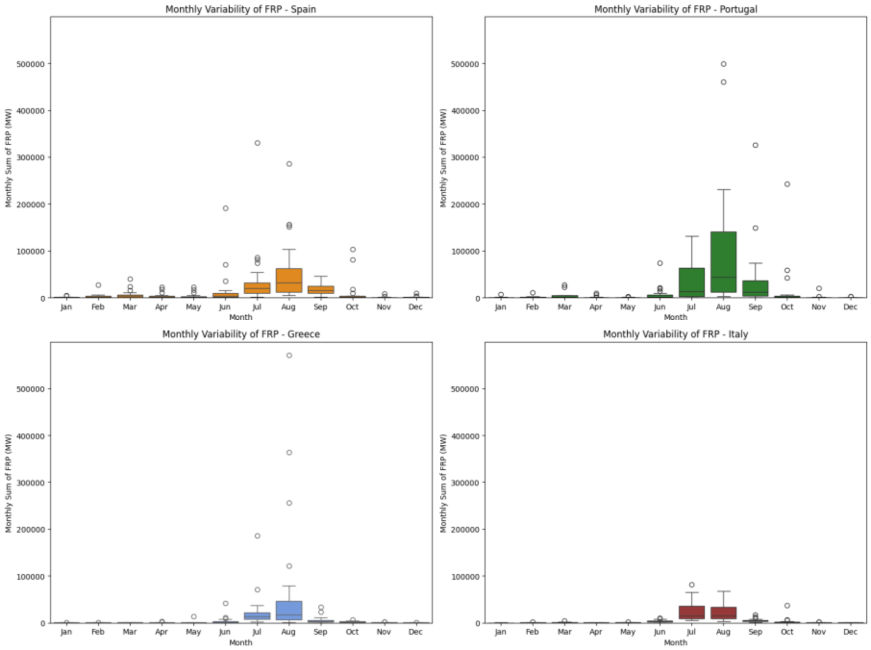
3.5.1.Pollutant Data

3.5.2.Wildfire Data

In Figure x, it is represented the time series of the daily sum of FRP values, in MW, across the years of the dataset for each region. Higher values indicate larger fires occurred in those years.

In Figure x, it shows the monthly variability across all years, and as expected the northern hemisphere summer months (July, August and September) show fire predominance.


Daily Sum of FRP by Region from 2000 to 2024
Spain
Portugal
Greece
Italy
Daily Sum of FRP (MW)
300000
250000
200000
150000
100000
50000
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year



Monthly Variability of FRP - Spain

Monthly Variability of FRP - Portugal

Monthly Variability of FRP - Greece

Monthly Variability of FRP - Italy

4.Methodology

Machine learning (ML) is a subset of artificial intelligence (AI) that allows systems to learn from data and improve their performance over time without explicit programming, playing a crucial role in extracting insights, detecting patterns, and automating processes across different topics. While all ML is considered AI, not all AI encompasses machine learning. As data generation grows exponentially, ML helps businesses analyse vast and large datasets to uncover trends and make predictions. It encompasses different learning types, including supervised, unsupervised, semi-supervised, and reinforcement learning – each suited for different tasks. **Supervised learning** relies on labelled data to train models, such as predicting wildfire occurrences based on historical data and meteorological variables, while **unsupervised learning** identifies patterns and clusters in unlabelled data, like grouping areas with similar wildfire risk levels. **Semi-supervised learning** combines both approaches to reduce labelling costs, iteratively improving models with pseudo-labelling. **Reinforcement learning**, on the other hand, involves trial-and-error learning with feedback, commonly used in optimizing fire suppression strategies or assessing the effectiveness of air quality control measures during wildfire events. ML uses algorithms like **neural networks**, which mimic the human brain to analyse complex data like images; **linear regression**, that predicts continuous outcomes by fitting a line to data points but may struggle with non-linear relationships; **logic regression**, used for binary outcomes like wildfire-induced health impacts; **clustering**, an unsupervised method that groups similar data points into clusters for tasks like mapping areas with similar air pollution profiles due to wildfires; **decision trees**, which make predictions using simple if-then rules; and **random forests**, an ensemble of decision trees that improve prediction accuracy by addressing limitation like overfitting. By training algorithms on data, ML models identify complex relationships, enabling automatic pattern discovery, predictive analysis and large-scale data processing. This makes it crucial to select the right learning approach based on data structure, resources, and application needs (Chen, 2024).

Commented [SM11]: Add a more detailed version

Machine learning models are built by training statistical algorithms on data rather than relying on predefined rules. The development process follows key steps: (1) **Data collection** – gathering and evaluating high-quality data, which may require extensive preprocessing and labelling; (2) **Algorithm selection** – choosing the appropriate approach (supervised, unsupervised, semi-supervised, or reinforcement learning) based on the problem; (3) **Data preparation** – cleaning, transforming, and structing data for efficient processing; (4) **Model training** – feeding the data into the algorithm, adjusting hyperparameters, and iterating to improve performance; (5) **Performance assessment** – evaluating accuracy using test data and refining the model accordingly; (6) **Fine-tuning** – further optimizing parameters and incorporating domain-specific data; (7) **Deployment and monitoring** – integrating the model into real-world application, tracking its performance, and making periodic updates to maintain accuracy and relevance. Continuous evaluation ensures that models remain effective and aligned with business or research objectives (Chen, 2024).

Atmospheric dispersion models are mathematical models used to simulate how air pollutants (gases and particles) spread in the atmosphere. These models take into account meteorological conditions, chemical reactions, and physical dispersion processes to estimate pollutant concentrations at different locations. There are two main types of dispersion models: Gaussian models used for local, near-source dispersion (e.g., air pollution from traffic); Eulerian and Lagrangian models that are used for regional or global scales, incorporating meteorological and chemistry (Holmes & Morawska, 2006).

Commented [SM12]: Identify caveats of these approaches to highlight the research gap

Model Development:

- Supervised Machine Learning Models: Implement and compare models (e.g., Random Forests, XGBoost, and Neural Networks) to classify pollution events and identify anomalies related to wildfire smoke.
- Spatial and Temporal Analysis: Use geospatial tools to map the spatial reach of pollution plumes, identifying the temporal dynamics of smoke dispersal patterns.

Evaluation and Validation:

- Validate the model using metrics such as accuracy, F1-score, and area under the curve (AUC) for classification tasks.
- Cross-reference predicted pollution events with historical extreme events (e.g., Portugal's 2017 megafires) to validate spatial and temporal accuracy. Analyse the impact of compound events (e.g. droughts and heatwaves) to the magnitude of fire-driven pollution events.

5.Results

Model Performance:

- Present the results of model performance, discussing the predictive accuracy in detecting pollution events and identifying contributing factors from FRE, FRP, and meteorological data.

6.Discussion

Spatial and Temporal Impact Analysis:

- Discuss the spatial and temporal patterns of wildfire smoke distribution. Evaluate how FRP and FRE data correlate with pollutant concentrations across affected regions.

Limitations and Future Improvements:

- Discuss potential limitations of the model, such as sensitivity to specific atmospheric conditions or data quality, and suggest directions for future improvement.

7.Conclusion

Summary of Findings:

- Summarize the effectiveness of remote sensing and meteorological data in detecting and mapping wildfire-induced pollution events.

Implications:

- Discuss the potential of this approach to support decision-making interventions and policy planning.

Future Work:

- Suggest advancements in real-time monitoring systems, the inclusion of additional data sources, and potential integration with other environmental and health monitoring frameworks.

References

AghaKouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasni, O., . . . Sadegh, M. (2020). *Climate Extremes and Compound Hazards in a Warming World.* Annual Reviews. doi:https://doi.org/10.1146/annurev-earth-071719-055228

Bento, V. A., Lima, D. C., Santos, L. C., Lima, M. M., Russo, A., Nunes, S. A., . . . Soares, P. M. (2023). The Future of extreme meteorological fire danger under climate change scenarios for Iberia. *Elsevier, 42*. doi:https://doi.org/10.1016/j.wace.2023.100623

Chen, M. (25 de November de 2024). *What is Machine Learning?* Obtido em 26 de February de 2025, de oracle.com: https://www.oracle.com/uk/artificial-intelligence/machine-learning/what-is-machine-learning/

Copernicus. (2025). *About Copernicus*. Obtido em 25 de February de 2025, de copernicus.eu: https://www.copernicus.eu/en/about-copernicus

Copernicus. (2025). *CAMS European air quality reanalyses - Overview*. Obtido em 26 de February de 2025, de ads.atmosphere.copernicus.eu: https://ads.atmosphere.copernicus.eu/datasets/cams-europe-air-quality-reanalyses?tab=overview

Copernicus. (2025). *ERA5 hourly data on single levels from 1940 to present - Overview*. Obtido em 25 de February de 2025, de cds.climate.copernicus.eu: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview

De Sario, M., Katsouyanni, K., & Michelozzi, P. (2013). *Climate change, extreme weather events, air pollution and respiratory health in Europe.*

Durao, R., Alonso, C., Russo, A., & Gouveia, C. (14-19 de April de 2024). *The role of fire radiative power to estimate fire-related smoke pollution.* Vienna, Austria: EGU General Assembly 2024. doi:https://doi.org/10.5194/egusphere-egu24-13237

Holmes, N. S., & Morawska, L. (2006). A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Elsevier*, 5902-5928. doi:https://doi.org/10.1016/j.atmosenv.2006.06.003

Instituto de Meteorologia. (23 de March de 2009). *Fire Radiative Power Pixel - MSG*. Obtido em 24 de February de 2025, de geonetcastamericas.noaa.gov: https://www.geonetcastamericas.noaa.gov/products/navigator/details/EO_EUM_DAT_MSG_FRP-SEVIRI.html

Knorr, W., Dentener, F., Lamarque, J.-F., Jiang, L., & Arneth, A. (2017). *Wilfire air pollution hazard during the 21st century.* Atmospheric Chemistry and Physics. doi:https://doi.org/10.5194/acp-17-9223-2017

Li, F., Zhang, X., Kondragunta, S., & Csiszar, I. (2 de May de 2018). Comparison of Fire Radiative Power Estimates From VIIRS and MODIS Observations. *Journal of Geophysical Research: Atmospheres*, 4545-4563. doi:https://doi.org/10.1029/2017JD027823

World Health Organization. (2025). *Air pollution*. Obtido em 25 de February de 2025, de who.int: https://www.who.int/health-topics/air-pollution#tab=tab_1

World Health Organization. (2025). *Wildfires*. Obtido em 24 de February de 2025, de who.int: https://www.who.int/health-topics/wildfires#tab=tab_1

World Meteorological Organization. (2025). *Extreme Weather*. Obtido em 24 de February de 2025, de wmo.int: https://wmo.int/topics/extreme-weather

Yao, J., & Henderson, S. B. (2013). An empirical model to estimate daily forest fire smoke exposure over a large geographic area using air quality, meteorological, and remote sensing data. *Journal of Exposure Science and Environmental Epidemiology (2014), 24*, 328-335. doi:doi:10.1038/jes.2013.87

Appendices