



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



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 Neurais, 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

Introduction	10
State of the Art	13
Data	14
Methodology	15
Results.....	18
Discussion.....	19
Conclusion	20
References	21
Appendices	23

List of Figures

Não foi encontrada nenhuma entrada do índice de ilustrações.

List of Tables

Não foi encontrada nenhuma entrada do índice de ilustrações.

List of Abbreviations

IPMA – Instituto Português do Mar e da Atmosfera

MODIS – Moderate Resolution Imaging Spectroradiometer

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

CAMS – Copernicus Atmosphere Monitoring Service

PM – Particulate Matter

Introduction

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 (O₃), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂). 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).

Wildfires are a derivative of a fire that is uncontrolled and can take large proportions because they are usually situated in rural 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 cause 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 PM_{2.5}, NO₂ and O₃ (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 impact the climate when they release large amounts of carbon dioxide (CO₂) and other greenhouse gases into the atmosphere (World Health Organization, 2025).

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). That being said, there is a concern that wildfires will become more aggravated over the years and with that, the air quality will degrade even more. 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.

Commented [SM1]: examples

Commented [SM2]: Add examples of the impacts and then pass to the smoke

Traditional approaches to 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. 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. 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.

Commented [SM3]: Find references

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 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).

Commented [SM4]: add some examples

The study will use data from two different sources that will be described in more detail further. Copernicus is the source for two different datasets – 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 European air quality reanalyses, which estimate annual air quality in Europe using data assimilation from eleven models, improving accuracy through an ensemble median approach (Copernicus, 2025).

MODIS

To obtain data related to wildfires... <https://modis.gsfc.nasa.gov/data/> or <https://sa-saf.eumetsat.int/en/data/products/fire-products/>

Commented [SM5]: Adicionar se for esta a fonte para o FRP e FRE

Fire Radiative Power (FRP) represents the radiative energy released by active fires, derived from satellite sensors. It provides insights into biomass combustion rates, helping estimate trace gas and aerosol emissions (Li, Zhang, Kondragunta, & Csiszar, 2018). Higher FRP values indicate more intense fires and greater smoke production, leading to larger emissions of particulate matter and other pollutants (Durao, Alonso, Russo, & Gouveia, 2024). Fire Radiative Energy (FRE), can be calculated by integrating FRP over a fire's duration, estimating the total biomass consumption (Instituto de Meteorologia, 2009).

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

State of the Art

- An empirical model to estimate daily forest fire smoke exposure over a large geographic area using air quality, meteorological and remote sensing by Jiayun Yao ([Yao & Henderson, 2013](#))
- Machine Learning & Big Data Analyses for Wildfire & Air Pollution Incorporating GIS & GEE by Abdullah Al Saim
- Multimodal Wildland Fire Smoke Detection by Siddhant Baldota
- Remote sensing and model analysis of biomass burning smoke transported across the Atlantic during the 2020 Western US wildfire season by Xavier Ceamanos
- Wildfire air pollution hazard during the 21st century by Wolfgang Knorr
-

Data

EXPLICAR ERA5 E CAMS DATA

Data Collection:

- Meteorological Data: Gather atmospheric data, such as temperature, humidity, wind speed, and direction, from sources like CAMS and monitoring stations.
- Remote Sensing Data: Use FRE and FRP data from MODIS, SEVIRI, and Sentinel-3 for monitoring fire radiative power and energy release.

Data Processing:

- Integrate meteorological and remote sensing data with pollution monitoring station data to create a comprehensive dataset.
- Preprocess data by cleaning and transforming it to address missing values, noise, and scaling for model input.

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).

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

Commented [SM6]: Add a more detailed version

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).

The study will have data from two/three different sources that will be described in more detail further. Copernicus is the European Union's Earth observation program 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). From Copernicus there are two sources that are going to be used to retrieve data – Climate Data Store (CDS) and Atmosphere Data Store (ADS). From CDS, the dataset that is going to be used is 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 (Copernicus, 2025). The other source, ADS, provides the dataset called CAMS (Copernicus Atmosphere Monitoring Service) European air quality reanalyses and it provides annual air quality for Europe based on unvalidated and validated observations. CAMS relies on eleven different air quality models that use data assimilation to merge observational data with model predictions. The final air quality estimate is based on the median ensemble approach, meaning the median value of all eleven models is taken. It somewhat improves accuracy since ensemble methods tend to perform better than any single model alone (Copernicus, 2025).

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 [SM7]: Reduce and move the most to methods

Commented [SM8]: Identify caveats of these approaches to highlight the research gap

Fire Radiative Power (FRP), in megawatts (MW), is one of the variables used to characterize wildfires, it is obtained “from the radiance at the 4- μ m 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) . Or <https://user.eumetsat.int/catalogue/EO:EUM:DAT:MSG:FRP-SEVIRI>

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.

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.

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.

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., Mazdiyasn, 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). *Wildfire 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